THE META OVERSIGHT BOARD
AND THE EMPTY PROMISE OF LEGITIMACY

Evelyn Douek*


ABSTRACT

The Meta Oversight Board is an audacious experiment in self-regulation by one of the world’s most powerful corporations, set up to oversee one of the largest systems of speech regulation in history. In the few years since its establishment, the Board has in some ways defied its many skeptics, by becoming a consistent and accepted feature of academic and public discourse about content moderation. It has also achieved meaningful independence from Meta, shed light on the otherwise completely opaque processes within the corporation, instantiated meaningful reforms to Meta’s content moderation systems, and provided an avenue for greater stakeholder engagement in content moderation decisionmaking. But the Board has also failed to live up to core aspects of its role, in ways that have gone underappreciated. The Board has consistently shied away from answering the hardest and most controversial questions that come before it—that is, the very questions it was set up to tackle—and has not provided meaningful yardsticks for quantifying its actual impact. Understanding why the Board eschews these questions, and why it has nevertheless managed to acquire a significant amount of institutional legitimacy, suggests important lessons about institutional incentives and the revealed preferences of stakeholders in content moderation governance. Ultimately, this Article argues, the current political environment incentivizes a kind of oversight that is formalistic and unmoored from substantive goals. This is a problem that plagues regulatory reform far beyond the Board itself, and shows that generalized calls for “more legitimate” content moderation governance are underspecified and may, as a result, incentivize poor outcomes.

* Assistant Professor, Stanford Law School. Many thanks to Anupam Chander, Ros Dixon, Thomas Kadri, Genevieve Lakier, Clare Mathias, Martha Minow, and Rory Van Loo. All oversights my own.

Electronic copy available at: https://ssrn.com/abstract=4565180
INTRODUCTION

Over the past three years, a single institution has adjudicated whether the President of the United States should be able to use one of his preferred channels of communication with an audience of over 35 million people,\(^1\) how much weight should be given the UK Metropolitan Police’s assessments of the dangers of certain music,\(^2\) whether COVID-19 misinformation should be suppressed online,\(^3\) and how to deal with inflammatory and threatening statements from the Prime Minister of Cambodia.\(^4\) The same institution has decided disputes that touch on some of the world’s most contentious subjects, from conflict between Israel and Palestine,\(^5\) the invasion of Ukraine,\(^6\) and the pervasiveness of gender-based violence.\(^7\) It has engaged with some of the hardest, most consequential questions about speech regulation in the modern world—questions about which there is profound, intractable societal disagreement. And it has done so despite a complete lack of formal legal authority.

The institution is the Meta (née Facebook\(^8\)) Oversight Board (the Board), and it sits in a twilight zone between being a decisionmaker exercising immensely consequential power, and being a made-up body that exists at the whim of its creator. Its “case decisions” look like court rulings, but they technically bind no one. It purports to interpret and apply state-created international human rights law (IHRL), but it has no state-given mandate to do so. The Board’s rulings have impact in the real world, and yet its power is somewhat a fiction. Created voluntarily by Meta as a form of self-regulation,

---

\(^1\) Case Decision 2021-001-FB-FBR, OVERSIGHT BD. (2021), https://www.oversightboard.com/decision/FB-691QAMHJ (Trump Suspension case).
\(^8\) For simplicity, this paper will refer throughout to the parent company of the platforms Facebook and Instagram as “Meta,” even when referring to documents or events that preceded the company’s formal name change from “Facebook.”
the Board’s decisions have force only insofar and for as long as Meta—the company the Board was set up to hold accountable—agrees they do.

And yet, despite this significant handicap on its authority, the Board has received a remarkable degree of attention and respect in the academic and media discourse around content moderation. Its decisions regularly get reported in leading media outlets like The New York Times,10 Washington Post,10 Reuters11 and The Guardian.12 This news coverage of the Board resembles, in tone and content, reporting on judicial decisions or administrative agencies. “Meta Oversight Board calls for Cambodian leader’s accounts to be suspended” reads one headline, for example.13 “Instagram told to reinstate music video removed at request of Met police” says another.14 Reporters often treat the Board, in other words, as a meaningful source of authority whose decisions are news that its readers should know about.

It’s not only the media that consistently and meaningfully engages with the Board in ways that both suggest and bestow a degree of legitimacy. Numerous academics and civil society groups have submitted public comments to the Board with respect to particular cases, including the ACLU,15 the United Nations Special Rapporteur on Minority Issues,16 the

---

14 Hern, supra note 13.
International Commission of Jurists,\textsuperscript{17} the Electronic Frontier Foundation,\textsuperscript{18} and many other leading institutions and individuals in the digital governance space and beyond.\textsuperscript{19} Government actors have done the same: American lawmakers have submitted comments to the Board,\textsuperscript{20} as have the U.K. Metropolitan police.\textsuperscript{21} The Israeli Supreme Court has referenced the Oversight Board in one of its judgments as an example of due process that Meta offers its users.\textsuperscript{22} The United Nations High Commissioner for Human Rights has approvingly noted the Board’s work in its annual report.\textsuperscript{23} The Board has also established academic cachet. It has been written about extensively.\textsuperscript{24} Its decisions are included alongside judicial opinions and legislative materials in the most prevalent legal databases Westlaw and Lexis+.\textsuperscript{25} The Board’s decision in the Trump Suspension case is included in a leading internet law casebook,\textsuperscript{26} and another opinion has even been the subject of a case note in the Harvard Law Review.\textsuperscript{27}

While this degree of attention to, and acceptance of, the Board might one
day seem completely unremarkable, it’s important to underline for readers from that future that this was far from inevitable when the Board was created. Early reactions to Meta’s decision to create the Board were extremely skeptical. Many suggested that the Board should not be engaged with at all, because to do so was either fruitless or, worse, would necessarily legitimize a corrupt institution. The Board was described as a “folly,”28 “dysfunctional by design” and “toothless[],”29 “a McGuffin,”30 having a “Potemkin quality,”31 and many other similar things in print—and probably some much worse off the record. Such skepticism is not surprising: the Board was the unilateral creation of a corporate, profit-driven entity, and was given a narrowly defined remit and no coercive legal power. These are not typically the characteristics that inspire unmitigated confidence or widespread public respect for a decisionmaking authority. As such, the Board started with a very low baseline of sociological legitimacy, and understandably so.

Yet only a few years later, skepticism of the Board has receded. Far from being dismissed as a mere sham, the Board’s existence is largely taken for granted, and its work is taken increasingly seriously. This Article asks how this change came about, whether the shift in its public and academic reception is justified by what the Board has achieved so far, and what lessons this holds for social media governance more generally. Understanding how the Board has come to occupy the position it does in public discourse matters not only because the Board oversees what is by volume one of the most expansive speech forums in history, but also because the Board experiment sheds light on broader dynamics in content moderation governance. How stakeholders have engaged with the Board reveals what they want and expect from such governance institutions. And as regulators, companies and netizens alike search for ways to bring greater oversight to content moderation systems, the Board holds both salutary lessons and cautionary tales for these institutional designers.

On the one hand, and contrary to many people’s expectations, the Board has not been entirely toothless. It has chafed against the extremely limited remit Meta originally gave it, pushed the limits of its authority, and in doing

31 Kevin Roose, Facebook’s ‘Supreme Court’ Tells Zuckerberg He’s the Decider, N.Y. TIMES (May 6, 2021), https://www.nytimes.com/2021/05/06/technology/facebook-oversight-board-trump.html.
so has sought out broader impact on Meta’s systems than merely deciding individual cases. It has as a result brought about meaningful reforms at Meta, some of which civil society and activists had been seeking for years to no avail. And it has shed light on some of the otherwise entirely opaque systems that make millions of speech decisions every day. It has been rewarded by becoming in many ways part of the platform governance establishment.

But the Board has also underdelivered in underappreciated ways and failed to fulfill some of its most important tasks. First, despite its design centering the Board’s role of public reason-giving, the Board has not explained the normative framework that guides its thinking about online speech governance. This is a huge missed opportunity—the central normative difficulty of content moderation is how to adapt principles created to constrain governmental power over expression to the very different context of private content moderation systems. This task is hard, but it is vital in an age where corporate power over everyday speech has never been greater. There is little in modern public discourse that is not affected by the decisions that platforms make—everything is a content moderation problem. Thus, developing a normative foundation for speech rights that goes beyond protection from governmental interference is essential if free speech principles are to be made meaningful in the platform era. The Board is uniquely placed to do this work as one of the most high-profile experiments in content moderation governance, made up of free expression experts who have access to an unending supply of content moderation cases to review and significant resources at their disposal. And yet, the Board has largely eschewed these normative questions. It has adopted IHRL as the basis for its decisions, but its application of IHRL to Meta’s platforms has been shallow and provides little insight into what understanding of free speech guides its work. This undermines not only the predictability of the Board’s decisions, but also the Board’s role as a source of public reasoning that explains and justifies the rules that govern a major segment of the online public sphere.

Second, if the Board’s normative contributions to content moderation debates have been underwhelming, its material impacts on online speech are altogether harder to quantify, and the task of quantification has been made more difficult by the Board’s emphasis on metrics that are easier to measure but poor proxies for actual impact. The Board tracks Meta’s implementation of its decisions and recommendations, but it does not track the more

32 See infra Part II.A.2.
33 See infra Part II.A.3.
34 See infra Part II.A.4.
35 See infra Part II.B.1.
important question of whether and how those changes actually improve people’s lived experience and the exercise of their rights. In short, the Board has adopted a formalistic approach to the kinds of changes it seeks from Meta, and a simplistic approach to measuring the results.

Third, in high-profile or especially hard cases, the Board has all too often ducked its responsibility to act as the decider of last resort and avoided giving a definitive answer altogether. This might have been thought to be the raison d'être of the Board. But instead the Board has found ways to be as uncontroversial as possible, even when deciding the most controversial cases. The result has been to prolong, rather than resolve, some of the most important and high-stakes arguments about online speech rules.

The Board’s failure to deliver in these important ways raises questions about why the Board has adopted the approach that it has, and why it has succeeded in garnering increasing sociological legitimacy regardless. This puzzle is the core question that this Article ultimately seeks to answer. What it shows is that “legitimacy,” although often invoked as the lodestar of successful governance, is an ambiguous indicator of what we might want from governance institutions, and its pursuit can create perverse institutional incentives. This insight is critical for platform governance institutional design in general, and not merely for the Board.

This Article proceeds in three parts. Part I describes the beginnings and ends of the Board. It briefly sets out the history of the Board’s creation and the design choices of its creators, and shows how the pursuit of “legitimacy” was core to the Board’s purpose. The Board was fundamentally a response to a growing perception that Meta’s content moderation practices were illegitimate, and that the exercise of such incredible power over so much speech should not be so unaccountable Therefore, while stakeholders articulated many things that they hoped the Board would do—push back against pressure to remove ever more content online, reduce Meta’s power (and responsibility), provide an avenue for more democratic input in content moderation governance, bring greater accountability and transparency to Meta’s decisionmaking, provide public reasoning and rationales for Meta’s most consequential and controversial decisions—ultimately all these goals were in service of making Meta’s content moderation “more legitimate.”

Part II then engages in a performance review. It canvasses the Board’s successes and the ways in which the Board has delivered on some of its promise as an institution. Then it turns to the Board’s shortfalls, and the ways it has engaged in a formalistic kind of governance that fails to provide firm foundations for, or strong evidence of, the benefits of its work. What this Part shows is that while the Board has achieved much more than many expected, it has also underdelivered in ways that are both significant and underappreciated.
Part III then gives an explanation for why the Board shies away from core parts of its role, and why this has not (yet) had reputational costs. Ultimately, it argues, what the Board’s (albeit qualified) success suggests is that the lack of any consensus as to what effective content moderation governance should do creates space for institutions to define success in ways that may suit them but may have little benefit for others. As long as the Board seems better than the available alternatives for decisionmaking (in this case, completely opaque and unaccountable power resting in the hands of Mark Zuckerberg), and exhibits some features of “good governance,” it appears to be a valuable improvement over the status quo. Indeed, it is an improvement. But that is a low bar, and we should demand more from the institutions that oversee some of the most important, and definitely the most expansive, speech forums in history. As an exercise in the revealed preferences of stakeholders in social media governance, then, the story of the Board is somewhat discouraging. While the Board has had some successes, ultimately it has so far carried out a form of governance that prioritizes form over substance. The danger is not that the Board’s choices are seen as, or even are, “illegitimate”—it’s that they are seen as legitimate, unconnected from any other outcome goal we might hope for content moderation governance to pursue and that these perceptions influence the design and actions of all content moderation governance, far beyond the Board.

I. THE BOARD’S BEGINNING AND ENDS

The origins of the Board reflect a pervasive anxiety about the “legitimacy” of content moderation governance. The Board was created at a moment when long-standing concerns about the unaccountable power platforms wielded over the digital public sphere had reached a crescendo, and was Meta’s attempt at placating these concerns. This Part briefly describes these origins and the design of the Board, before turning to examine its goals and how it was intended to meet them. What this story shows is that the mechanisms by which the Board should improve content moderation governance were always somewhat vaguely articulated. But the core contribution the Board was intended to make was always clear: build trust and legitimacy for Meta’s decisionmaking about the difficult, contestable and

consequential online speech decisions that the company has to make every day.

A. The Design

The fact that the Board was an attempt by Meta to shape public discourse was clear from the start. Meta’s CEO Mark Zuckerberg first publicly floated the idea of a “some sort of structure, almost like a Supreme Court” for his company in a podcast interview in April 2018. This was a little over a week before Zuckerberg appeared at the first of the many congressional hearings in which lawmakers lambast him about Meta’s decisionmaking, and the proximity of these two events is surely no coincidence. As the timing suggests, the Board is a creature born of the techlash, and understanding this context sheds light on the particular public pressures that Meta was responding to in creating it. In 2018, Meta was busy trying to contain the damage to its reputation caused by revelations about Russian influence operations on its platforms during the 2016 US Presidential elections, the Cambridge Analytica scandal, and a growing general sentiment that tech platforms had become too big, too powerful, and too unaccountable. As pressure on the company continued to mount, Zuckerberg formally announced in a blog post in November 2018 that Meta would create an independent body that would hear user appeals against the company’s content moderation decisions, kicking off a process that culminated in the first Board members being announced in May 2020.

The final blueprint of the Board depicted a body intended to comprise forty members, who have a maximum of three three-year terms. (In practice, the Board has never had anywhere near 40 members—it has had a maximum of 23 members to date.) These members hear appeals from users

---


42 Charter, Art 1, §1; Bylaws, art 1, §1.4.

43 Charter, Art 1, §3.
or take referrals from Meta itself. The Board’s Charter states that the Board should “determine whether [Meta’s content enforcement decisions] are consistent with [Meta’s] content policies and values.” In doing so, the Board should “pay particular attention to the impact of removing content in light of human rights norms protecting free expression.”

Despite the name “Oversight Board,” which might be thought to imply sweeping oversight of Meta generally, the jurisdiction of the Board—the kinds of decisions and cases it is empowered to review—has always been extremely limited, and explicitly does not cover some of the most consequential parts of the company’s operations. Within the ambit of its authority, though, The Charter states that the resolution of each case (that is, the decision about the individual piece of content in question) is “binding” (although no enforcement mechanism is specified), and Meta should implement it promptly, both respect to the individual piece of content at issue in the case and also with respect to “identical content with parallel context.”

In every decision, the Board may also include a policy advisory statement making recommendations for Meta’s future policy development. However, Meta is not bound by any policy guidance the Board gives, nor is Meta bound to follow the Board’s advice if it refers a case to the Board requesting a general policy advisory statement. Thus, the Board’s input on policy “will be considered as a recommendation” and Meta’s only obligation is to publicly respond.

But the Board is not mere trompe-l’oeil or a façade without substance. Meta made an initial commitment of $130 million dollars to fund the Board through an independent trust, and contributed an additional $150 million to the trust in 2022. The resumes of the Board’s inaugural members are impressive, and Meta has since asked them to review some of its most high-profile and controversial decisions, starting with the decision of how to handle former President Donald Trump’s account in January 2021.

---

44 Charter, Art 2.
45 Charter, Art 2, §2.
46 Charter, Art 2, §2.
48 Charter, Art 4.
50 Id; Bylaws, Art 2, §2.3.
53 Nick Clegg, Referring Former President Trump’s Suspension From Facebook to the
B. The Objectives

Trying to understand the purpose of the Board only by looking at the ways it is explicitly described in its constituent documents and other public-facing materials does not get one very far. Meta and the Board’s websites both declare that the Board was created to “help” Meta “resolve some of the most difficult questions around freedom of expression online: what to take down, what to leave up and why.”\(^\text{54}\) This is accurate the most literal sense, but does not answer the deeper question about what kind of “help” the Board is intended to provide. The lack of engineers amongst the Board’s members makes it clear, for example, that the Board was not intended to help review the code of Meta’s automated content moderation tools. So what exactly is the deficiency in Meta’s decisionmaking that the Board was created to address? And how do the choices of the Board’s institutional designers go towards addressing them?

The Board’s design and context shows that the substantive decisionmaking of the Board is a relatively minor part of its intended benefits. Instead, the fundamental goal of the Board is to create legitimacy for Meta,\(^\text{55}\) in the sense of making Meta’s decisions become seen as “justified, appropriate, or otherwise deserving of support for reasons beyond fear of sanctions or mere hope for personal reward.”\(^\text{56}\)

For Noah Feldman, the law professor that first proposed the idea of a Facebook Supreme Court to executives at Meta,\(^\text{57}\) the Board did have at least a partially substantive purpose. Feldman originally conceived of the Board as a direct response to the threats to free speech in the social media age, and “the pressure that the platforms face to limit expression in order to satisfy engaged, committed advocacy groups.”\(^\text{58}\) Therefore, Feldman’s goal for the Board was explicitly substantive and values-laden, not merely procedural: the

---


\(^{55}\) Douek, supra note 37, at 18.

\(^{56}\) Richard H. Fallon, Legitimacy and the Constitution, 118 HARV. L. REV. 1787, 1787 (2004–2005); See also Tom R. Tyler, Procedural Justice, Legitimacy, and the Effective Rule of Law, 3 CRIME AND JUSTICE 283, 307 (2003) (“Legitimacy is the property that a rule or an authority has when others feel obligated to defer voluntarily.”).

\(^{57}\) Klonick, The Facebook Oversight Board, supra note 37, at 2449.

Board should be “an effective counterweight to censorship.” But the institutional design of the Board was intended to deliver these substantive goals in a way that reduced the decision costs to Meta, and the benefits of “legitimacy” pervade Feldman’s original memos arguing for the Board’s creation. The Board, he wrote, should become a “durable institution to deliver principled, reasoned decision-making that would be widely understood as legitimate.” The public-facing, judicial model of the Board was essential in order to “capture the legitimacy benefits of decisional independence.”

These procedural and legitimacy benefits were attractive to Meta. When Zuckerberg was first proposing and explaining his decision to set up the Board, he did refer to the importance of its role in protecting free expression. But he “also recognize[d] the reality of keeping people safe” and did not center the Board’s role in pushing back against censorship. Instead, when Zuckerberg set out to sell the Board to the public, he focused on the procedural aspects of its institutional design. He strongly emphasized its intended independence, writing that the Board would “prevent the concentration of too much decision-making within our teams” and “provide assurance that these decisions are made in the best interests of our community and not for commercial reasons.” Thus, for Zuckerberg, the Board was intended to outsource the decision-making, not necessarily to make it better or more speech protective. The charitable reading of this is that Zuckerberg wanted to benevolently relinquish power to ensure that justice was not only done, but also seen to be done. The more cynical take is that Zuckerberg wanted to throw content moderation’s hottest potatoes to other decisionmakers and get them off his hands.

Either way, there was interest alignment between Zuckerberg and other stakeholders. There have long been concerns about the astonishing power social media companies wield over global communications, concentrated in the hands of a few. In 2008, Jeffrey Rosen wrote in The New York Times about the example of the deputy general counsel of Google, Nicole Wong, who was nicknamed “The Decider” because of her control over what would and would not be allowed on some of the most precious real estate on the internet, suggesting that “Wong and her colleagues arguably have more influence over the contours of online expression than anyone else on the

59 Id. at 102.
60 Id.
61 Id. at 110.
63 Zuckerberg, supra note 41.
64 Id.
planet.”65 This raised the “increasingly urgent” question, said Rosen, “Can we trust a corporation to be good?”66 Nothing much had changed by 2012, when Rebecca MacKinnon wrote about the lack of traditional forms of legitimating constraints on the power of tech platforms and asked, again, “How do we make sure that people with power over our digital lives will not abuse that power?”67 As the companies in question only grew in size and importance these concerns became more pressing, and they reached new heights in the aftermath of the 2016 US presidential election.68 The fundamental problem for Meta in that moment was foreshadowed by Professor Tim Wu over a decade and a half before, in the context of one of Meta’s competitors: “To love Google, you have to be a little bit of a monarchist … [T]hey live and die on trust, and as soon as you lose trust in Google, it’s over for them.”69 In the techlash, people’s trust in platforms hit a nadir, and the platforms were looking for ways to get it back. The Board was the most unique and high-profile attempt, purporting to finally provide an answer for how to build back confidence in Meta’s corporate monarchy: separation of powers.

To be clear, the interest alignment only went so far. Lest one think that the Board was a genuine answer to critics’ longstanding fears of Meta’s unchecked power over speech, the Board’s limited remit, or “jurisdiction,” is a clear reminder that Meta’s goal in setting up the Board did not match its rhetoric. As Nick Clegg, Meta’s President of Global Affairs, wrote when welcoming the Board’s first members, “we have created and empowered a new group to exercise independent judgment over some of the most difficult and significant content decisions.”70 “Some” was right. The areas of decision-making that the Board actually had oversight over was, and remains, extremely limited.71 The Board does not have the power to review decisions Meta makes about accounts, advertising, events, groups, amplification or the platforms’ features and affordances, for example. When the Board was first

---

66 Id.
69 Rosen, supra note 66.
70 Clegg, supra note 42.
71 I wrote about this far too often during the process leading up to the Board’s creation. See, e.g., Douek, supra note 37, at 39–46; Douek, supra note 48; Evelyn Douek, The Facebook Oversight Board Should Review Trump’s Suspension, LAWFARE (Jan. 11, 2021), https://www.lawfareblog.com/facebook-oversight-board-should-review-trumps-suspension.
created, it did not even have the power to review cases in which Meta decided to leave content up as opposed to when it took content down. And while this has since been remedied, there are few signs the Board’s “jurisdiction” is set to expand any further.

Thus, in many ways, Meta’s creation of the Board resembles authoritarian governments’ use of courts. Authoritarian rulers maintain courts because courts can provide a patina of legitimacy, they can provide a measure of predictability which can be useful for attracting commercial investment in particular, and they allow for the deflection of controversy away from the ruling regime. But in order to get these benefits without too great a cost, such rulers often confine the scope of courts’ jurisdiction such that “[a] relatively independent judiciary may be preserved but simply excluded from domains significant to the authoritarian regime.” As a result, the actual trust garnered by such institution tends to be limited, and the mainstream initial reception to the Board reflected this skepticism. But in order to get any of the legitimacy and trust dividends that Meta hoped for from the Board, it needed to meaningfully relinquish some power. The cost/benefit balance that Meta struck in the Board’s design was opting for fairly robust independence mechanisms within an extremely limited jurisdiction.

There are of course many different kinds of legitimacy, and different theories about how to establish it, and the Board’s model was left unspecified and undertheorized. One mechanism through which the Board might bring

73 Douek, supra note 37, at 10.
77 Martin Shapiro, Courts in Authoritarian Regimes, in RULE BY LAW: THE POLITICS OF COURTS IN AUTHORITARIAN REGIMES 326, 331 (Tamir Moustafa & Tom Ginsburg eds., Cambridge University Press 2008).
80 See, e.g., Brenda Dvoskin, Expertise and Participation in the Facebook Oversight
legitimacy was by increasing the Meta’s accountability and transparency,81 in the sense that it would require Meta to provide at least a bare minimum amount of information about, and justification for, its decisions.82 This is a thin definition of accountability, but even so it is one that is sorely lacking across the content moderation industry.

Some also hoped the Board would build legitimacy by mitigating long-standing concerns over the democratic deficits of content moderation.83 This is similar to the goal of outsourcing power away from Meta as the sole decision-making, but is more specific. It is not only that such significant power should not lie solely in the hands of a single corporate entity, but that it should be in some meaningful way informed by broader public participation.84

But most fundamentally, the Board’s design is centered around creating legitimacy by replicating a form of speech governance that is “afforded a degree of presumptive legitimacy”85—that is, by acting like a court. This was,


81 See, e.g., Oreste Pollicino & Giovanni De Gregorio, Shedding Light on the Darkness of Content Moderation: The First Decisions of the Facebook Oversight Board, VERFASSUNGSBLOG (2021), https://verfassungsblog.de/fob-constitutionalism/ (hoping the Board will “increase transparency in content moderation, while making Facebook more accountable for its choices.”); David Wong & Luciano Floridi, Meta’s Oversight Board: A Review and Critical Assessment, MINDS & MACHINES 1 (Oct. 2022) (observing that the Board’s significant strengths include “its ability to enhance the transparency of content moderation decisions and processes”).


83 Klonick, The Facebook Oversight Board, supra note 37, at 2499 (stating that the Board “signifies a step towards empowering users by involving them in private platform governance”); Edward Lee, Virtual Governments Special Issue: Governing the Digital Space, 27 UCLA J.L. & TECH. 1, 5 (2022) (arguing that platforms “operate in the shadow of a democratic deficit”); 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.J. 939, 991 (2020) (stating that “Facebook and other major platforms could go much further in ensuring more structured multistakeholder participation in the development and revision of their content moderation rules.”).

84 What exactly that public participation should look like is not usually made clear, as discussed further below at infra XX.

of course, the initial framing that Feldman and Zuckerberg suggested for the Board, and it has had remarkable staying power, “successfully shift[ing] the discourse around the [Board] into a register that invited comparisons between Facebook and the legal systems of democratically founded states, rather than to classic forms of industry-led informal regulation and/or lobbying.”

The most significant feature of judicial decisionmaking that the Board’s institutional design replicates is the obligation to provide public rationales for its decisions. As Feldman argued in his original proposal, “[t]he key to making this private-courts approach work is recognizing that there is no magic-bullet solution to balancing competing values … . The advantage enjoyed by real-life constitutional courts is that they openly address difficult cases, and so derive credit and legitimacy from being principled.”

The duty to publish reasons might seem self-evident to lawyers accustomed to this practice, but was not inevitable. The Board could have been set up to provide independent but in camera review of Meta’s rules—many platforms have trust and safety advisory boards or similar bodies that play this kind of role. By contrast, the Board’s designers made a conscious choice to require public reasons, along with the directive that such decisions will have precedential effect. This gives the Board an important explanatory role, and reflects a hope that the Board will provide substantial and coherent public justifications for Meta’s otherwise opaque rules and decisionmaking.

The Board itself has, in appearance at least, leaned into this conception of its role. The Board’s decisions clearly ape the norms and forms of a judicial institution. The Board has stuck to an IRAC (Issue, Rule, Analysis, Conclusion) template in its decisions, familiar to every law student, and

---

86 Michael Veale et al., AI and Global Governance: Modalities, Rationales, Tensions, 19 ANN. REV. LAW. SOC. SCI. 14 (2023) (citations omitted); see also Kadri, supra note 37, at 171–74; Josh Cowls et al., Constitutional Metaphors: Facebook’s “Supreme Court” and the Legitimation of Platform Governance, NEW MEDIA & SOC. 15 (2022).

87 Bylaws arts 3.1.7; 3.2.

88 Global Feedback and Input on the Facebook Oversight Board for Content Decisions: Appendix, supra note 59, at 102.

89 Frederick Schauer, Giving Reasons, 47 STAN. L. REV. 633, 633 (1995) (“The practice of providing reasons for decisions has long been considered an essential aspect of legal culture.”).


91 Charter art 2, §2.

92 Douek, supra note 37, at 66.

93 Kadri, supra note 37, at 174 (giving examples of how Board members have embraced the judicial analogy).
applied it mechanically and faithfully in every case. It makes “findings,” cites “sources of authority,” summarizes parties’ “submissions,” and invokes the decisionmaking rubric of IHRL. It “overturns” or “upholds” Meta’s original enforcement actions, and has started releasing “summary decisions.” In all these ways, the Board is seeking to invoke the legitimacy afforded to legal decisionmaking.

Thus, the Board has many subsidiary goals: to make content moderation more speech protective; more independent; more transparent and accountable; more democratic; to provide public reasoning. But at its root, these goals are all in service of the Board’s overarching purpose: to establish legitimacy as a source of authority for important and difficult disputes about online speech rules.94

II. REVIEWING THE BOARD’S PERFORMANCE

According to the Board, it is doing a very good job indeed of achieving its goals. The Board’s inaugural members have completed their first three-year term, and every member that wanted to extend their tenure was renewed for a second.95 The Board’s Trustees, who oversee the process for approving a Board member’s reappointment,96 thus appear happy with the Board’s overall performance. The Board’s own assessment is also unequivocal. In the Board’s eyes, “early successes in holding Meta accountable demonstrate the Board’s viability and provide a self-regulatory framework for extending and improving our operations in the future.”97 Through its work, “the Board has

94 Douek, supra note 37, at 18 (“Legitimacy is central to the FOB experiment.”); Klonick, The Facebook Oversight Board, supra note 37, at 2427 (“Facebook’s creation of the Oversight Board is an investment in building user trust”); Monroe E Price & Joshua M. Price, Building Legitimacy in the Absence of the State: Reflections on the Facebook Oversight Board, 17 INT’L J. COMM. 3315, 3319 (2023) (“The aspiration is that the insertion of this new element—the Oversight Board—becomes a productive and effective attribute of legitimacy”); Chinmayi Arun, Facebook’s Faces, 135 HARV. L. REV. F. 236, 238 (2022) (“the Oversight Board was created to bring it more legitimacy”); Gilad Abiri & Sebastian Guidi, From a Network to a Dilemma: The Legitimacy of Social Media, 26 STAN. TECH. L. REV. 92, 115 (2022) (“The idea, in short, was that the Board would enhance Meta’s legitimacy”); Platform://Democracy: Perspectives on Platform Power, Public Values and the Potential of Social Media Councils 7 (2023) (arguing social media councils like the Board “can provide more legitimacy to the rules and algorithmic practices of platforms”); Haggart & Keller, supra note 81 (discussing the different ways the Board could establish legitimacy).


96 Charter, art 1, §3.

demonstrated” the importance of its role.98

This Part reviews the Board’s performance so far and whether these claims are justified. First, I look at the Board’s successes, which are not insubstantial. At a minimum, the Board has succeeded at being a functional institution in the most literal sense, by hearing cases, making decisions and issuing recommendations to Meta, and it has succeeded in avoiding the failure scenario most commonly predicted for it—it has not been captured by Meta and has asserted its independence in significant ways. The Board has also succeeded in some of the more ambitious goals people had for it—it has made Meta’s content moderation more accountable, transparent, and participatory, and has garnered public attention and respect for doing so. But the Board’s claims of success need to be qualified. The second half of this Part turns to the Board’s missed opportunities and the ways in which it has failed to live up to its potential—its lack of attention to developing a normative framework for thinking about private content moderation, its measurement of impact based on metrics that are poor proxies for material benefit, and its persistent avoidance in the hardest cases.

Overall, the story of this Part is ambiguous. The Board is perhaps much more successful than many might have imagined. But even if it did not fail in expected ways (being captured by Meta; becoming forgotten by content moderation stakeholders; having its decisions consistently overruled), it has failed to live up to its potential in other ways that have not received enough attention.

A. The Board’s Success Stories

When the Board was established, no one knew what to expect. It was an unprecedented experiment in governance for a private company—or, put another way, “a bit odd.”99 It was therefore impossible to predict what it might achieve. Would the Board become a model for the industry? Or would it fall apart and join the graveyard of other all-but-forgotten experiments in social media governance, like the time Facebook tried letting users vote on its rules?100 Only a few years in, a final verdict is still premature. But so far, it seems the Board is, in many ways, working. This subpart reviews some of the success stories of the Board. But each of these stories is also a story of

---

98 Tweet, @OVERSIGHTBOARD (Nov. 50, 2022), https://twitter.com/OversightBoard/status/1588513975188484099.
limitation—every achievement of the Board also reveals the constraints on what it can do or has done.

1. Fulfilling its Basic Functions

To start with, the Board has succeeded at performing its role in the most literal sense. The functional purpose of the Board is to hear cases, issue opinions and give Meta recommendations. Over the past three years the Board has done exactly that. This is more of a minimum condition for success than a dramatic achievement, but with any experimental institution there is no guarantee of even this form of success.

One way of reviewing an institution’s output is by looking at some headline statistics,\textsuperscript{101} and indeed the executive summary of the Board’s Annual Reports tend to focus on this kind of data in summarizing the Board’s performance. For example, a reader opening the Board’s first Annual Report will be informed that in its first year the Board received over 1.1 million appeals, published 20 decisions, overturned Meta’s decisions 70% of the time, made 86 policy recommendations, which Meta took steps towards implementing two-thirds of the time,\textsuperscript{102} and so on.\textsuperscript{103}

Numbers like these are clearly intended to tell a story: there is high demand for the Board (over a million appeals!), the Board is being tough on Meta (overturning so many of its decisions!), the number of recommendations it makes far outstrips the bare numbers of decisions it makes (86 vs. 20!), and its power is evident in how often Meta has committed to reform as a result of the Board’s work (two-thirds of the time!). The exact numbers have changed in subsequent years, but the overall gist, and the story they are intended to tell, has remained the same.

It is important to put this story in context, however. Pushing a little on these numbers reveals a more complicated picture than the one that the Board wants to portray. Let’s take a few illustrative examples.

First, the Board often points to the fact that many users have lodged appeals to the Board to indicate “enormous pent-up demand” for additional review mechanisms.\textsuperscript{104} The Board’s annual report for 2022 proudly proclaims that “[o]n average, the Board received a case every 24 seconds.”\textsuperscript{105}

The Board highlights this figure to suggest that there is “ongoing demand from users to appeal Meta’s content moderation decisions to an independent

\textsuperscript{101} The Statistics, 136 Harv. L. Rev. 500 (2022).
\textsuperscript{102} Oversight Brd., supra note 98, at 6–7.
\textsuperscript{103} The Board’s subsequent Annual Reports have continued this practice. See, e.g., Oversight Board 2022 Annual Report, supra note 20, at 8–9.
\textsuperscript{104} Oversight Brd., supra note 98, at 4.
\textsuperscript{105} Oversight Board 2022 Annual Report, supra note 20, at 30.
This figure of course does show that there are some users who remain dissatisfied with Meta’s rules and internal appeals processes, but by itself this number actually tells us little about the importance of the Board or the general public demand for such an institution. In reality, in context the number is actually astonishingly small. The Board received an average of 3,537 appeals per day in 2022, while Meta actioned 35,608,178 pieces of content a day in the last quarter of that period (a figure that doesn’t include the number of times Meta decided to leave content up after someone flagged it for review, which could also be appealed to the Board), and received 140,597 internal appeals every day. That is, users only appealed what would be significantly less than 0.0001% of Meta’s decisions to the Board, and only 2.5% of users who appealed decisions to Meta went the additional step of appealing to the Board. The level of demand for the Board looks very different when viewed in relative, rather than absolute, terms.

Even this figure lacks context. The Board’s citation of raw numbers does not indicate the nature or quality of these appeals. Content moderation experience suggests that these appeals would be a very noisy indicator of problems in Meta’s enforcement systems. Meta has previously stated, for example, that when it gave users the opportunity to provide more information to support their appeals, only 2% of the information people provided was useful. Thus, the simple fact that there are a lot of appeals does not in itself make the case that the Board is necessary to satisfy pent up and sincere demand for an additional appeals mechanism.

The Board is struggling to meet the demand that there is, however. This is not a dynamic unique to the Board—the scale of content moderation is a fundamental problem that any system of content moderation governance must reckon with, and suggests the limits of an approach based around ex post, individual appeals. But even accepting that the Board can only decide

---

106 Id. at 32.
107 Id. at 30.
108 Meta Quarterly Community Standards Enforcement Report [to be updated at time of publication].
109 Brian Fishman, Dual-Use Regulation: Managing Hate and Terrorism Online before and after Section 230 Reform, BROOKINGS (Mar. 14, 2023), https://www.brookings.edu/articles/dual-use-regulation-managing-hate-and-terrorism-online-before-and-after-section-230-reform/ (“One of the most dispiriting early lessons I learned at Meta was that user reports, paradoxically, are both critically important and wildly unreliable.”).
a limited fraction of the appeals it receives, the Board is underdelivering. We
can see this if we look at the unimpressive statistics about how the Board has
met the overwhelming demand for it. In total, the 20+ member Board issued
45 decisions in the first three years of operation.\textsuperscript{112} By comparison, in the
2021 term alone, the 9-member Supreme Court issued full decisions in 60
cases.\textsuperscript{113}

The paucity of issued opinions is not the consequence of a paucity of
funding. As Helfer and Land note, out of all the international human rights
tribunals, only the European Court of Human Rights has a budget that is
anywhere close to the size of the Board’s, yet the ECHR reviews more cases
by several orders of magnitude per year than the Board (36,000 vs. 20 in
2021).\textsuperscript{114} Board decisions are also significantly shorter and generally less
detailed than a typical appellate court decision. For people earning six-\figure
salaries,\textsuperscript{115} the Board’s output is incredibly low. Worse yet, these relatively
few and short decisions often arrive after the deadlines set for them in the
Board’s own bylaws.\textsuperscript{116} In a sign of the quasi-make-believe nature of the
Board, the fact that essentially every decision is now issued in breach of the
procedures laid out in the Board’s notionally constituent documents appears
to trouble no one involved in the project. The Board simply notes the reasons
why decisions are late (which include members taking leave and scheduling
challenges) in footnotes in the Board’s transparency reports.\textsuperscript{117}

While it is true that much of the impact of the Board will come through
its broader recommendations rather than its individual decisions, the Board
was purposefully set up in an adjudicative model, and deciding individual
cases is a core part of its role. As one tech reporter put it, “It’s true that the
board’s actions go beyond the three cases it decided [that quarter]. … At the
same time, I can’t be the only person to feel like the board is slacking.”\textsuperscript{118} So
even as the Board emphasizes overwhelming demand for its work, its output
suggests that it is struggling to meet that demand. Meta has implicitly

\textsuperscript{112} As at June 30, 2023.
\textsuperscript{113} The Statistics, supra note 102.
\textsuperscript{114} Laurence R. Helfer & Molly K Land, The Meta Oversight Board’s Human Rights
\textsuperscript{115} Allana Akhtar, Facebook is Reportedly Paying Its Oversight Board 6-Figure Salaries
Only for It to Tell the Company to Solve Its Own Problems, BUSINESS INSIDER (May 6, 2021),
\textsuperscript{116} Evelyn Douek & Tia Sewell, Meta’s Oversight Board Often Turns in Its Homework
\textsuperscript{117} See, e.g., Oversight Brd., Q1 2023 Quarterly Transparency Report 13 n 8 (Jun.
\textsuperscript{118} Casey Newton, The Oversight Board Spins Its Wheels, PLATFORMER (Oct. 20, 2022),
acknowledged the problem of low and slow work product from the Board, noting its hopes and expectations that the Board will “significantly increase its output and impact”\(^{119}\) by starting to take expedited reviews and issue summary decisions.

Although the Board could never review anything more than a tiny fraction of the number of appeals that users bring to it, there would appear to be another simple way to at least increase the baseline from the currently incredibly low output: appoint more Board members. The Board originally intended to be forty members,\(^{120}\) but in its three years, the Board has only appointed new Board members three times.\(^{121}\) In April 2023, it announced it would not be appointing many more: “While we originally expected the Board to reach 40 Members, three years of operations has shown us that, in practice, the optimal number of Members allowing for timely, regular, and effective deliberation and decision-making, is 26.”\(^{122}\) This very specific figure (twenty-six, not one more or less!) is offered without explanation, and no mention is made of the countervailing considerations, like the fact that the Board appears to be unable to manage its workload, or the impacts on the ability to secure representation from certain geographic regions. Again, the figures given in the Charter and bylaws for the composition of the Board appear to be seen by the Board as a suggestion only. Importantly, perhaps tellingly, at least one effect of this decision not to appoint more members is that the power of the original Board members is not diluted by new membership.

Another metric that is somewhat misleading is the Board’s statistics about how often it “overturns” Meta. The way this figure is presented suggests that the Board is being more confrontational with Meta than it really is. When the Board announces that it has “overturned” Meta, the suggestion is that the Board and Meta disagreed on the outcome of a case but the Board’s decision prevailed. However, in many instances in which the Board says it has “overturned” Meta, the Board and Meta do not disagree about the right outcome at all. This is because frequently, once the Board selects a case, Meta reviews the content in question and finds a mistake in its original enforcement

\(^{119}\) Meta Q1 2023 Quarterly Update on the Oversight Board 8 (2023).

\(^{120}\) Charter, Art 1, §1; Bylaws, art 1, §1.4.


action.\textsuperscript{123} It therefore agrees with the Board that the decision was wrong, and
reverses its enforcement decision. In these cases, though, the Board will still
announce it has “overturned” Meta’s original decision.\textsuperscript{124} To be clear: it’s not
that the Board isn’t serving a valuable function in these cases—the Board is
identifying errors that would have remained uncorrected if it weren’t for the
Board selecting the case for review, and highlighting the existence between
Meta’s policies on paper and how they are enforced. But identifying mistakes
is a different role from forcing Meta to change its normative judgments. The
former is more akin to auditing Meta’s systems rather than helping Meta
“resolve some of the most difficult questions around freedom of expression
online.”\textsuperscript{125} For the Board, focusing on a statistic about “overturn” rates
defined so broadly makes it look “tougher” on Meta and helps bolster its
public narrative of vigorous and independent oversight, which is how this
figure has often been understood.\textsuperscript{126}

What these few examples show is that there is no simple way of
measuring the Board’s work, and the statistics the Board gives about its
performance need to be viewed critically and in context. These figures are
selected to tell a particular story, and to show the Board in the most positive
light. There is nothing inherently wrong or unexpected about the Board or
Meta trying to make themselves look as good as possible, but it requires
approaching their statements with caution and with an understanding of the
picture the handpicked figures are intended to paint. And it underscores why
an in-depth study of the Board’s work is necessary before it can be properly
evaluated.

2. The Board’s Assertions of Independence

The Board has been successful in defining its identity as an independent
entity, and not merely one acting at the behest of its creator. A core feature
of the Board—indeed, perhaps the most novel aspect of it as an experiment
in corporate governance—is that it would exercise “independent
judgment”\textsuperscript{127} and not be beholden to Meta. But the self-regulatory nature of
the Board naturally led many people, particularly at first, to express concerns
about whether it could be sufficiently independent, given it was created and

\textsuperscript{123} Oversight Board 2022 Annual Report, supra note 20, at 34 (“Meta deemed its original
decision in 32 out of 50 cases shortlisted in 2022 [64\%] to have been incorrect”).
\textsuperscript{124} See, e.g., Case 2022-005-FB-UA (Mention of the Taliban in news reporting); 2022-
004-FB-UA (Colombian police cartoon case).
\textsuperscript{125} The Oversight Board, supra note 55; Oversight Board, supra note 55.
\textsuperscript{126} Wong & Floridi, supra note 82, at 8 (suggesting that a “notable strength of the OB is
its assertiveness, manifested in its willingness to overrule Meta”).
\textsuperscript{127} Oversight Board, supra note 55.
funded by the entity it is intended to oversee. However, the first few years of the Board’s existence have largely dispelled critiques of the Board model on the grounds that it would inevitably be captured and compromised by Meta. The Board has asserted a significant degree of independence in its first few years, including in ways that depart from the vision that Meta originally held for the Board and its role.

To be clear, Meta did not give the Board completely free rein. Indeed, the remit of the Board ensures it cannot function as a comprehensive oversight body, by only allowing it to review a small portion of Meta’s decisionmaking. Nevertheless, Meta did give the Board substantial independence by putting the Board’s funding into an irrevocable trust and placing the power to remove Board members exclusively in the hands of independent trustees. This of course was not selfless. Meta could not get the legitimacy dividends it wanted without such independence. For this reason, within its limited remit, the mechanisms protecting the Board’s independence are robust. This has protected the Board’s decisional independence, but has also given the Board the opportunity to depart from the precise plans Meta had for it.

Perhaps most consequentially, the Board’s recommendations have been broader than originally envisioned. Meta has remarked that “[t]he size and scope of the board’s recommendations go beyond the policy guidance that we first anticipated when we set up the board, and several require multi-month or multi-year investments.” The recommendations have, amongst other things, targeted the systems underlying Meta’s enforcement mechanisms (“Ensure people can appeal decisions taken by automated systems to human review” and “Inform people when automation is used to take enforcement action against their content”), asked Meta to submit to further independent scrutiny and assessment (“Engage an independent entity not associated with either side of the Israeli-Palestinian conflict to conduct a thorough examination”) and demanded further transparency around

---


130 Oversight Board Bylaws, Section 2.1.2.


132 2020-004-IG-UA (Breast Cancer Symptoms & Nudity Case).

133 *Id.*

134 2021-009-FB-UA (Al Jazeera Post on Israel & Palestine Case).
sensitive matters (“Formalize a transparent process on how it receives and responds to all government requests for content removal”\textsuperscript{135}). Such recommendations require more dramatic restructuring and resources to implement than the simple policy recommendations or rule-changes it seems Meta had intended.

The Board has also asserted the power to interpret its own founding documents and the procedures they set out. It has developed what might be described as a mootness doctrine that requires Meta to submit to review (and thus have to answer questions and receive recommendations) even in cases where Meta admits a mistake in the individual case in question. In one of its first cases, Meta argued that the Board should decline to take the case because it had already restored the post in question, agreeing that it had made an error in originally removing the content.\textsuperscript{136} But the Board rejected that argument, saying Meta had misinterpreted the Board’s Charter and that “[f]or Facebook to correct errors the Board brings to its attention and thereby exclude cases from review would integrate the Board inappropriately to Facebook’s internal process and undermine the Board’s independence.”\textsuperscript{137} The Board’s capacity to reject Meta’s interpretation of the Charter that \textit{Meta itself wrote} illustrates a kind of independence and aggressiveness that allows the Board to, within limits, define its role for itself.

The Board’s centering of IHRL as the authoritative body of norms governing its (and therefore Meta’s) decisionmaking is also an expression of its independence. The Board’s Charter only requires the Board to “pay particular attention” to “human rights norms protecting free expression,”\textsuperscript{138} and does not reference IHRL directly at all. The Bylaws make reference to “international human rights principles” and “standards,”\textsuperscript{139} but do not require the Board to apply IHRL directly. Nevertheless, the Board has elevated IHRL as its primary source of authority, citing it in every decision,\textsuperscript{140} and even overturning one of Meta’s decisions because it found that the company’s own community standards did not comply with IHRL.\textsuperscript{141} That is, the Board overruled the authority under which it was created based on an external body of norms.

\textsuperscript{135} \textit{Id.}

\textsuperscript{136} 2020-004-IG-UA (Breast Cancer Symptoms & Nudity Case).

\textsuperscript{137} \textit{Id} (emphasis in original).

\textsuperscript{138} Charter, Art 2, §2.

\textsuperscript{139} Bylaws, Art 2, §2.3.2; Art 1, §1.4.4.


\textsuperscript{141} 2021-004-FB-UA (Navalny Protests Case).
Finally, the Board has engaged in a perhaps unanticipated amount of what might be described as “Oversight Board Overspeech,” following Josh Chafetz’s description of Congress’ use of its oversight mechanisms to communicate with the broader public as “Congressional Overspeech.”142 The Board has actively courted public attention. Its co-Chairs wrote a New York Times op-ed announcing its creation,143 it has an Instagram account,144 a LinkedIn account,145 and a very active Twitter account that retweets positive coverage of the Board,146 or sends out cautionary warnings to Meta in high-stakes moments such as the invasion of Ukraine.147 The Board got press coverage for issuing “a strong reprimand” to Meta for giving misleading answers in one case,148 and announced a meeting with a Facebook whistleblower to get more information (although never releasing information about what was said at the meeting).149 Its members constantly appear at public events, turning up to such events as UN General Assembly,150 the Aspen Ideas Festival,151 the Paris Peace Forum,152 and annual Austin festival SXSW.153 They give interviews to journalists extolling their work for feature stories.154

---

142 Josh Chafetz, Congressional Overspeech, 89 FORDHAM L. REV. 529, 536 (2020).
143 Catalina Botero-Marino et al., Opinion | 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.
144 https://www.instagram.com/oversightboard/
145 https://www.linkedin.com/company/64981442/admin/
146 https://twitter.com/OversightBoard
147 Oversight Brd., Tweet, TWITTER, https://twitter.com/OversightBoard/status/1499045053142081543 (last visited Dec. 12, 2022) (“The Oversight Board is closely following the situation in Ukraine and how Meta is responding.”).
154 Steven Levy, Inside Meta’s Oversight Board: 2 Years of Pushing Limits, WIRED
communications with social media platforms, two Board members wrote opinion pieces in mainstream outlets explaining how the Board had dealt with a case raising the same issues (and recommended a better solution). The Board has a newsletter, promoting its activities, and releases quarterly transparency reports extolling its achievements in getting Meta to enact reforms.

Through these and similar activities, the Board is making an argument about its purpose and utility. In academic work, one of the Board’s original co-Chairs wrote that in traditional legal systems “the practice of constitutional law is the practice of persuading diverse citizens to share the priorities of the adjudicator.” Professor Jamal Greene clearly believes persuasion to be an important part of governance, and this is in keeping with the Board’s public relations efforts. But the Board is approaching this task in ways that are not typical of what one might expect from an august “Supreme Court”-like body. Courts do not typically have Twitter accounts that retweet compliments of them, or use the royal “we” to boast that “the results we obtained so far show we are making progress.” Board representatives rarely (if ever) prominently discuss the obvious shortcomings of its institutional design that hamstring its ability to be a comprehensive form of oversight. The Board has become an activist institution, aggressively promoting its own virtues, and acting sometimes more like an influencer than a governing body. It insists in its publications, through its members, and through its social media presence, that it is performing well and should be at the center of discourse about content moderation governance.

3. Prompting Reforms to Meta’s Systems

The Board has been very aggressive in exercising its power under the Charter to provide optional policy advisory statements in addition to individual decisions in each case, giving multiple such recommendations

---


157 Tweet, @OVERSIGHTBOARD (Nov. 14, 2022), https://twitter.com/OversightBoard/status/1587553721826906113.

in every decision. Critics had suggested that the Board could only have limited impact on Meta because the format of individual appeals meant that the Board would only review a small fraction of Meta’s decisions. The Board’s practice of issuing sweeping recommendations that target Meta’s underlying systems has been its answer to this critique.

It has emphasized the importance of this policy guidance to its work as an institution, insisting its “top priority” is to make sure that Meta implements the Board’s recommendations. The Board issues quarterly transparency reports tracking the comprehensiveness of Meta’s responses to every single recommendation the Board has made. Notwithstanding the fact that the “non-binding” status of recommendations means that the Board has even less authority to require Meta implement them than it does to require Meta to comply with its individual case decisions, the Board itself says that Meta has made progress in implementing the vast majority of reforms the Board has suggested.

In reality, the situation is more complex than the Board describes. Although the headline figures are impressive (Meta has only declined to implement 42 out of 214 recommendations!), there is a dramatic variance in the importance of the reforms that the Board has been responsible for precipitating. Some of the reforms the Board has pushed Meta to make have been in response to fairly obvious problems. That the Board’s push was necessary to make the company adopt them speaks more to the Meta’s negligence rather than the Board’s vigilance. For example, the Board recommended that Meta translate its Community Standards into Punjabi, one of the most-spoken languages in the world. Such a recommendation hardly requires expertise in freedom of expression to recognize as important, but the Board highlighting the issue does appear to have made Meta act more quickly than it would have otherwise. In another case, the Board found that one of Meta’s internal policies wasn’t being enforced because it was lost for three years. Again, it is hardly a profound insight that this is not best practice. The Board has an obvious and positive impact in cases like these, but perhaps

---

159 See, e.g., Klonick, The Facebook Oversight Board, supra note 37, at 1490.
160 Oversight Brd., supra note 98, at 60.
162 Oversight Brd., supra note 118, at 18.
163 2021-003-FB-UA (Modi & Indian Sikhs Case); Case on Punjabi Concern over the RSS in India, META TRANSPARENCY CENTER (Jul. 13, 2022), https://transparency.fb.com/oversight/oversight-board-cases/punjabi-concern-over-the-rss-in-india/.
164 2021-006-IG-UA (PKK Founder Abdullah Öcalan Case).
not the kind of impact that should really require years and $280 million to bring about. In such cases, the Board is playing the role of a glorified auditor rather than an appellate review body. Identifying points of failure can be important, but is not necessarily the original vision for the institution.

Other reforms sound good, but their impacts are unclear. For example, a common recommendation in the Board’s decisions is that Meta should add further language and clarity to its various Community Standards.\textsuperscript{165} The Board asserts that these recommendations are designed to further the IHRL principle of “legality” (that rules should be clear and accessible), but whether they actually make a difference to users or make content moderation any more predictable and consistent is an open question.\textsuperscript{166} As I will return to below, these kinds of recommendations evince a formalism in the Board’s thinking that reflects legalistic intuitions about the nature of rule-making rather than empirically demonstrated impact.

But there are certain reforms that that the Board has brought about that indisputably represent meaningful exertions of authority, too. For example, after long-standing concerns about Meta’s enforcement practices in Israel and Palestine and potential disparate impacts on different populations,\textsuperscript{167} the Board prompted Meta to conduct and release a human rights impact assessment of its operations in the region which concluded that unintentional bias had led to adverse impacts on Palestinian and Arabic speaking users.\textsuperscript{168} Meta also committed, in response to Board recommendations, to begin notifying users if their content is removed due to an extra-request from a government actor, and to include information in its transparency reporting about how many such requests it receives.\textsuperscript{169} Both of these initiatives are examples of reforms in areas that civil society had long been raising concerns about for years without getting traction.\textsuperscript{170} This suggests that the Board has

\begin{itemize}
    \item\textsuperscript{165} See, e.g., OVERSIGHT BOARD Q1 2022 Quarterly Transparency Report.Pdf 21–22.
    \item\textsuperscript{166} For an argument that these kinds of recommendations are futile, see Evelyn Douek, \textit{The Siren Call of Content Moderation Formalism} (Aug. 2022).
    \item\textsuperscript{169} Case Regarding the Support of Abdullah Öcalan, Founder of the PKK, META TRANSPARENCY CENTER (Jul. 13, 2022), https://transparency.fb.com/oversight/oversight-board-cases/support-of-abdullah-ocalan-founder-of-the-pkk.
    \item\textsuperscript{170} See, e.g., Open Letter to Facebook, Twitter, and YouTube: Stop Silencing Critical Voices from the Middle East and North Africa, ACCESS NOW (Dec. 17, 2020), https://www.accesnow.org/press-release/facebook-twitter-youtube-stop-silencing-critical-voices-mena/ (“Palestinian activists and social media users have been campaigning since
some meaningful capacity to influence Meta that other outsiders lack.

The Board’s recommendations have also become more sophisticated and targeted over time. One of the biggest areas of improvement for the Board over the course of its first term has been the shift in its thinking from a highly individualistic analysis of the issues before it to more systemic thinking. That is to say, the Board has moved from very case-specific, fact-intensive decision-making,171 to decisions that are more focused on Meta’s underlying systems for content moderation enforcement and platform design choices.172 Early decisions turned on matters like comparing slightly different translations of individual posts,173 or different inferences about the poster’s individual intent.174 The Board’s later decisions move away from such fact-specific analysis, acknowledging the impossibility of getting every decision right at scale, and looking for discriminatory patterns of enforcement,175 and using individual errors to identify “systemic breakdowns” requiring reform.176 Because of the trade-offs and complexities created by the volume and speed of content moderation, this move from an individualistic to a more systems-thinking-based approach is necessary to address the system design choices that matter at scale.177 The shift in the Board’s thinking is marked, shows growth in members’ understanding of the systems they are overseeing, and increases their potential for impact.

There have also been some significant losses for the Board, however, including (or perhaps especially) in high-profile cases. Meta refused to take up in any meaningful way one of the most consequential recommendations in the Trump Suspension case, to “review its potential role in the election fraud narrative that sparked violence in the United States on January 6, 2021

2016 to raise awareness around social media companies’ censorial practices.”); Scott Craig & Emma Llansó, Pressuring Platforms to Censor Content is Wrong Approach to Combating Terrorism, CENTER FOR DEMOCRACY AND TECHNOLOGY (Nov. 5, 2015), https://cdt.org/insights/pressuring-platforms-to-censor-content-is-wrong-approach-to-combating-terrorism/ (“Company transparency reporting can also help to provide more information about content removal requests from governments.”).

171 For a critique of the Board’s early decisions along these lines, see Evelyn Douek, The Facebook Oversight Board’s First Decisions: Ambitious, and Perhaps Impractical, LAWFARE (Jan. 28, 2021), https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical.

172 See, e.g., discussion of importance of systems and design in 2021-012-FB-UA (Indigenous Art and Kamloops School Case) and 2022-004-FB-UA (Colombian police cartoon case).

173 See, e.g., 2020-002-FB-UA (Myanmar Hate Speech Case).

174 See, e.g., 2021-005-FB-UA (Armenian Hate Speech Case).

175 See, e.g., 2022-003-IG-UA (Reclaiming Arabic Words Case).

176 See, e.g., 2022-001-FB-UA (Knin Cartoon Case).

177 Douek, supra note 112.
and report on its findings.”\textsuperscript{178} It similarly did not adopt the Board’s recommendation to provide a specific transparency report about Community Standards enforcement during the COVID-19 pandemic.\textsuperscript{179} And it withdrew a request for policy advice concerning content moderation issues related to Russia’s ongoing war with Ukraine, against the Board’s wishes.\textsuperscript{180} That is, in three of the most controversial and consequential content moderation subject areas in the last few years, the limits of the Board’s power over Meta were on stark display. In other less visible cases, Meta “repeatedly push[es] back the deadline for implementation” and the Board is unable to do anything than highlight Meta’s slow-walking.\textsuperscript{181}

And once again, it is important to be aware of the underlying incentives at play when interpreting the data that the Board and Meta release. Both organizations have good reason to paint as glowing a picture of the Board’s accomplishments as they can, in order to try convince outsiders of the benefits of the Board as an institution and reap the legitimacy dividends. This is why it is important to remain skeptical about Meta’s claims about the Board’s impacts. It is impossible to tell how many of the initiatives Meta says are a response to the Board’s recommendations would not have happened without it. Some of the changes—such as Meta launching a Transparency Center—seem likely to have occurred regardless.\textsuperscript{182} Moreover, it is not always obvious what it means when the company says it is implementing the Board’s recommendations in part or in full. Originally, Meta also counted itself as “committing to action” in response to the Board’s recommendations when its only response was to point to actions it was already taking. After I criticized this as inflating the impact of the Board,\textsuperscript{183} Meta implicitly acknowledged this was misleading by introducing a new category of response, “work Meta already does.” But Meta’s categorization of its responses often remains very generous. For example, when the Board

\begin{footnotesize}
\begin{enumerate}
\item\textsuperscript{178} \textit{Case on Former President Trump’s Suspension from Facebook}, META TRANSPARENCY CENTER (Jul. 13, 2022), https://transparency.fb.com/oversight/oversight-board-cases/former-president-trump-suspension-from-facebook.
\item\textsuperscript{181} \textit{Oversight Board 2022 Annual Report}, supra note 20, at 24.
\item\textsuperscript{182} Douek, supra note 180.
\item\textsuperscript{183} Id.
\end{enumerate}
\end{footnotesize}
recommended that Meta “increase its investments in digital literacy programs.” Meta classified itself as implementing the Board’s recommendation “in Full” because it was committing to “continue to globally roll-out” its existing programs—-not exactly the “increase” in investment the Board had recommended.

As I will discuss further below, assessing the tangible impact of the reforms that the Board has brought about is a much trickier—and much more important—question than the mere fact that the Board has prompted Meta to initiate reforms. But the basic fact remains that the Board, an institution with no formal legal power, has succeeded in making one of the most powerful corporations in the world change the way it operates.

4. Making Meta More Transparent and Accountable

The Board has brought some (albeit limited) forms of transparency and accountability to Meta’s content moderation which, despite being perhaps the most pervasive system of speech regulation in history, was previously almost entirely opaque. When the Board selects a case, it sends Meta a series of questions, and Meta has answered the vast majority. The Board sometimes includes parts of these answers in its decisions and through this process the Board has revealed new information in a range of areas, from how Meta’s database of images that get automatically removed across its platforms works, to the quality assurance systems the platform has in place, to the way the company makes allowances for newsworthy content in moments of political protest. Thus, the Board has imposed accountability at least in the thin sense of requiring Meta to explain itself. The Board has also imposed accountability through third parties, by, for example, prompting Meta to undertake a human rights impact assessment in Palestine. In these ways, the Board has helped shed light on systems that were previously completely dark to outsiders, and forced Meta to explain (and often rethink) its decisions.

---


185 See infra Part II.B.2.

186 _Oversight Board 2022 Annual Report_, supra note 20, at 38 (reporting that Meta answered 86% of the Board’s questions fully, and that the answers were increasingly comprehensive).

187 2022-004-FB-UA (Colombian police cartoon case).

188 See, e.g., 2022-005-FB-UA (Mention of the Taliban in news reporting case); Policy advisory opinion on Meta’s cross-check program (2022).


190 This definition of accountability draws on, e.g., Bamberger, supra note 83, at 404; Freeman, supra note 83, at 664; Mashaw, supra note 83, at 117.

191 Allison-Hope et al., supra note 169.
But there have also been real limits to how much accountability the Board has been able to impose, not least the fact that the Board still has jurisdiction over only the tiniest fraction of decisions that Meta makes.\textsuperscript{192} Meta can also refuse to answer the Board’s questions for any of the broad reasons given in the Bylaws (such as if Meta “determines that the information is not reasonably required”\textsuperscript{193}) and the Board cannot contest Meta’s determinations or compel it to disclose anything it doesn’t want to.\textsuperscript{194} In some cases, it has later become clear that Meta was withholding relevant information or selectively disclosing information to the Board in ways that was misleading.\textsuperscript{195} Meta has also in some cases rejected Board recommendations that would have led to greater accountability, such as investigations into its content moderation during the pandemic or in the lead up to January 6,\textsuperscript{196} and has simply withdrawn cases from the Board’s review on sensitive topics, with little explanation.\textsuperscript{197}

Even in these cases though, the Board has provided a focal point for other stakeholders’ demands for accountability. For example, lawmakers have pointed to Meta’s refusal to implement recommendations from “its own Oversight Board” as a sign of the company’s failures.\textsuperscript{198} They have also used the Board’s decisions as moments of leverage to encourage Meta to enact

\textsuperscript{192} See Douek, supra note 48 The Board’s jurisdiction has been marginally expanded since that piece was written, to include cases where individuals are appealing for content to be taken down, but the overall point still stands.

\textsuperscript{193} Bylaws art 2, §2.2.2.

\textsuperscript{194} Arun, supra note 95, at 261–62.


\textsuperscript{196} Douek, supra note 196; Douek, supra note 180.

\textsuperscript{197} Meta Withdraws a Policy Advisory Opinion Request Related to Russia’s Invasion of Ukraine, supra note 181.

reforms.199 And the fact that Meta had mislead the Board on its so-called X-Check program for high-profile users was prominently reported on,200 which was no doubt a significant reason Meta later referred the X-Check program to the Board.201

Thus, the scope of accountability that the Board can impose is limited (only a fraction of the kinds of decisions Meta makes), the kind of accountability thin (simply the requirement for Meta to explain itself), and the consequences for transgression slight (a public rebuke). That said, the Board has proven even this form of accountability has value. The Board has forced Meta to provide information and justification for decisions that previously went entirely unexamined, and helped create further levers for stakeholders to pressure the company. This has positive externalities for broader debates about content moderation, as the complexities and trade-offs inherent in running a system of speech regulation at the speed and size of a major platform are opened up for public examination. To the extent that the Board’s review process often reveals previously undiscovered or unacknowledged mistakes on Meta’s behalf, which it then corrects,202 this also shows a form of accountability and the benefits of Meta being forced to explain itself, even to a minimal degree.203

5. Making Content Moderation More Democratic

For some, the Board represented an opportunity to make “digital spaces more democratic,” and “represents one of the first attempts to open up the decision-making of a commercial platform to the ‘outside.’”204 But what does it mean to make content moderation “more democratic”? While a decision will generally be considered “democratically responsive to the extent it

200 Schechner, supra note 196.
202 See, e.g., Oversight Brd., supra note 118, at 9 (“It is noted that Meta found its original decision to have been incorrect in 63% of cases the Board shortlisted ... The Board continues to raise with Meta the questions this poses for the accuracy of the company’s content moderation and the appeals process the company applies before cases reach the Board.”
203 For further exploration of this point, see Glen Staszewski, Reason-Giving and Accountability, 93 MINN. L. REV. 1253, 1280 (2009) (“Because public officials must provide public-regarding justifications for their decisions, other participants in the process have incentives to articulate their claims in public-regarding terms as well. As a result, relatively selfish policy options may be discarded.”).
204 Platform, supra note 95, at 6.
reflects and expresses the popular will," the way that the popular will should be distilled and translated into specific content moderation decisions is usually left under-specified. It is impossible to say whether the Board has made content moderation more democratic without a more particularized conception of what that would mean. But it is at least true that it has provided an avenue for a broader set of stakeholders to comment on Meta’s rules and decisions.

The Board has broadly interpreted its discretion to “request and receive information from a global pool of outside subject-matter experts” to allow it to issue a call for public comment at the time it announces that it has taken a new case. The Board has called this process “crucial” for achieving its goals of improving Meta’s processes and stated that “[o]n numerous occasions, public comments have shaped our decisions and our recommendations to Meta.” Through this mechanism, then, the Board has created an avenue for anyone and everyone to express their views on Meta’s decisions. To the extent that calls for democratization simply reflect the desire for broader participation in content moderation policy and decisions, the Board could, for this reason, be considered successful. Indeed, as noted earlier, a wide set of stakeholders—including leading civil society organizations, academics, and politicians—have taken advantage of this process. And it is not just elites—the broader public has sometimes also engaged with the Board in this way. In the Trump Suspension case people seized the opportunity and the Board received 9,666 public comments, but this remains a significant outlier in terms of volume of engagement.

While the Board is therefore creating a venue for more people than ever to have their views on specific content moderation issues heard, the creation of a public comment process also is not inherently or necessarily a pro-democratic force. The long-standing concerns about skewed participation in public comment processes that favor well-resourced or powerful interest groups in the context of the administrative state apply equally to the

206 For an important discussion of this point, see Brenda Dvoskin, Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance, VILLANOVA L. REV. (2023) (arguing that “prevailing assumptions about civil society participation in content moderation are wrong”); see also Rachel Griffin, Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability, 14 TRANSNT’L LEGAL THEORY 46, 69–73 (2023).
207 Bylaws art 3.1.4.
208 Oversight Brd., supra note 98, at 51.
209 2021-001-FB-FBR (Trump Suspension Case).
210 Mendelson, supra note 206, at 1357–58 (summarizing the literature on these concerns).
Board’s process, in which wealthier and English-speaking regions have comprised the majority of comments.

There is also the question of what the Board does with these comments once it receives them. Indeed, whether public comments have any affect (let alone a democratic one) on the Board’s work is hard to tell, especially from the outside. In some cases, public comments appear to have had an impact. In the Gender identity and nudity cases, for example, the Board made special note of the high level of public comment, especially from the minority community likely to be particularly affected by the decision (in that case, trans and non-binary people). It was the first time the Board described the way public comments had affected its deliberation in detail, noting that it takes “comments seriously as part of its deliberations,” while understanding that “these comments may not be representative of global opinion.” In that case, however, the Board found the public comments’ evidence of the disparate impact of Meta’s policies persuasive, and this clearly informed the Board’s decision to recommend that Meta reform its rules.

References to particular public comments seem to be increasing in the Board’s more recent cases, but overall the Gender identity and nudity cases are an outlier as a rare instance where the public comment played a significant role in the Board’s explicit reasoning—the vast majority of the time the Board does not engage with public comments beyond summarizing them at a very high level. Even in the Trump Suspension case, where the Board received nearly ten thousand comments, it is not clear whether these comments affected the Board’s decision in any way. The Board did not make reference to any such impact, beyond noting that it was “grateful for the many thoughtful and engaged public comments that it received.” (It also did not mention that it probably received many less thoughtful comments, too.)

Nor is it clear how much weight the Board gives public comments compared with the other forms of input the Board solicits. Every decision has a note at the end that the Board commissioned independent research from, most often, the University of Gothenburg. The Board has never revealed the

---

211 Dvoskin, supra note 81, at 9.
213 Mendelson, supra note 206, at 1359 (making the same point in the context of administrative agencies).
214 2022-009-IG-UA and 2022-010-IG-UA (Gender identity and nudity cases).
215 Id.
216 As Brenda Dvoskin has summarized it, “Despite the Board’s interest in stakeholder engagement initiatives, the view from nowhere has, so far, prevailed in the Board’s deployment of the IHRL project”: Brenda Dvoskin, Expert Governance of Online Speech, 64 HARV. INT’L L.J. 85, 128 (2023).
217 Id.
substance of this research or referenced it in a decision. In some cases, it organizes “virtual roundtable[s]” with advocacy groups, presumably hand-picked.\textsuperscript{218} It has also cited to a single academic article, once.\textsuperscript{219} But presumably (hopefully) the Board’s work is also informed by other academic research, even if not referenced in its decisions.

In sum, while it is clear that the Board has created new ways for people to express opinions about Meta’s rules, how this has altered the power dynamics of decision-making—and whether it has done so for the better—is much more ambiguous. This is not, of course, a problem that is unique to the Board. Many governance regimes face the same challenge, and there is no single model of how best to incorporate public input.\textsuperscript{220} But the Board could and should be much more transparent about how it makes use of public comments,\textsuperscript{221} as well as other sources of input, lest ultimately it undermine faith in its own public responsiveness. The Board’s public comment process should not be “to public participation as Japanese Kabuki theater is to human passions—a highly stylized process for displaying in a formal way the essence of something which in real life takes place in other venues.”\textsuperscript{222} The Board has acknowledged the limits of its own democratic credentials, as a hand-picked group of experts, by opening up its cases for public comment, but it should be more consistent about showing that this consultation is genuine and not merely performative.

6. Legitimacy Building

The Board’s perhaps most important achievement is the way it has managed to transform the tenor of public engagement with it from highly critical to often accepting, even respectful. I have already noted the media reporting, the extensive engagement with its public comment process, and its inclusion in sets normally reserved for formal legal authorities such as legal databases and case notes.\textsuperscript{223}

The academic literature on the Board also takes the institution remarkably seriously. There are numerous articles analyzing its decisions and use of IHRL.\textsuperscript{224} Scholars (and Board members) have suggested that, through “norm

\textsuperscript{218} See, e.g., 2021-016-FB-FBR (Journalism on Sexual Violence Case).
\textsuperscript{219} 2021-004-FB-UA (Navalny Protests Case).
\textsuperscript{221} Mendelson, supra note 206, at 1380.
\textsuperscript{222} E. Donald Elliott, Re-Inventing Rulemaking, 41 DUKE L. J. 1490, 1492 (Jun. 1992).
\textsuperscript{223} See Introduction.
\textsuperscript{224} See, e.g., Barata, supra note 141; Note: Case Decision 2021-004-FB-UA Oversight Board Finds a Facebook’s Rule Application Violates International Human Rights Law,
diffusion,” the Board could directly influence the jurisprudence of human rights bodies, and as well as for national courts and regulators. It has been called a source of “transnational constitutional advice” and “remarkable work product,” and a body which could signal a “new wave of transnational hybrid adjudication.” Newton and Martha Minow have pointed to the Board as a salutary guide for effective self-regulation that the industry more broadly should learn from. David Fontana and David Schleicher have suggested that the Board represents an emerging form of governance that “could be commercially valuable” for many other private firms, far beyond the social media industry, to replicate. Laurence Helfer and Molly Land have compared the Board to an international human rights tribunal and argued that the Board can, and already has, “serve[d] as an important check on Meta and to significantly advance the promotion and protection of rights online.” Jonathan Zittrain has cited the Board as an experiment of the kind necessary to bring greater process and legitimacy to digital governance. Chinmayi Arun has noted the Board’s limitations but concludes that the Board is meaningfully independent, “highly influential” and plays a “very significant role” in bringing accountability to Meta. The Board gets lots of attention from the blogosphere, and still one influential commentator has

supra note 28; Pollicino & Gregorio, supra note 82; Wong & Floridi, supra note 82.


227 Miloš & Pelić, supra note 227.


230 Fontana & Schleicher, supra note 100, at 10.

231 Helfer & Land, supra note 115, at manuscript at 64.


233 Arun, supra note 95, at 262–63.

234 See, e.g., Alexa Koenig, Meta’s Oversight Board Recommends Major Advance in International Accountability, JUST SECURITY (Jun. 22, 2023),
declared that the Board “has not got nearly the credit it deserves.”\textsuperscript{235} I myself have added and am right now adding to the pile of articles about the body.\textsuperscript{236} The literature is, as they say, burgeoning.

Naturally, not all reviews of the Board’s work are positive. But all influential institutions are of course subject to critique, and the tenor of even many of the critical assessments of the Board have changed from being downright dismissive to often being serious and sincere.\textsuperscript{237}

Legitimacy is hard to define, and even harder to measure. Still, this kind of serious engagement from a wide diversity of stakeholders surely suggests that the institution is regarded with some respect—at least enough to make the cost of engagement worthwhile.\textsuperscript{238} For a fledging institution, created by a corporation that was itself at the peak of a legitimacy crisis of its own, this is a somewhat remarkable achievement.

\textbf{B. The Board’s Formalistic Approach}

The Board’s establishment of legitimacy is made more remarkable by the fact that it has been neglecting some core aspects of its role. While the above noted many limits on the Board’s success stories due to its institutional design, this subpart explores the ways that the Board has under-delivered, even taking its institutional constraints as a given. While the structural critiques of the Board are important, it is vital to appreciate the Board’s


\textsuperscript{236} I’ll let you be the judge of whether that bolsters its legitimacy or not.


\textsuperscript{238} Price & Price, \textit{supra} note 95, at 3322.
missed opportunities within its current design because ultimately, as Part III below will argue, these failures are the result of institutional incentives and political dynamics that apply much more broadly than to the Board alone. That is, if the nature and source of these failures is not adequately appreciated, they are likely to be replicated by other content moderation governance institutions, even if they do not have the same structural weaknesses.

This subpart focuses on three key ways in which the Board has failed to fulfill its potential. First, the Board has neglected the task of developing a coherent normative framework for its work. Its decisions mimic the forms of judicial opinion-writing, but the Board’s reason-giving has been relatively superficial and failed to engage with the most difficult questions that content moderation governance raises. Second, the Board has not clearly defined its goals, and the metrics that it measures as indicators of its impact are poor proxies for material benefits. It is thus impossible to tell the real-world impacts of the Board’s work and therefore meaningful accountability—for either Meta or the Board—remains illusory. Finally, in the most difficult cases, the Board often avoids giving a clear answer, refusing to take a clear stance on some of the most controversial questions that content moderation raises. The result is an institution that observes many of the formalities of good governance, but pays less attention to its substantive outcomes.

1. Rulings Without Reason

The Board’s function as a reason-giving institution is core to its mission—its job is not only to decide cases, but to publicly explain its decisions.239 This reflects the aspiration that “[t]he principle of publicity—or public reason-giving—allows for notice, guidance, and prediction, all essential to the rule of law.”240 But despite the centrality of reason-giving to the Board’s design and purpose, the Board’s reasoning has been thin. It has prioritized form over substance, simplicity over depth, and failed to seriously grapple with the most significant questions it was set up to answer.

The Board’s reasons have mechanistically followed a uniform template. They begin with a recitation of the relevant facts and submissions, followed by citations to the general rules the Board will apply. Then the opinions evaluate the merits of Meta’s original decision under Meta’s policies, values and what the Board considers its human rights responsibilities to be. Finally, there are the Board’s conclusions and recommendations. The format never changes, and the style is dry and rote, with the Board preferring clarity

---

239 See supra Part I.B. See also Charter, supra note X, Introduction (“The board will operate transparently and its reasoning will be explained clearly to the public”).

and consistency over rhetoric. The decisions are also economical—the analysis has grown gradually longer, but remains brief by the standards of most legal decisions, especially ones of such complexity and consequence. There are no dissents. To the extent that any Board members disagree with a decision, this is usually briefly noted at a high level in a sentence or two. As a matter of aesthetic impression, then, the Board’s decisions exude competence and professionalism perhaps, but not panache.

The formal veneer of these decisions obscures the ways in which the Board has neglected deeper questions about its role and the basis of its decisionmaking. Maybe the hope is that if the decisions look like rigorous legal decisions, that’s enough. The Board has opted for simplicity and consensus, rather than complexity and argumentation, leaving it with relatively thin theoretical foundations for its work. This makes the Board’s reasoning ultimately unsatisfying. The Board performs the role of reason-giver, but often gives very little by way of meaningful justification for its decisions.

The most significant manifestation of this is in the Board’s development (or lack thereof) of its analytical framework. A central problem for content moderation decisionmakers in search of legitimacy is that there is no existing normative framework for their decisions. No prior body of precedent deals with how corporate speech regulators operating at massive scale should govern platforms that are privately owned but significantly affect the public interest.241 Indeed, the Board was created in part because there were no established norms, agreed modes of analysis, or pre-existing public forums for developing a coherent body of well-reasoned and principled content moderation decisions. The Board’s public reason-giving was therefore intended to fill this gap, by developing a normative framework upon which to base private platform speech regulation.

The Board’s answer to this problem was to adopt IHRL as its framework. As it has since described it, “[a] defining theme of the Board’s work is our conviction that Meta will make content moderation decisions in a fairer, more principled way if it bases them on the international human rights standards to which it has committed itself.”242 The way the Board has applied IHRL is by cutting-and-pasting the three-part test under IHRL for state restrictions on speech into Board decisions, testing Meta’s rules for legality, a legitimate aim, and necessity and proportionality.243

241 See, e.g., Packingham v. North Carolina, 137 S.Ct. 1730, 1737 (2017) (observing that social media platforms are “for many are the principal sources for knowing current events, checking ads for employment, speaking and listening in the modern public square, and otherwise exploring the vast realms of human thought and knowledge.”).
242 Oversight Board 2022 Annual Report, supra note 20, at 51.
243 David Kaye, Report of the Special Rapporteur on the Promotion and Protection of
The Board’s decision to invoke IHRL is not terribly surprising, although it wasn’t inevitable. The Board’s founding documents direct it to pay particular attention to international human rights norms and principles, and do not mention any other source of law, but they stop short of mandating that the Board apply IHRL directly. The ambivalence of the founding documents makes sense given, as I will discuss in detail shortly, it is not clear how IHRL should be applied in this context. There has nevertheless been a growing movement in academia and civil society for platforms to adopt IHRL as their normative framework. The Board may have been heeding these calls when it turned to IHRL. But it also may simply not have considered any alternatives.

Formally, the Board has justified its invocation of IHRL primarily by citing reports of a UN Special Rapporteur who endorsed this approach. The Board did not discuss the authoritative weight of such reports (which are persuasive but subsidiary authority under IHRL). The Board has also cited Meta’s voluntary commitment to respect human rights in accordance with the United Nations Guiding Principles on Business and Human Rights (UNGPs) as a reason to rely on the IHRL framework. In later decisions, the Board justifies its use of the IHRL framework self-referentially. In most recent cases, it provides as the reason to use the IHRL framework the fact that “[t]he Board has employed the three-part test based on Article 19 of the [International Covenant on Civil and Political Rights (ICCPR)] in all of its decisions to date.” This self-referential justification is now a ritual incantation in almost every decision, with the Board referring to its own prior decisions as a “source of authority” which requires the use of the IHRL analytical framework.

In other words, the Board has treated the decision to adopt the IHRL...
framework as the basis of all its decisionmaking as a simple and self-evident choice. In fact, this framing belies much complexity. Contrary to what the Board’s cursory analysis might suggest, IHRL does not neatly map on to the problems content moderation raises.\textsuperscript{252} Crucially, IHRL, like First Amendment law, is a body of norms intended to constrain public authorities. It cannot simply be transposed from state-based jurisprudence and applied to the practices of private companies without interrogation of the very meaningful differences between these two contexts. Both private power and public power can threaten the speech rights of others, but the nature of the threat is different, and the normative framework for protecting speech in the two contexts must necessarily differ. Therefore, if IHRL is to be applied to content moderation by private companies, the key question is \textit{how such rules differ in the context of a private company versus when they are being applied to a state.}

The characteristics of public and private speech regulators differ in several relevant and significant respects. What that means for their responsibilities is a deep and difficult question that is beyond the scope of this paper, but it’s enough to note the contours of the problem and the obvious issues with which the Board is failing to engage. First, governments and platforms can threaten different sanctions. A government can lock you up, fine you, or change your legal status. Companies cannot do these things. This does not mean their actions cannot have profound consequences, though—platforms can, in some cases, cut someone off from their livelihoods, their social circles, or prevent them from receiving everyday modern services like deliveries or cloud storage. Second, the social meaning of being sanctioned by a public actor and a private one may differ. State sanction usually carries with it a level of social stigma that may not attend sanction by a corporation. Third, while both governments and platforms may regulate speech to suit their own interests, their motives for doing so may (and likely in many cases do) differ. Free speech jurisprudence is especially vigilant about governments suppressing particular ideologies or opinions they dislike,\textsuperscript{253} and while companies may also have political agendas, the dominant concern is usually

\textsuperscript{252} For more extensive analysis on this point, see douek, supra note 246; Dvoskin, supra note 217, at 113–24; EMILY B. LAIDLAW, REGULATING SPEECH IN CYBERSPACE: GATEKEEPERS, HUMAN RIGHTS AND CORPORATE RESPONSIBILITY 292 (2015); Molly K. Land, Regulating Private Harms Online: Content Regulation under Human Rights Law, in HUMAN RIGHTS IN THE AGE OF PLATFORMS 285, 292 (Rikke Frank Jørgensen ed., 2019); Susan Benesch, But Facebook’s Not a Country: How to Interpret Human Rights Law for Social Media Companies, 38 YALE J. REG. ONLINE BULLETIN 86 (2020); Rachel Griffin, Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality, 2 EUROPEAN LAW OPEN 30 (Cambridge University Press Mar. 2023).

\textsuperscript{253} Rosenberger v. Rector and Visitors of the University of Virginia, 515 U.S. 819, 829 (1995) (“Viewpoint discrimination is … an egregious form of content discrimination.”)
that the company’s commercial interests will skew their decision-making. 

Fourth, states and companies have different kinds of expertise when it comes to assessing the harms and benefits of different kinds of speech restrictions. The government will be in a position to assess national security or foreign relations considerations, for example, that a private company simply will not. Conversely, a corporation might have more insight into the specific ways its products are exploited, such as manipulative influence operations or commercial scams, and greater capacity to intervene in a targeted way. Fifth, the costs of entry and exit for affected parties are different—Meta’s platforms may be dominant, but it is still easier to set up an account on a different social networking service than it is to move country.254 Sixth, the tools available to public and private speech regulators to enforce their rules are very different. Just as a matter of sheer technical capacity, platforms have far more ability to be flexible and nimble in their rule-enforcement, and the online environment means they have more tractability and visibility into their speech environments than offline regulators. Finally, and significantly, corporations have their own speech (or if you prefer, “business”) rights and interests, which surely entitle them to some latitude a state should not have. But the question is: how much?

These important differences in how and why states and corporates regulate speech mean that developing a normative framework for corporate content moderation raises new questions that cannot easily be answered by a pre-existing normative framework designed for states—whether that normative framework be IHRL, or the First Amendment, or anything else. It is not an exaggeration to say that the central challenge of creating a normative foundation for platform content moderation is to account for to the ways in which companies can regulate speech that a government cannot. If it were not for this difference, platforms should simply be required to observe the restrictions that are placed on government actors. That is, old precedents could just be applied directly. Of course, there would be difficult questions about how such precedents apply to new facts, as there always are, but there would be no deep theoretical knot to untangle. But a knot there is, and untangling it is the core challenge for the Board if it wants to create a body of principled content moderation rules. It cannot simply invoke IHRL and suggest that it readily provides the answers.

Instead of facing this challenge head-on, the Board’s decisions only tend to confirm that the work that IHRL does in this context is largely symbolic.255


255 Douek, supra note 246, at 56–58 (noting that IHRL will not provide determinate answers to most difficult content moderation questions); Griffin, supra note 253, at 50 (noting that often “Human rights provisions are not doing much work here beyond providing general rhetorical support for policy arguments influenced by other political views about...
The Board has barely discussed the differences between state and private speech regulators—it has not even discussed the ways in which the UNGPs explicitly distinguish a State’s and a company’s obligations under IHRL. Indeed, it more often emphasizes the way companies and states might face similar questions, rather than their differences. In no case has the use of IHRL meaningfully constrained the Board, nor could it. This is not a weakness of IHRL alone—because there is so little precedent about how to think about free speech interests in the context of private platforms, no body of pre-existing norms answers the difficult questions for making rights-respecting content moderation decisions. This indeterminacy means that the Board could start from IHRL, Meta’s values, the First Amendment, or first principles, and ultimately the utility of each of these frameworks would run out in the same way: on most questions, there is no pre-existing precedent, and the Board will simply be left making a judgment call. But it is a weakness in IHRL that needs to be addressed if it is to be made into a meaningful normative framework for content moderation decisions.

The Board initially acknowledged the centrality and importance of its role as the translator of existing norms to this new context. In one of its first decisions it noted that the human rights obligations imposed on private companies are not the same as those imposed on governments, and concluded accordingly that “clarifying the nature” of IHRL obligations on Meta was the “principle task facing this Board.” As Helfer and Land observe, “the Board is uniquely placed to explain, through an accretion of case-specific decisions on content moderation and broader policy recommendations, how international rules that bind states apply to private social media companies.” And yet it has never really attempted that task. It has not clarified the nature of Meta’s obligations under IHRL at all. Instead, the conclusory way the Board adopted IHRL framework is mirrored in the conclusory way it applies it in individual cases.

From time to time the Board still notes in its decisions that the framework that applies to a private company will necessarily be different from that which applies to a government. But these comments tend to be cursory and are how online media should be governed.

256 These principles distinguish States’ “duty to protect” from a businesses “responsibility to respect” IHRL: UNGPs, supra. See Dvoskin, supra note 217, at 99.
258 2020-003-FB-UA (Armenians in Azerbaijan Case).
259 Helfer & Land, supra note 226.
260 See, e.g., 2021-002-FB-UA (Zwarte Piet Case) (upholding a Facebook rule that the Board said would be impermissible under IHRL); 2021-016-FB-FBR (Journalism on Sexual Violence Case) (stating that “[a]lthough the ICCPR does not create the same obligations for Meta as it does for states, Meta has committed to respecting human rights”) (emphasis
becoming less frequent. More often the Board does not advert to the difference at all. The Board instead regularly cites rules that apply to states and then proceeds to apply them to Meta, without comment, often using general and passive language like “restrictions on freedom of expression must [meet IHRL criteria].”\footnote{See, e.g., 2021-009-FB-UA (Al Jazeera Post on Israel & Palestine Case); 2021-008-FB-FBR (Brazilian Authority on COVID Lockdown Case).} eliding that the authorities they cite to are exclusively state-based jurisprudence. In some cases, in what might be a Freudian slip, the Board gives incantations like “[t]he principle of legality under IHRL requires rules used by states to limit expression to be clear and accessible,”\footnote{2021-007-FB-UA (Situation in Myanmar & Profanity Case) (emphasis added). \textit{See also} 2021-008-FB-FBR (Brazilian Authority on COVID Lockdown Case).} without noting that, well, Meta is not a state and so this rule does not technically apply. Far from working through the underlying normative questions, typically the Board conclusorily asserts that certain outcomes are supposedly required by IHRL but without explaining why.

A few concrete examples may be helpful. A particularly flagrant example of shallow reasoning is the Russian Poem Case. The case concerned a Facebook post comparing the Russian army in Ukraine to Nazis, and which included a photo of what appeared to be a dead body, face down in the street. The Board summarily held that Meta could not put a warning screen over the photo—a screen that any user could click through—because such a screen was not “necessary” under IHRL given that the photo did not show any “clear visual indicators of violence.”\footnote{2022-008-FB-UA (Russian Poem Case).} No authority is cited for this proposition. No reasoning is offered. No justification is given for why IHRL would prevent a private company from giving users advance notice before they see what appeared to be a murder victim. No citation is given for the proposition because none exists—how warning screens should be used on private platforms is of course not a question IHRL authorities have considered before.

This illustrates one of the difficulties of applying IHRL in the context of content moderation: almost all issues are novel. How should Article 19 of the ICCPR apply to warning screens in social media newsfeeds? What about de-amplification of content? Or turning off certain account features, such as the ability to livestream, as sanction for breaking the rules? Prior IHRL authorities simply have not grappled with these questions. The Board therefore cannot rely on IHRL as is, without further explanation. It is not that the principles underlying IHRL \textit{cannot} guide content moderation decision-
making—the nature of common law reasoning is to take old cases and apply them to or distinguish them from new situations. But to do this requires work—work that the Board is just not doing in its decisions.

Conversely, there have been a few cases where the Board has found that Meta can *depart* from what IHRL would require of a state because it is a private company. But in these cases, too, the Board has failed to provide much explanation as to why. For example, the Board ruled that Meta could remove content depicting blackface, including portrayals of Zwarte Piet, a Dutch cultural tradition, because such content creates a discriminatory environment for Black people, and this was so even though IHRL “would not allow a state to impose a general prohibition on blackface.”264 A majority of the Board found that because there was some evidence of severe harms at a societal and individual level stemming from Zwarte Piet and other stereotypes, “this justified Facebook adopting a policy that departs from the human rights standards binding states.”265 But the majority did not explain or justify this decision in any way. What exactly about the private status of Meta justified the departure from IHRL norms binding states in this case? Is it something to do with the different sanctions or social meaning of Meta’s rules? Is it something to do with the unique features of the online environment? Is this a one-off exception, or some generalizable principle? The Board simply stated the outcome, leaving any deeper questions unanswered.266

The superficiality of the way the Board has adopted and applied the IHRL framework manifests in the Board’s reasoning more broadly. For example, despite adopting the norms and forms of judicial decisionmaking, the Board has generally been quite casual in its use and development of precedent. It invokes and purports to follow IHRL, but it has never discussed the facts of any IHRL decision. The Board’s use of its own decisions as precedent has also been perfunctory. It has never suggested anything other than that the application of a prior decision is clear, and has ignored inconvenient tensions. In the policy advisory opinion in which the Board concluded that Meta should continue to remove Covid-19 misinformation, for example, the Board did not even mention that in one of its first decisions it had overturned Meta’s decision to take down a post containing misinformation because the harm it could cause was not sufficiently “imminent.”267

Or take three cases that raised the question of when Meta can or should

---

264 2021-002-FB-UA (Zwarte Piet Case).
265 Id.
266 Dvoskin, *supra* note 217, at 133 (similarly noting that “in the Black Pete case, the Board’s adoption of the IHRL project actually harmed the coherence and clarity of its decision.”).
267 2020-006-FB-FBR (Claimed COVID cure case).
add warning screens to violent content. In one case, the Board upheld Meta’s decision to restore a graphic video that depicted a wounded body in Sudan and apply a warning screen, stating that “[t]he warning screen does not place an undue burden on those who wish to see the content while informing others about the nature of the content and allowing them to decide whether to see it or not. The warning screen also adequately protects the dignity of the individual depicted and their family.”268 Five months later, in the Russian Poem case described above, the Board overturned Meta’s decision to apply a warning screen to an image of a dead body in Ukraine, saying that “the use of a warning screen inhibits freedom of expression and is not a necessary response in this instance.”269 The latter case cited to the former case in its “sources of authority,” but otherwise did not discuss it or attempt to resolve the inconsistencies. A minority of the Board disagreed with the majority, stating that “Meta may err on the side of prudence” in such cases, but the minority also did not invoke the prior decision. It is not that the cases are necessarily inconsistent. The Board could have, for example, suggested that a difference in the level of graphic gore in the two cases led to its different conclusions. Instead, the Board simply did not discuss the prior case at all. In a third case, the Board changed tack again and upheld Meta’s decision to add a warning screen to a video that appeared to document war crimes committed by Azerbaijan in which Armenian prisoners of war and casualties were visible.270 The Board relied on the Sudan graphic video case and held that “providing users with the choice of whether to see disturbing content is a proportionate measure” which showed “respect for the rights of the prisoners and their families.”271 In this decision, the Board at least acknowledged the inconsistency with the Russian Poem case, but did not explain in any detail how to reconcile the decisions, beyond perfunctorily suggesting that the content was less graphic in the earlier decision.

By citing authorities but not meaningfully engaging with them, the Board gives the appearance of creating a body of precedent, without doing the hard work of reconciling cases that such a task entails. Some of the Board’s “doctrinal” incoherence, of which more examples could be given, no doubt stem from the differing make-up of different panels. But the Board has opted for unsigned opinions, and without knowing who makes decisions, the result is that the Board as a whole has the appearance of contradicting itself. It also means no individual member of the Board pays a reputational cost for poor work product.

A number of the Board’s practices have, intentionally or not, allowed this

268 2022-002-FB-MR (Sudan Graphic Video Case).
269 2022-008-FB-UA (Russian Poem Case).
270 2023-004-FB-MR (Armenian prisoners of war video case).
271 Id.
lack of deep reasoning to fly under the radar. First, the fact that it is deciding so few cases means the Board does not need to confront the task of reconciling its decisions often. This very fact illustrates why deciding more cases is important—it is through cases that the Board should be explaining its reasoning and how to weigh competing considerations in different contexts. An essential part of the Board’s role is to play a discursive role in the ongoing, ever-changing debate about online speech rules. It cannot do this through intermittent, scattershot engagement.

Second, the shallowness of the Board’s reasoning is both less noticeable because of, and compounded by, the way the Board has decided to present minority opinions. Despite the complexity of the issues and the size of the Board, minority opinions are comparatively rare, with one being expressed in only 14 out of 39 cases. More importantly, the presentation of minority opinions is usually brief, low on detail, and does not note the size or composition of the minority. Thus, even when there is internal disagreement, the Board does not describe the nature of the disagreement in a way that helps readers understand the theoretical complexities of the cases. In these ways, the Board has opted for simplicity over depth, and presenting a unified front over ventilating points of contention. Perhaps this made sense in the Board’s early cases—providing the institution an opportunity to establish its authority as a cohesive institution rather than highlighting the differences of its members. But three years on, the lack of meaningful debate even amongst the Board members only serves to highlight the overall shallowness of the opinions.

Third, the Board’s decisions cannot be appealed, do not have to be interpreted by lower courts, cannot be invoked by litigants seeking to make their own cases, and are generally not embedded in a broader community of stakeholders that carefully read the Board’s jurisprudence as whole. For all the attention that the Board gets, little of it engages with the Board’s work at a level of detail that would highlight these kinds of inconsistencies or tensions. As a result, generally speaking, the Board’s reasoning and modes of analysis remain relatively unexamined.

The above critiques of the Board’s lack of normative framework and doctrinal coherence are not mere academic nitpicking or law review quarterbacking. Superficial reasoning undermines the guidance the Board’s decisions give to Meta in future cases (pity the poor content moderator who has to determine when applying a warning screen violates IHRL in the

---

272 See supra Part II(A)(1).
273 [As at July 30, 2023].
274 Helfer & Land, supra note 115, at 50 (similarly noting that “Allowing individual Board members to write separate dissenting or concurring opinions will help identify areas where norms are in flux and would help promote dialogue about those junctures.”).
Board’s eyes, let alone why!), to users about the rules that apply to them, and to regulators for understanding these complex systems. It makes disagreeing with the Board’s decisions harder, when the bases for those decisions are obscure. And it also undermines the justification for the Board itself, if it fails to engage with one of the most important tasks it was set up to perform—starting to provide a common language and framework for thinking about how to justify platform rules that are different from public free speech laws in a coherent and thoughtful way. Instead, the Board’s reasons have all the trappings of jurisprudence, but too often they lack the underlying substance.

2. Formalistic Measures of Impact

If the Board’s guiding normative framework is unclear, the criteria for assessing the Board’s material impacts—that is, the affects its work has in the world—are even more so. What impact the Board’s decisions have had on the world directly is almost impossible to say, and the Board itself does not seem especially interested in answering the question.

It is somewhat striking how little material impact comes up in the discourse around the Board. There will often be a flurry of commentary and news articles when the Board hands down a case, but little—if any—reporting on Meta’s responses to, or implementation of, the Board’s recommendations. But the Board’s decisions are the least consequential part of the entire process in a material sense, especially because its recommendations are non-binding. All that matters, if the Board is to have more than rhetorical impact, is what Meta does in response. And yet it is the moment of decision that draws public attention. This dynamic is not unique to the Board and illustrates a deeper problem in how content moderation governance is evaluated: anecdotes about individual cases and decisions capture the imagination far better than questions about system design.275 The story of the Board overturning Meta, rebuking it, putting it in its place, is far more attention-grabbing than, say, whether Meta publishes error rates for its Media Matching Service bank of violating images.276 But in substance, the latter matters far more for systemic improvement. And what happens as a result of the publication of those error rates (i.e., does it lead to some other form of accountability or reform) matters even more still.

There are a few problems with the way the Board appears to think about impact. The first relates to the Board’s emphasis on ever-more procedure as the indicia of effective governance. The Board’s recommendations often reflect an assumption that giving users more detailed rules, more extensive reasons for decisions, more opportunities to appeal, and in general more due

275 Douek, supra note 112, at 559–60.
276 2022-004-FB-UA (Colombian police cartoon case)
process, will necessarily result in a better and more legitimate system overall. But this assumption reflects a legalistic procedural fetishism that depends on abstract notions of legitimacy and accountability not always borne out in practice, and does not acknowledge the costs of piling on procedural requirement after procedural requirement.277 The Board’s intuitions that Meta’s content moderation needs more individual due process cannot be merely assumed, and the Board should treat them as hypotheses that are subject to testing and revision. Instead, the Board’s interpretation of IHRL has been extremely formalistic, and unconcerned with practical impacts. For example, the Board’s interpretation of the principle of “legality” under IHRL that requires “rules that limit expression to be clear and publicly accessible” has resulted in many recommendations that Meta add additional wording to its already voluminous Community Standards.278 But the Board has not questioned whether this actually results in better outcomes for any users or the public more broadly.

Another category of problems arises as a result of the difficulty of quantifying impact. The Board does track the results of its recommendations, but in a narrow way. The Board tracks Meta’s implementation of its recommendations, but not the effects of that implementation. As evidence of its importance, the Board notes that “Meta has publicly recognized how our recommendations are changing its behavior.”279 Tracking implementation alone, however, shows the Board’s impact on Meta, not the Board’s impact on the world. When Meta adds definitions to “non-medical drugs” and “pharmaceutical drugs” to its Community Standards, or starts giving users more detailed notifications for specific policy violations for hate speech,280 does that actually have meaningful effects on how speech flows through its platforms, let alone society more broadly? And… are they good effects? While the Board’s impact on Meta may be relevant to the question of the Board’s impact on the wider world, it is not a perfect or even necessarily reliable proxy. Without more, knowing only that Meta made some change in

277 Nicholas Bagley, The Procedure Fetish, Mich. L. Rev. 345, 369 (2019); see also the discussion in Douek, supra note 112, at 577–78; Elettra Bietti, A Genealogy of Digital Platform Regulation, 7 Geo. L. Tech. Rev. 1, 45 (2023) (“Governance efforts and law-like procedures … become a façade with little intrinsic value and detrimental effects for the public”).
278 See, e.g., Kayyali & York, supra note 235 ("Out of the Oversight Board’s 18 recommendations to date, nearly every one of them echoes the Santa Clara Principles’ recommendation to disclose “detailed guidance to the community about what content is prohibited.”"). [Will update number closer to publication.]
279 Oversight Board 2022 Annual Report, supra note 20, at 19.
its practices is not evidence of positive material impact.

But the Board is not incentivized to measure its material impact in this way for three reasons: first, broader material impact is hard to measure; second, there is no normative agreement on what impact is positive, making outcome goals difficult to define; third, it is easier for the Board to sell a story of success if it focuses on the narrower question of its impact on Meta than some broader question about its effects in the world.

First, the challenges of measuring the impact of various speech rules and governance practices are substantial. Determining even the first-order effects of speech decisions is hard, let alone the flow-on effects of any changes in a complex system. Every system of speech governance has struggled with it. Take the example of hate speech—perhaps one of the most controversial free speech issues, and one on which you might think that there is a perfect natural experiment: the United States is unique in its almost blanket protection of hate speech. Research has not been able to settle long-standing debates about the relationship between hate speech laws and hate speech, or the effects of hate speech itself.

But such difficulty is not sufficient reason to ignore the importance of substantive social impact as a metric of good governance. The consequences of rules and their enforcement have for society at large “are the broadest and most diffuse of all effects, but in the final analysis, they are the most important.” The whole project of speech governance rests on an assumption that institutions matter because they affect the world. The Board’s existence and decisions assume that the ways in which speech rules are written and enforced matters in some material way. Perhaps the best available measures of impact will always be indeterminate, disputed and contingent, but they should not be neglected for this reason alone. Working out the substantive impact of various decisions requires at least starting with

---


the right questions. What should matter to the Board is not only whether Meta implements its recommendation to give users better notifications about hate speech violations, for example, but also whether better notifications in some way benefit not only users but the speech environment more broadly. Focusing on the former at the expense of the latter is goal displacement, with the Board becoming focused on what is measurable rather than what is impactful.\textsuperscript{284} The disruption of the platform era offers an opportunity to start answering questions about speech governance that have typically been answered by intuition alone. Because much speech online is tractable and searchable, rather than ephemeral, there is now more data available than ever before about the way speech flows through society.\textsuperscript{285}

The task of measuring impact is made even harder by ambiguity in defining the Board’s relevant constituency. The first question must be: who exactly is the Board acting in service of? Is it Meta’s users, or society more generally? Zuckerberg seemed to envision the former, writing that “[i]n the case of our board of directors, our board of directors is accountable to our shareholders, [the Board] would be focused only on our community.”\textsuperscript{286} The Board also suggests that it “was established … to promote the rights and interests of users.”\textsuperscript{287} And many of its recommendations focus on Meta’s due process obligations to the people directly affected by decisions rather than the diffuse impact Meta makes in society. But users’ interests will not always align with the broader public interest. And the Board’s current emphasis on users’ interests and rights—although perhaps an attractively more modest domain—is somewhat in tension with the UNGPs it purports to apply, which emphasize business enterprises’ responsibility to address any adverse human rights impacts.\textsuperscript{288} A proper application of these principles thus requires the Board to be attentive to Meta’s (and its own) impacts on society beyond Meta’s users.\textsuperscript{289} The Board needs to be clearer about its community of interest in order to determine if it is serving it well.

But this leads to the second problem, because evaluating (not just

\textsuperscript{285} First, in \textit{Social Media and Democracy} \textsuperscript{8} (Nathaniel Persily & Joshua A. Tucker eds., 2020) (“For the first time in human history, we have real time records of millions – if not billions – of people as they discuss politics, share information about politics, and organize politically. Each of these actions simultaneously produces an archived, digitized record.”).
\textsuperscript{286} Id. at 3 (emphasis added).
\textsuperscript{287} Id. at 3 (emphasis added).
\textsuperscript{288} Id. at 3 (emphasis added).
measuring) performance on the basis of real-world impact requires some notion of the normative good that the Board is trying to achieve. Alas, defining the normative goals of speech governance is even more fraught than trying to measure its impact.

Recall, for example, that Feldman’s goal that the Board should be “an effective counterweight to censorship.” On this view, the measure of success is whether the Board has decreased the amount of content that Meta takes down. This is a metric that can be measured. And, indeed, it’s possible that on this metric the Board has been quite successful. In cases where the Board overturned Meta (that is, changed the substantive outcome), in all but three the Board ordered a post to be reinstated that had initially been taken down. In fact, in its entire corpus of cases, the Board has only thought that a post should be taken down eight times. It seems, then, that the Board itself views its role in the same vein as Feldman—to push back on Meta’s removals. If this is your conception of the substantive good, then the Board may be “working.” But of course, hardly everyone agrees that the problem with content moderation is that platforms take down too many posts. Many believe the main problem with Meta’s content moderation is that it leaves far too much harmful content up.

Whatever the Board’s own goals for its substantive impact, it has not been explicit about them, let alone measured progress towards them. Instead, the Board has shown relative disinterest in this project. For example, one of the key avenues of potential impact of the Board is Meta’s obligation to implement the Board’s decisions with respect to individual pieces of content to “identical content with parallel context.” But neither the Board nor Meta have publicly tracked how broadly Meta has carried out this obligation, and thus even the direct reach of the Board’s decisions remains unknown.

---

290 Global Feedback and Input on the Facebook Oversight Board for Content Decisions: Appendix, supra note 59, at 102.
291 2022-001-FB-UA (Krin cartoon case); 2023-001-FB-UA (Brazilian general’s speech case); 2023-003-FB-MR (Cambodian Prime Minister case). [As at July 30, 2023]
292 [Not including Trump – Krin; Armenians; Zwarte Piet; Tigray]
293 In 2023-001-FB-UA (Brazilian general’s speech case), the Board suggested for the first time that it would “pay special attention to Meta's application of its decision to identical content with parallel context that has remained on the company's platforms,” but did not specify how it would do so.
294 Charter, art 4 (“In instances where Facebook identifies that identical content with parallel context – which the board has already decided upon – remains on Facebook, it will take action by analyzing whether it is technically and operationally feasible to apply the board’s decision to that content as well”): see also Bylaws art 2.3.1.
establish meaningful systemic impact on Meta’s content moderation, it needs to track far more than whether has restored the few dozen pieces of content that the Board has examined directly or whether Meta has ticked the box on procedural reforms. The Board may be aware of this weakness in its reporting. In the transparency report for the first quarter of 2023, the Board included additional data about the impacts of two prior recommendations for the first time, stating that Meta had shared data showing that more content was being sent to human review than would have otherwise as a result of the Board’s recommendations. But this is a great illustration of exactly my critique—the mere fact that more content was sent for human review cannot, without more, merely be assumed to be beneficial, unless your definition of benefits revolves entirely around proceduralism.

The final reason the Board may not be tracking its broader impact is more cynical: the results may be less impressive compared with the Board’s direct impact on Meta alone. Being able to show that Meta took X number of concrete steps in response to its rulings helps the Board establish its credentials in its role as watchdog, and is far less indeterminate and risky than trying to measure downstream impacts.

3. Persistent Strategic Avoidance

Surprisingly, for an institution set up to “help” Meta “resolve … what to take down, what to leave up and why,” the Board has a clear pattern of ducking the most controversial questions that come before it. In the most difficult cases, the Board is very eager to tell Meta it did something wrong, but much less eager to tell Meta how to get things right. This is a much more risk averse approach: Meta cannot ignore, and people cannot disagree with, a decision if no decision is made. But the persistence with which the Board does this undermines its utility as a decisionmaker of last resort that offers some measure of finality in content moderation debates.

Perhaps the clearest example was the Board’s decision with respect to the deplatforming of then-President Trump. Instead of answering the question of whether Trump’s account on Facebook should be reinstated, the Board instead rebuked Meta for the way in which it had made the decision and sent it back to Meta, giving Meta six months to make a new decision. By avoiding making the hard call itself, “[t]he realist take is that the [Board] was trying to avoid controversy and prioritize its own legitimacy.” In a sense,

295 Oversight Brd., supra note 118, at 17.
296 The Oversight Board, supra note 55; Oversight Board, supra note 55.
297 2021-001-FB-FBR (Trump Suspension Case).
298 Id.
299 Evelyn Douek, It’s Not Over. The Oversight Board’s Trump Decision Is Just the
this was a politically deft move. A non-decision with respect to Trump’s account forestalled the public outrage that might have followed a substantive decision either way. It also sidestepped being too confrontational with Meta. By slapping Meta over the knuckles for procedural deficits in its decision about Trump’s account, the Board looked like it was holding Meta accountable without actually testing the limits of its coercive authority—Meta could not ignore or contradict the Board’s decision if the Board didn’t make any decision. But while politically deft, the decision was also an abdication of duty. Making this kind of hard decision was squarely within Board’s job description to “exercise independent judgment over some of the most difficult and significant content decisions.” The Board ducked its responsibility in order to avoid being made to look weak or becoming unpopular.

At the time, the decision was analogized to *Marbury v. Madison* because it was the first big display of the Board’s power. The analogy is apt, but not for that reason. By contrast with the Supreme Court’s powers of judicial review, it was never in doubt that the Board had the formal power to issue a decision in the *Trump Suspension Case*—indeed, such decisions were the reason it was created. But the analogy fits as an example of judicial statecraft in which the Court “protected itself from immediate political attack by finding that under the particulars of the case it was without jurisdiction to act.”

This was merely the first example of what would become a regular practice of the Board sending the most high-profile and difficult decisions back to Meta to make for itself. The Board’s decision about Covid-19 misinformation is another example. In July 2022, Meta asked the Board for advice about “whether Meta’s current measures to address COVID-19 misinformation under our harmful health misinformation policy continue to be appropriate” in light of the changing pandemic conditions. Nearly a year

---

300 Clegg, supra note 42 (“we have created and empowered a new group to exercise independent judgment over some of the most difficult and significant content decisions.”).
301 5 U.S. (1 Cranch) 137 (1803).
304 Policy advisory opinion 2022-01 (Removal of COVID-19 misinformation).
305 Nick Clegg, *Meta Asks Oversight Board to Advise on COVID-19 Misinformation*.
later, in April 2023, the Board issued its opinion—but not an answer. Its advice was that Meta should maintain its current policy, but also “should begin a process to reassess each of the 80 claims it currently removes, engaging a broader set of stakeholders.” To which one might respond: wasn’t that exactly what Meta was asking the Board to do? It’s not hard to guess why the Board might not want to give a definitive to the answer, though. How platforms have dealt with Covid-19 misinformation has been one of the most politically controversial topics in content moderation. By again rebuking Meta for inadequate procedures and consultation about its policies but not providing a definitive answer on what the policy should be, the Board was able to perform toughness without wading into a political quagmire.

This dynamic is not confined to the most high-profile cases. The Board took the same approach in a perhaps less obviously consequential and lower-profile context—Meta’s policies around nipples. How Meta should treat nipples on its platforms is one of the earliest and most enduring content moderation controversies. It raises (perhaps surprisingly) difficult normative questions. From its earliest days, Meta made a business decision to generally ban adult nudity. Such a rule aligns with many users’ preferences, and prevents Meta’s platforms from becoming flooded with porn. As a business decision this might be a simple call, but as a speech decision it is far more complicated. A ban on female-presenting nipples perpetuates certain social norms about sexualization that many disagree with, stigmatizes practices such as breastfeeding, and has disparate impact on already marginalized communities such as women, societies that traditionally go bare-chested, and LGBTQ individuals. A broad-based ban on adult nudity of the kind Meta has historically had could never be imposed by a state in accordance with IHRL, or under most national constitutions. The question of how to treat nipples therefore raises thorny questions about how to reconcile a platform’s right to define what kind of product experience it wants to provide its users in accordance with its business interests, on the one hand, with traditional free speech norms, on the other. But the Board punted this


Policy advisory opinion 2022-01 (Removal of COVID-19 misinformation).


decision too, when it came up in the *Gender Identity and Nudity Cases*.\(^{309}\) Rather than recommend a policy to Meta, it simply said that Meta should conduct a human rights impact assessment and then “define clear, objective, rights-respecting criteria to govern the entirety of its Adult Nudity and Sexual Activity policy consistent with international human rights standards.” But surely defining such criteria was a core aspect of the Board’s intended role.

The same political dynamics may also partially explain the Board’s low case-load. Fewer cases means fewer opportunities for confrontation or mistake. It is notable how rarely the Board has attracted controversy, given how many controversial cases it no doubt could have selected. As the Board itself emphasizes, it receives many, many appeals, and presumably could take cases on almost any topic. But the Board never has to take a case it doesn’t want, and it has avoided touching many of content moderation’s livest wires.

From the perspective of evaluating the Board as a governance institution, this pattern seems like a failure to meet its responsibility. But as a matter of politics, the Board’s desire to avoid controversy and side-step hard calls is rational. The Board is *not* actually incentivized to weigh in on intractable normative and moral debates, despite this being its ostensible purpose; instead, the Board is incentivized to perform the role of solemn governance institution, without upsetting too many people. It should therefore not be surprising that the Board has focused so much on the appearance of its judgments and its public communications, and less on its substantive productivity and impact. It is these incentives to which the next Part turns.

### III. Legitimacy’s Empty Promise

The story of the Board’s success is therefore complicated and, this Part argues, ultimately suggests that the pursuit of sociological legitimacy for its own sake creates perverse incentives. Legitimacy lacks substantive content— in some ways, that’s the very point. Agreeing on the substantive goals of content moderation is hard; agreeing we should have respect-worthy institutions is less so. But this breeds complacency about what governing institutions can and should achieve. The Board’s focus on legitimacy building has come at the cost of a more ambitious, and substantive, vision for content moderation governance.

The positive version of the story is that the Board has attained far more public attention and acceptance than many expected. For Meta and the Board, this is an indisputable win. Meta’s creation of the Board and outsourcing of certain kinds of authority evinces Meta’s belief that reestablishing public trust was more important to its business than any particular content moderation policies.
decision or rule. That is, that legitimacy matters and has independent value to Meta. The subsidiary goals that people had for the Board were all instrumental to this central aim.

But for the rest of us, the sociological legitimacy of the Board—indeed, of any platform governance institution—should depend on the achievement of other ends, not be an end in and of itself. And here, the story is more complicated. The previous Part showed that while the Board has unquestionably brought some benefits, it has also neglected core aspects of its role.

Why has the Board neglected these tasks? It seems hard to make sense of. Answering difficult normative questions was the reason the Board was created, and the Board was designed with this task in mind. The Board is not made up of data scientists or engineers, and it does not perform ongoing audits of Meta’s systems. It is comprised of scholars, judges, human rights activists—that is, the kinds of people who are better placed to opine on the normative problems of content moderation, rather than its technical challenges. And yet, the Board has eschewed this kind of theory almost entirely. At the same time, the Board has hardly forsaken theory in the interests of a more practical and empirically grounded approach. The Board does not define, let alone measure, its goals in concrete terms. And despite all the big talk about “help[ing] [Meta] answer some of the most difficult questions around freedom of expression online,” often the Board dodges clear answers in the most fraught cases. This Part argues that these choices—initially baffling—make sense once the Board is understood as a political actor responding to its political environment.

Understanding this political environment is important for debates far beyond Meta and the Board. The Board of course matters on its own terms. Billions of people use Meta’s products, and the company makes millions of speech decisions every hour. A mechanism to bring transparency and accountability to this expansive global system of speech regulation is presumptively important. And this is all the more true given Meta and the Board’s ambitions to grow the Board’s oversight and influence. Board members have from its earliest days said they hope the project can expand and ultimately include reviewing content moderation decisions from platforms beyond Meta’s. Professor Thomas Kadri has carefully documented the ways Meta and the Board have made their more expansive

---

310 Oversight Board, supra note 55.
vision for the institution clear.\textsuperscript{312} Both have pushed for the Board to become a frame of reference for excellence in content moderation governance across the industry.\textsuperscript{313} Mark Zuckerberg explicitly pointed to the Board as an example of the kind of regulation that lawmakers should be thinking about,\textsuperscript{314} and Meta representatives have mentioned it repeatedly in Congressional hearings.\textsuperscript{315} Meta’s website proclaims the hope that “the board can serve as a model for the future of content governance across our industry.”\textsuperscript{316} The Board has made it clear that it is “interested in working with companies that share our belief that transparent and accountable content governance, overseen by independent bodies, is an essential part” of creating a rights-respecting online public sphere.\textsuperscript{317} Clearly, the Board hopes to directly impact far more than merely Meta’s services.

But, perhaps more fundamentally, the lessons of the Board experiment matter because debates around the Board provide a kind of microcosm through which broader debates about content moderation can be examined. The increasing public, academic and civil society acceptance of and engagement with the Board reveals something important about the broader political environment in which the Board operates. And, importantly, the political incentives to which the Board is responding will also exist for any content moderation governance institution. What the Board experiment shows is that “legitimacy” can be won through the presence of formalistic indicia including the appearance of reasoning, consultation, or certain other elements of due process. These indicia are seen as intrinsically beneficial, even without any further inquiry into its effects, and they are increasingly being mandated across the industry.

This juridical and legalistic model of governance that the Board exemplifies in particular is a tool increasingly favored by regulators.\textsuperscript{318} As

\begin{itemize}
\item \textsuperscript{312} Kadri, supra note 37, at 176–85.
\item \textsuperscript{313} Id. at 176–85.
\item \textsuperscript{315} See, e.g., Senate Judiciary Subcommittee Hearing Social Media Platform Algorithms and Amplification (Apr. 27, 2021) (Testimony of Monika Bickert, VP of Content Policy at Facebook discussing the progress made by the Oversight Board); House Energy and Commerce Subcommittee on Communications and Technology and Consumer Protection Hold Joint Hearing on Misinformation and Disinformation on Online Platforms (Mar. 25, 2021) (Testimony of Mark Zuckerberg mentioning the Oversight Board as an accountability mechanism). See also Kadri, supra note 37, at 195.
\item \textsuperscript{316} Creating the Oversight Board, META TRANSPARENCY CENTER (Jan. 19, 2022), https://transparency.fb.com/oversight/creation-of-oversight-board/.
\item \textsuperscript{317} Oversight Board 2022 Annual Report, supra note 20, at 62.
\item \textsuperscript{318} Douek, supra note 112.
\end{itemize}
Rachel Griffin has argued, the Board’s work and much regulation, especially in Europe, reflect a line of thought that believes in the legitimating power of rule of law procedural norms. Yoel Roth has observed the same phenomenon, bemoaning the “ever-more-formalized set of requirements around platform conduct” that tech companies and their regulators are turning to in order to try build trust.

Indeed, legislatures are writing and enacting laws that would require of all platforms the kinds of things that the Board’s recommendations to Meta seek to get Meta to adopt. For example, the European Union’s Digital Services Act—set to be one of the most globally influential pieces of legislation regarding content moderation—shares many of the same goals as the Board’s recommendations. It includes requirements for extensive procedural protections in every case: platforms need to provide reasons for any content removal, a right of appeal open for six months in all cases, a human in the loop for all appeals, and a further right of appeal to a third-party arbitrator. The Board itself has expressed an interest in becoming a formal part of this regime. Laws in Texas and Florida similarly require individual notice, appeal and reasons. The Irish government commissioned an expert group to advise on the feasibility of establishing an individual appeals and complaints mechanism for platform content decisions and that group recommended such a mechanism be established. The German Federal Court of Justice has also ruled that Facebook must provide every individual user notice if a post is deleted, reasons for that deletion, and an opportunity to reply. The Indian government has passed a law creating government-appointed “Grievance Appellate Committees” to hear user

319 Griffin, supra note 207, at 56–58.
322 DSA art. 17.
323 DSA art. 20.
324 DSA art 20(6).
325 DSA art. 21.
326 Kadri, supra note 37, at 194; Levy, supra note 155.
327 H.B. 20 § 120.103 (Tx. 2021); S.B. 7072 (Fl. 2021).
challenges to company decisions, creating a government-run mechanism resembling the Board. Thus, the idealized due process model of governance which the Board epitomizes can be found in many contexts, and so the Board experiment holds important lessons for these other experiments in platform governance. This Part turns to those lessons.

First, I give a realist account of the Board’s incentives to show how institutional incentives have played an underappreciated role in the Board’s work. The starting point is understanding that the Board is a servant of two masters: its creator, Meta, at whose mercy it exercises any power at all, and the public, whose approval it was created to court. The Board’s precarious status depends on it walking a tightrope between these two sets of interests. The Board’s decisions are therefore never simply a judgment about the merits of any individual case—they are always, consciously or not, also a political judgment of how best to manage the (often conflicting) interests of these stakeholders. This leads the Board to prioritize bolstering its own reputation over actual impact.

Second, I examine the political environment that the Board operates within to ask why this has incentivized the outcomes that it has. Often, the relief that someone is doing something has overtaken the need to carefully examine exactly what that something is, and allowed the Board to reap rewards for pursuing a formalistic style of governance.

Finally, I suggest that what this shows is that legitimacy is an ambiguous indicator of successful governance. Sociological legitimacy is sought after by those in power, but is not a standalone good. At best, it is something that can be harnessed to enact substantial reforms. At worst, it can be used to breed complacency about the need for such reforms altogether. Accordingly, we should be far more skeptical about generalized calls for more “legitimate” content moderation governance and instead articulate more specific normative ends.

A. The Importance of Institutional Incentives

The Board’s institutional incentives have played an underappreciated role in dictating the Board’s work and tempering its material impact. For the Board, as a self-regulatory institution with no legal or other independent

331 This subpart will discuss the Board and its members somewhat as a monolith. Of course, different members will have different values, incentives and goals. This subpart does not engage in individual speculation, but is interested in more general and generalizable dynamics.
authority, and no sword or purse of its own, its power lies in establishing sociological legitimacy. But the fact that the Board is a creature of self-regulation makes its goal of establishing its legitimacy more challenging, because it was designed by and continues to exist only at the whims of the actor it was created to constrain. Of course, Meta wants the Board to succeed—it did not establish the Board to fail. But it wants the Board to succeed in very particular ways, and this skews the Board’s work.

Meta and the Board’s interests are aligned to an extent. Both want the experiment to work, and both want the Board’s influence to grow. But Meta’s goodwill towards the Board will surely not be infinite. Ultimately, Meta only cares about the Board’s legitimacy and influence to the extent that it redounds to Meta’s own benefit. The goodwill of the regulated entity would not necessarily matter as much in most oversight relationships—but it matters a great deal when the oversight institution exercises power at the grace of the entity being overseen.

It is therefore important to remember Meta’s actual interests in the Board experiment. Meta did not create the Board in order to get yet another form of substantive input on its content moderation decisions, but in order to shift public perception. The substantive work of the Board is of little benefit to Meta, who regularly consults with stakeholders to get the perspective of experts and affected communities. The distinctive feature of the Board is not the nature of the expertise that the Board might provide, but the fact that provides that feedback in a public-facing way. It is the performance of oversight that is important to Meta. The return Meta wants for its $280 million investment is enhanced perceived legitimacy of its community standards, and reduced costs, both commercial and reputational, of making unpopular decisions. Meta’s support—and obedience—of the Board therefore is contingent on these returns being greater than the costs of complying with the Board. If that compliance becomes too costly, there is no way of forcing Meta to continue to pay attention to the Board.

This therefore makes public approval extremely important to the Board too, and not only for personal reputational reasons but also for institutional ones. The fact that Meta could start to ignore the Board tomorrow means that the Board’s power and continued existence depend on maintaining a baseline of public support. The Board therefore needs to constantly court public approval in order to increase the reputational costs to Meta of ignoring the

---

332 THE FEDERALIST No. 78 (Alexander Hamilton).
333 See supra Part I.
334 Douek, supra note 37; Kadri, supra note 37, at 189.
Board’s recommendations altogether.

But creates a predicament for the Board, because public faith in the Board relies on the Board looking like a meaningful accountability mechanism and not a mere sham. The Board therefore needs to look tough on Meta to garner public approval. But if the Board is too tough on Meta then Meta will simply ignore it, which would set off a downward spiral as it would make the Board look weak, which would cause the Board to lose public approval, which makes it easier for Meta to ignore it in future. Thus, to be successful, the Board must walk a fine line between appearing to bring Meta to heel while not getting too far out over its own skis. It must hold Meta accountable, but also know its own limits. As one Board member explained, “One of the dilemmas of the initial stages of the Board is the feeling that if we play too hardball, they will shut us up entirely. So we have to develop a collaborative modus operandi, we can’t be too tough, lest this experiment kind of fail on inception.”

Martin Shapiro explains the legitimacy paradox in the context of courts trying to rein in authoritarian governmental abuses in the following way: if a court challenges a regime to the extent that it gets ignored, the court loses its legitimacy, and if it doesn’t challenge the regime at all it also loses its legitimacy. But even “[i]f they manage things just right and maintain some perceived legitimacy, they lend that legitimacy to the authoritarian regime of which they are a part precisely because they are a part of it.” This means that the Board necessarily co-opted into bolstering Meta’s own legitimacy, not merely its own. And it also means that the Board cannot risk the weakness that comes with being too publicly controversial.

Ultimately, this explains the Board’s reluctance to meaningfully engage with the project of developing a normative framework for its decisions, focus on easy-to-measure metrics of success rather than necessarily meaningful ones, and pattern of side-stepping hard calls. The Board is not incentivized to provide answers, so much as it is incentivized to slap Meta over the wrists.

Take the development of a normative framework for thinking about the private regulation of speech in a rights-respecting and coherent way. Such a task is hard and will be controversial. As the Board is confronted with this task, there is little the Board can fall back on to justify its decisions beyond first principles, making it especially vulnerable to criticism for its own normative choices. Instead of showing its work, therefore, the Board simply presents many of its conclusions as compelled by IHRL, when in fact what IHRL would require in the context of most content moderation decisions is


337 Shapiro, supra note 78, at 334.
almost entirely under-specified and indeterminate. That is, the Board is making political choices but obscuring them. And it is deferring to another day the work of creating a principled framework that creates predictability and can reconcile the various tensions in its work.

This is a form of strategic avoidance—postponing the decision of contentious issues that could threaten the Board’s institutional viability—\textsuperscript{338}\—that is often used by newer institutions uncertain about their power. For such institutions, “[a]voidance is a means of cultivating … legitimacy. A [decisionmaker] can simply avoid deciding contentious, politically divisive issues that, by creating powerful opponents with the capacity to rein in (or oppose) the [decisionmaker]’s actions, could threaten its institutional legitimacy.”\textsuperscript{339} This also gives the decisionmaker time to shore up support for its authority and decisions before engaging in confrontation.\textsuperscript{340} The Board is reigning in the ambitions of its decisions and muting their political significance in order to bide its time and try establish a base of public support before becoming more confrontational.

This also explains the Board’s emphasis on doing so much public relations and institutional promotion—whether through its own corporate communications team and accounts, or whether through the regular appearance of members of the Board at various events with stakeholders in digital governance present, even as the Board’s case load continues to dwindle, and it keeps missing deadlines for its work. Convincing the public of the institution’s worth is as important to the Board as the actual work of governing.

The Board’s practice of public relations and strategic avoidance has so far seemingly served it well. Whether the Board could have achieved the power building that it has while being less avoidant is of course impossible to know. But as time goes on, such an approach gets riskier. If the Board continues to duck all the most difficult questions, and neglects its job of building a more robust normative framework for content moderation decisions, it will consign itself to irrelevance. Meta will no longer refer its most difficult decisions to the Board if it can expect the Board to simply keep dodging the hardest questions. That is, if the Board becomes “known to avoid politically divisive issues, it may lose its authority to decide controversial cases.”\textsuperscript{341} And Board’s practice of continually recommending more due process mechanisms in place of giving substantive answers to questions will be less popular in a less favorable economic environment in which many

\textsuperscript{338} Delaney, \textit{supra} note 241, at 4.
\textsuperscript{339} \textit{Id.} at 10.
\textsuperscript{340} Dixon & Issacharoff, \textit{supra} note 308, at 687.
\textsuperscript{341} Delaney, \textit{supra} note 241, at 15.
platforms are cutting back, not further investing, in trust and safety.\textsuperscript{342}

But perhaps most importantly, the public’s interest and trust in the institution should wane if it cannot more convincingly demonstrate its value-added. The failure to give compelling and comprehensive reasons for its decisions and build out its normative framework is a failure to live up to its potential. It undermines the predictability of the Board’s decisions, and their value as a guide for other decisionmakers. It also makes it increasingly hard to credit the Board members’ expertise and unique qualifications to be making the kinds of decisions they are tasked with if they do not display that expertise in the decisions themselves. And it maybe suggests the real limits of self-regulation in general. If the Board never feels emboldened enough to stand up to Meta or public opinion in high-stakes moments, it might be a sad indication of the constraints on what self-regulatory institutions might be able to achieve. Yet, despite all this, public interest and trust in the Board seems to be trending up, not down.

\textbf{B. The Revealed Preference for Performative Governance}

The institutional incentives to perform solemn governance without rustling too many feathers exist because the formalistic approach to governance is working. That is one of the most important lessons of the Board experiment—the reception of the Board suggests that the pursuit of legitimacy through procedural mechanisms can be effective.\textsuperscript{343} There is something about the court-like model that appeals to stakeholders. It invokes the familiar state-based model of speech governance, and no doubt benefits from legitimacy by association. It a focal point for people to express their outrage, while also giving and the appearance at least of finality once a decision is issued. It provides the show of Meta being reprimanded (“Oversight Board Criticizes Meta for Preferential Treatment”\textsuperscript{344} reads one headline; “Facebook Is Rebuked by Oversight Board Over Transparency on Treatment of Prominent Users”\textsuperscript{345} says another), even if the real-world costs

\begin{itemize}
  \item \textsuperscript{345} Schechner, \textit{supra} note 196.
\end{itemize}
to Meta of such a reprimand are low. And the implementation of recommendations means that something is being done to reform the current systems. These all seem to scratch the itch of dissatisfaction with the prior regime of Meta’s unilateral and completely unaccountable system of governance.

It’s important to be clear: the Board is not being performative in the way that its most trenchant critics warned when it was being established. It is not a “sham” or mere “spin.” It has not convinced lawmakers that legal regulation of Meta is no longer necessary, or prevented continued trenchant critique of Meta’s decisionmaking. The Board is meaningfully independent and is apparently prompting Meta to make changes to its systems, often of the kind civil society and others have been calling for from platforms for years, and indeed that regulators are drafting laws to bring about now. By making recommendations to Meta that bring about some of these reforms through self-regulation, the Board is governing, and in exactly the way that many stakeholders have asked for. It is receiving sociological legitimacy precisely because it is far less performative than many cynics predicted. Or, perhaps more specifically, it is performative in a way that is seen as legitimate.

C. Sociological Legitimacy as a Poor Marker of Success

Ultimately, the Board should be a reminder that legitimacy is not a standalone good, and social media governance should not be thought to be successful based simply on whether it attains sociological legitimacy amongst elite constituencies. This kind of legitimacy obviously matters to companies like Meta, but it does not translate into material benefits for others. Worse, focusing on legitimacy as the metric that needs to be maximized can divert attention for the need for other kinds of reforms. And, as in authoritarian regimes, a veneer of legitimacy can provide cover for

346 McNamee & Ressa, supra note 31.
348 For example, Meta did not seem to receive less critique than Twitter and YouTube for its decision to reinstate Donald Trump’s accounts, even though Meta went through the extensive process of getting the Board to review the decisions, while the other two companies made sudden and seemingly arbitrary decisions accompanied by little public reasoning.
349 Making this point in another context, see Ryan D. Doerfler & Samuel Moyn, Democratizing the Supreme Court, 109 CAL. L. REV. 1703, 1745 (2021) (“most approaches to legitimacy define it in terms of partisan neutrality rather than rights protection.”).
continued engagement with an otherwise objectionable institution. 351 What
this article has shown is that institutions can establish legitimacy even as they
neglect core parts of their job.

Ultimately, calls for “more legitimate” social media governance are both
accurate and too abstract to be useful as a guide for institutional behavior.
They respond to the understandable intuition that the prior regime of
unilateral power over content moderation is normatively undesirable, but
avoid the much harder and ultimately inherently contestable question of what
would be better. The lack of a specific vision of what content moderation
governance should be aiming for has created the conditions in which the
Board pursues the performance of governing as much, if not more, than its
substance.

In many ways this should not be surprising. This dynamic of the pursuit
of procedural reforms in the name of “good governance” or rights protection
and potentially at the expense of effective governance has been seen in areas
as diverse as administrative law 352 and international law. 353 But content
moderation governance should learn from those mistakes, not replicate them.

The point is not that there are some obviously correct criteria of good
governance that the Board should be using instead. The point is that the Board
was set up in order for debates about what those criteria should be to take
place, and it has instead eschewed those difficult questions or tried to forestall
such debate rather than encourage it. The Board has not defined the
constituency it is serving, or the yardsticks by which it measures its own
success. It highlights metrics in its public communications—like the number
of times the Board “overturns” Meta or how often Meta “takes action” based
on the Board’s recommendations—that are poor proxies for actual impact,
but good proxies for the kinds of things that attract headlines. It doesn’t
encourage debate or disagreement with its decisions, giving the briefest and
most abstract possible account of minority views or disagreement within the
Board itself in its decisions.

The Board experiment also makes clear that oversight institutions cannot
and do not operate in a vacuum, and respond to the incentives created by their
political environment. This means that setting up oversight institutions is only
the first step in creating an ecosystem of accountability for content
moderation. Watchdogs also need to be watched, and made to show their
work. Accountability and oversight need to be a dynamic and ongoing

351 Moustafa & Ginsburg, supra note 77, at 6.
352 Bagley, supra note 279.
353 David Kennedy, The International Human Rights Movement: Part of the Problem?,
movement to the legal formalization of rights and the establishment of legal machinery for
their implementation makes the achievement of these forms an end in itself”).
processes involving a broad base of stakeholders. The Board has not so far been subjected to the level of critical examination commensurate with its status and influence. Part of the problem is surely lack of competition—the Board currently has a monopoly as the only example of public-facing content moderation oversight and precedent-based decision-making. Perhaps the creation of other oversight institutions, such as those that will spring up to serve companies who need to offer recourse to an independent third-party arbiter under Europe’s DSA, will spur improvements. But part of the problem is also that content moderation discourse has always been more interested in spectacle than substance. For years, content moderation debates have focused on anecdata and individual decisions, rather than interrogating the underlying systems and the pragmatic trade-offs that such speech ecosystems entail. This is an environment that will naturally focus on headlines about the Board rapping Meta over the knuckles, rather than pushing for further answers about the normative assumptions and trade-offs that underpin the Board’s decisions, too. The Board could and should be a mechanism for changing that dynamic, rather than playing into it.

CONCLUSION

Has the Board improved content moderation governance? We might say, “compared to what?” To say that the Board is better than nothing, and achieved some things, is a start. But it is harder to evaluate whether what the Board has achieved is meaningful, whether it is commensurate with the level of investment and attention it has been given, or whether it is the best we can hope for from content moderation oversight. Importantly, the answers to those questions require much more critical engagement with the Board’s work than has generally been undertaken so far, paying much more attention to the politics and incentive structures of the Board itself, and demanding greater proof of work. When the Board was being created, the discourse was dominated with concerns about Meta’s incentives, and how to ensure that the Board was sufficiently insulated from its creator to be a meaningful accountability mechanism. But this Article has argued that the Board’s own institutional incentives and political environment are impacting its work in significant, and underappreciated, ways, and it is time to pay more attention to them.

The critiques in this Article of the Board should not therefore be understood as critiques of the Board alone. They are critiques of the larger discourse of content moderation which after first dismissing the Board entirely has come to assume its significance and accept its influence rather uncritically, while failing to ask enough questions about what, exactly, the Board is or should be trying to achieve.