Meta's Private Speech Governance and the Role of the Oversight Board: Lessons from the Board's First Decisions

Ruby O'Kane*

25 STAN. TECH. L. REV. 167 (2022)

ABSTRACT

The Oversight Board, Meta's project over three years in the making, began handing down its decisions in early 2021. Envisioned as a "Supreme Court"-type body for Meta's speech governance regime, this new institution has been plagued with doubts from its inception. Analysis of the Board's first decisions reveals its ingenuity, its potential, and its willingness to criticize its maker. However, its decisions also reveal the Board's institutional pitfalls. The Board is a key institutional player in the emerging private governance of online speech and speakers. It has, largely of its own volition, created a methodology for its decision-making based on International Human Rights Law (IHRL) and norms, particularly those around freedom of expression and limitations on this freedom according to the principles of proportionality, necessity, and legitimacy. However, this produces an internal incoherency within Meta's speech governance regime; the Board promulgates an IHRL-based rights adjudication framework, whereas Meta itself, through both human and automated moderation, adopts a probabilistic method, where rights abrogation is accepted as inevitable and built into content moderation processes and technologies. This internal tension, which has come to the fore through the Board's reasoning and decision-making, evinces the relatively constrained role the Board can play in mandating the improvement of Meta's corporate standards and practices in terms of freedom of expression outcomes, or in providing increased accountability and transparency outcomes for users and the broader public. The Board's institutional constraints serve to maintain the status quo of Meta's private governance regime, which serves to exclude certain speech and speakers,

*LLB (Hons I)/BInst (UNSW); Tipstaff, New South Wales Court of Appeal. I am grateful to the UNSW Faculty of Law & Justice, including Associate Professor Daniel Joyce for his helpful comments and guidance. Thank you to my family and friends for their unwavering support during the research and writing of this piece, and throughout my studies. I also thank the *Stanford Technology Law Review* editorial team for their comments and exceptional work editing this piece. Thank you to all the wonderful scholars in this field, whose passion inspired me to keep reading and writing. The opinions expressed in this publication are those of the author only. They do not purport to reflect the opinions or views of anyone else, including past or present employers.

marginalize certain voices and amplify others, mirroring and reinforcing the existing dynamics of social, political, and economic hierarchy and stratification which exist offline.

TABLE OF CONTENTS

I. INTRODUCTION

Almost two decades after its founding, Meta has become one of the most profitable and controversial tech giants in the world. It was in the wake of controversies fueled by the perceived exploitation of user data and revelations about Facebook's impact on electoral and democratic processes that the Oversight Board ("the Board") was born.[1] Although the Board's genesis lay outside Meta, it was quickly adopted and formalized by Meta's leadership in 2018, and became fully operational at the start of 2021.[2] From

[1] *See generally* Adrian Chen, *Cambridge Analytica and Our Lives Inside the Surveillance Machine*, NEW YORKER (Mar. 21, 2018), https://perma.cc/LQ6Q-2TPP; SELECT COMM. ON INTEL., 2 RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION: RUSSIA'S USE OF SOCIAL MEDIA WITH ADDITIONAL VIEWS, S. REP. NO. 116-290 (2019).

[2] Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2448-50 (2020). *See, e.g.*, Access Now et al., *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, https://perma.cc/8W6K-LGAK.

the Board's inception, Meta framed it as a way for users to hold the company accountable for its decisions and policymaking,[3] ensuring the "final judgment call" on what constitutes "acceptable speech" on its platforms is not at Meta's sole discretion, but reflects "the social norms and values" of its platforms' users.[4]

The idea of "final judgment calls" elides the true nature of adjudication of online speech by social media companies, which involves increasingly complex self-regulatory frameworks and behemoth systems of governance. Platforms, straddling and unimpeded by national borders, operate across various jurisdictions with no clear overlord or master.[5] Filling this so-called "governance gap"[6] in the online realm, platforms create their own systems of private speech governance.[7] Platforms conduct governance projects through terms of service and "platform law,"[8] a combination of community standards, rules, and policies.[9] Although as a matter of *law*, terms of service are mere contract, they perform an important governance function by articulating rights and responsibilities of both users and platforms and establish a balance of powers and rights.[10] Speech is reviewed by automation or by a "privatized bureaucracy" of human content moderators[11] at various points in the content cycle: ex ante (between content upload and publication), ex post proactive (proactively removing),

---

[3] Klonick, *supra* note 2, at 2427, 2446. For an excellent summary of the conceptualization and creation of the Board, see *id.* at 2448-65; see also evelyn douek, *Facebook's "Oversight Board": Move Fast with Stable Infrastructure and Humility*, 21 N.C. J. L. & TECH. 1, 6 (2019).

[4] Ezra Klein, *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next*, VOX (Apr. 2, 2018, 6:00 AM EDT), https://perma.cc/HBU9-VLPC.

[5] Agnes Callamard, *The Human Rights Obligations of Non-State Actors*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 191, 215 (Rikke Frank Jørgensen ed., 2019).

[6] EMILY B. LAIDLAW, REGULATING SPEECH IN CYBERSPACE: GATEKEEPERS, HUMAN RIGHTS AND CORPORATE RESPONSIBILITY 95 (2015); Rikke Frank Jørgensen, *Rights Talk: In the Kingdom of Online Giants*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS, *supra* note 5 at 163, 182.

[7] Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2012-15 (2018).

[8] David Kaye (Special Rapateur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Op. and Expression*, U.N. Doc. A/74/486 (Oct. 9, 2019), https://perma.cc/T9TH-RBHX.

[9] Balkin, *supra* note 7, at 2021. *See also Community Standards*, META, https://perma.cc/8MWY-7SZ3.

[10] *See* Edoardo Celeste, *Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?*, 33 INT'L REV. L. COMPUT. & TECH. 122, 138 (2019); Nicolas P. Suzor, *The Responsibilities of Platforms: A New Constitutionalism to Promote the Legitimacy of Decentralized Governance*, ASSOC. OF INTERNET RESEARCHERS ANN. CONF. (Oct. 2016), https://perma.cc/5XDY-9GVG; Nicolas Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, 4 SOC. MEDIA + SOC'Y 1, 1-11 (2018).

[11] Balkin, *supra* note 7, at 2028.

ex post reactive (following flagging by users).[12] Additionally, particularly controversial or high profile cases are escalated through the "tiers" of internal review teams to the company's executives.[13] Rules of private governance are also built in, and expressed through, code and algorithms that determine how speech is organized and presented to users;[14] a process that is mirrored in the enforcement stage of content moderation through automated review, flagging, and removal of content.

Freedom of expression was previously characterized as a dualist model of speech governance centered around the "classical struggle between government censorship and citizens."[15] Today, platforms' ability to govern speech by determining what can and cannot be expressed on their platform and by implementing this through complex technological and human enforcement mechanisms, has led them to emerge as the "New Governors" of online speech.[16] These new systems of governance, invariably informed by the behemoth scale of platforms in terms of users and content posted, pose risks to freedom of expression online. The sheer size of the content moderation project for platforms creates problems of over-enforcement or over-blocking (filtering and taking down otherwise permissible content that might create backlash or be controversial to "err on the side of caution"[17]) or under-enforcement (not removing offending content because it is missed by automation or human moderators).[18]

Moreover, this regulation is characterized by significant accountability and transparency deficits.[19] The values to which platforms are "beholden"—namely, profit, and the promotion of user engagement—have provided only very indirect sources of accountability to users.[20] Platforms

---

[12] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1599, 1635-39 (2018).

[13] *Id.* at 1639-41. *See also* MATTHIAS C. KETTEMANN & WOLFGANG SCHULZ, SETTING RULES FOR 2.7 BILLION: A (FIRST) LOOK INTO FACEBOOK'S NORM-MAKING SYSTEM: RESULTS OF A PILOT STUDY (Hans-Bredow Institut 2020); *Corporate Human Rights Policy*, FACEBOOK, https://perma.cc/Z3T3-JAR7. For an extensive summary of Meta's content moderation process, see Klonick, *supra* note 2, at 2428-35. For a summary of the appeals process, see *Appealed Content*, META, https://perma.cc/3GP3-KEVF.

[14] Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939, 946 (2020).

[15] Klonick, *supra* note 2, at 2446. *See generally* Jack M. Balkin, *Old School/New School Speech Regulation*, 127 HARV. L. REV. 2296 (2014); Balkin, *supra* note 7; Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149 (2018).

[16] Klonick, *supra* note 12, at 1663.

[17] evelyn douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality & Probability*, 121 COLUM. L. REV. 759, 827 (2020); Balkin, *supra* note 7, at 2017.

[18] douek, *supra* note 17, at 809.

[19] Klonick, *supra* note 12, at 1666. *See generally* Suzor, *Digital Constitutionalism, supra* note 10.

[20] Klonick, *supra* note 12, at 1666.

have typically provided limited transparency around how they create and enforce their rules or adjudicate complaints and hard cases.[21] Further, platform self-regulation is not "wholly public spirited."[22] The "economic logic of advertiser-driven social media," requires continuous expansion of either membership or user attention often measured as time spent on the platform.[23] This dictates that platforms constantly search for "new opportunities for profits and property accumulation that can only be achieved through shutting down or circumscribing" speech.[24] Self-regulation creates opaque normative systems of speech governance that are informed by the idiosyncratic business interests and models of platforms, yet are presented as fundamentally based in broader social values and norms.[25]

The Board's establishment reflects changing understandings of the private governance of online speech, including the potential for such institutions to provide an avenue for appeal for users that lies outside the bureaucracy of the platform itself. Such institutions can also remedy some of the above deficits and provide a form of public reasoning, accountability, and transparency, lending legitimacy to these private governance regimes. Meta framed the creation of the Board as a delegation of Meta's decision-making authority for important, difficult cases to an independent body who could impartially adjudge content in a way that "reflects the social norms and values of people all around the world."[26] The reality is a more complex landscape of interplays between Meta and the Board, as well as between the Board and the broader normative frameworks operating to structure and inform new theories about how content moderation should be conducted.

The Board sits within an integrative institutional framework. Its budget, appointment and removal of members, and administration is run by a non-charitable trust: "The Oversight Board Trust," with trustees appointed by Meta. The Trust received funding of $130 million as an irrevocable grant, designed to fund the operation of the Board for six years.[27] The Board itself

---

[21] *Id.* at 1668; Jørgensen, *supra* note 6, at 166.
[22] Balkin, *supra* note 7, at 2023.
[23] *Id.* For more about Meta's adoption of this targeted advertising business model, see generally SHEERA FRENKEL & CECILIA KANG, AN UGLY TRUTH: INSIDE FACEBOOK'S BATTLE FOR DOMINATION (2021).
[24] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U.L. REV. 1, 14 (2004).
[25] David Kaye, *supra* note 8, at 14.
[26] Klein, *supra* note 4; Klonick, *supra* note 2, at 2449-51.
[27] META, OVERSIGHT BOARD TRUST AGREEMENT, § 2.2 (2019), https://perma.cc/P7AU-MLYG; Brent Harris, *An Update on Building a Global Oversight Board*, META (Dec. 12, 2019),

is a limited liability company (the Oversight Board LLC), of which the Trust is the sole member, formed by the Trust for the purpose of establishing and ensuring the operation of the Board, including managing finances, and appointing and removing Board Members.[28] The Board consists of a minimum of eleven members, (there are currently twenty seated members and an envisioned future of forty members), each serving for a maximum of three terms or nine years.[29] Members must be familiar with issues concerning "digital content and governance," and many members specialize in law, human rights, journalism, or technology.[30]

The Board has authority to produce binding decisions regarding either specific content that Meta has removed or content that Meta has determined should remain online, following either a request from a user who has exhausted the internal appeals processes or a referral from Meta.[31] The Board also has a separate jurisdiction, either when ancillary to a particular case or upon Meta's request, to give "policy guidance" regarding Meta's policies. These do not appear to fall within the Board's binding powers: its power to instruct Meta to allow or remove content or to uphold or reverse a decision enforcing its policies.[32] Case selection, deliberation processes, and production of final decisions are at the Board's discretion, beyond the Charter's requirement that the Board choose cases "that have the greatest potential to guide future decisions and policies." These selection processes are outlined in the recently updated Bylaws.[33]

Scholars have begun to analyze the Board's Charter and institutional set-up, including its level of independence from Meta and implications for procedural fairness and transparency.[34] However, since it began handing down decisions in January 2021, few have had the opportunity to consider the Board's decisions and the institutional interplay between Meta and the now-operational Board.[35] This piece will consider the Board's approach to

---

https://perma.cc/J32R-L6UG. *See also* META, OVERSIGHT BOARD CHARTER, art. 5, §§ 1-2 (2019), https://perma.cc/F84P-ZWRS; Klonick, *supra* note 2, at 2467-68.

[28] *See* META, OVERSIGHT BOARD TRUST AGREEMENT, *supra* note 27, at § 2.2; META, OVERSIGHT BOARD LLC LIMITED LIABILITY COMPANY AGREEMENT, art. 2, § 2.2, art.5, § 3, art.7 (2019), https://perma.cc/M98N-9E5L.

[29] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, §§ 1, 3.

[30] *Id.* at art. 1, § 2; *Meet the Board*, OVERSIGHT BD., https://perma.cc/R35S-LUN4 (last visited May 3, 2022).

[31] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 1; Klonick, *supra* note 2, at 2463.

[32] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 4; *see also id.* at art. 3, § 7.3.

[33] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 2, § 1; META, OVERSIGHT BOARD BYLAWS, at art. 1, § 3 (2021), https://perma.cc/6Y65-RJP3.

[34] *See* Klonick, *supra* note 2; douek, *supra* note 3.

[35] douek has considered similar questions in blog and short-form formats. *See, e.g.*, evelyn douek, *How Much Power Did Facebook Give Its Oversight Board?*, LAWFARE (Sept. 25, 2019, 8:47 PM), https://perma.cc/F8G7-LB5Q; evelyn douek, *Facebook's*

online speech governance and the methodology it has adopted in adjudicating appeals and providing policy advice. This piece will analyze the Board's role within Meta's existing speech governance system and its ability to address some of the key concerns raised with private governance of online speech. It will argue that the role and benefits of the Board, although numerous and broader in scope than even Meta anticipated, are undermined by the design and operation of Meta's platforms and by the distribution of policy and decision-making power among different actors within the speech governance system. Although the Board has thus far proved itself a force to be reckoned with, there are systemic and institutional barriers preventing it from providing users with a meaningful appeal or remedial avenue to enforce their freedom of expression rights and protect their participation in online speech on Meta's platforms.

Part II will explore the normative frameworks guiding Meta's speech governance. The Board has adopted an International Human Rights Law ("IHRL") framework from its earliest decisions, judging Meta's actions according to obligations arising under the United Nations' Guiding Principles on Business and Human Rights ("UNGPs") and the balancing methodology outlined in International Covenant on Civil and Political Rights ("ICCPR") Article 19. This differs significantly from the normative frameworks guiding speech governance at other levels of Meta's bureaucracy. Speech governance is guided by probabilism at Meta's "lower" bureaucratic levels: human "crowdworker" moderators and automated enforcement.[36] The sheer number of Facebook users forces speech adjudication to be an algorithmically-moderated and statistically-gauged process with the goal of reducing errors to an "acceptable" rate.[37] This scale produces conflicting normative frameworks within Meta's speech governance system: an individualistic human rights-based approach propagated and implemented by the Board and a probabilistic systemic approach underpinning the majority of enforcement and decision-making.

Part III will examine the implications of this discord and how it reflects an emerging trend of interaction between the Board and Meta. The Board has asserted itself within Meta's policy sphere, attempting to use its

---

Responses in the Trump Case Are Better Than a Kick in the Teeth, but Not Much, LAWFARE (June 4, 2021, 4:32 PM), https://perma.cc/Z2JM-K3T2; evelyn douek, *The Oversight Board Moment You Should've Been Waiting For: Facebook Responds to the First Set of Decisions*, LAWFARE (Feb. 26, 2021, 1:00 PM), https://perma.cc/3LFR-H2JM [hereinafter douek, *Facebook Responds to the First Set of Decisions*]. Oversight Board decisions and Meta responses up to July 2021 are considered in this paper.

[36] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 121 (2018); Klonick, *supra* note 12, at 1634-41; douek, *supra* note 17, at 793.

[37] douek, *supra* note 17.

advisory jurisdiction to improve policymaking, better align policies with human rights standards, and mitigate accountability and transparency deficits. However, as this piece will argue, Meta concurrently attempts to limit the Board's ability to improve the policymaking, design, and operation of Meta's speech governance system. It does so by limiting the Board's jurisdiction and scope, restricting both its binding jurisdiction and its review of algorithmic decision-making, curtailing its focus to politically sensitive cases, and limiting its size. Meta also restricts the Board by refusing to cooperate with the Board's requests for information and by mistranslating the Board's policy recommendations. The Board's contributions are thus restricted to indirect accountability and limited transparency benefits, thereby providing very limited opportunity for users to enforce their freedom of expression rights that are otherwise unprotected by Meta's "platform law."[38]

Yet, the design and technologies underpinning Facebook as a platform and product, Meta's content moderation and curation, [39] and their advertising business model further restrict platform users' speech rights by systemically excluding and silencing certain speech and speakers. While over- and under-enforcement, erroneous removal, deplatforming, and other errors are portrayed as neutral occurrences, Meta's probabilistic speech governance is far from neutral. It is characterized by traditional socio-politico-economic hierarchies and dynamics of exclusion and inclusion that undermine the very benefits the internet and platforms were supposed to provide: equal accessible participation in speech.

II. ENUNCIATION AND ENFORCEMENT OF SPEECH NORMS

Online speech governance is an "iterative," "law-making," and "norm-generating" process, not a value-neutral one.[40] This Part will explore the normative frameworks adopted by both the Board and Meta in their respective speech adjudication, as well as how these frameworks interact within Meta's broader speech governance system. As a "governor" of its online space,[41] Meta regulates speech on Facebook explicitly (through the creation and enforcement of its Community Standards and policies) and implicitly (through rules and norms "baked into" the code and algorithms constituting the platform).[42] The Board also engages in this governance

[38] David Kaye, *supra* note 8, at 3, 19.
[39] GILLESPIE, *supra* note 36, at 41; Tarleton Gillespie, *Governance of and by Platforms*, *in* SAGE HANDBOOK OF SOCIAL MEDIA 254, 254 (Jean Burgess et al. eds., 2017).
[40] Klonick, *supra* note 12, at 1663.
[41] *Id.*
[42] Mike Ananny, *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of*

process through its decisions and has chosen rather emphatically to adopt an IHRL framework. However, this IHRL framework differs fundamentally from Meta's normative governance framing. Meta adopts a "systemic balancing" of users' speech characterized by a probabilistic, statistical understanding of speech adjudication. This creates two diverging frameworks for speech governance, where Meta's system-focused approach conflicts with the Board's understanding of individual speech rights.

A. IHRL Framework

1. Content Moderation as IHRL

An IHRL approach to speech adjudication involves balancing individual speech rights with broader public interests, safety, and the rights and interests of other groups or individuals. IHRL as it relates to private corporations is embodied in the UNGPs.[43] Companies must respect human rights, mitigate adverse human rights impacts, and provide remedial mechanisms for aggrieved users. The UNGPs do not reflect "mere social expectation" but are part of a soft-law "system of public governance."[44] Platforms must "establish principles of due diligence, transparency, accountability and remediation that limit platform interference with human rights through product and policy development."[45] In March 2021, Meta

Online Expression and Platform Governance, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://perma.cc/T8NW-LJA8; GILLESPIE, *supra* note 36; Jørgensen, *supra* note 6, at 181.

[43] U.N. Working Grp. on Bus. and Hum. Rts., *The UN Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect, Remedy" Framework* (2011), https://perma.cc/G7C3-DHRZ. In particular, it is embodied in Pillar II 'The Corporate Responsibility to Respect Human Rights' and Pillar III 'Access to Remedy.' Pillar II consists of Guiding Principles 11 to 24 and requires businesses to respect human rights (at a minimum those expressed in the International Bill of Human Rights and principles relating to fundamental rights set out in the International Organisation of Labour's Declaration on Fundamental Principles and Rights at Work) in two ways: businesses 'should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.' *Id.* at 13. Pillar III requires states to facilitate state and non-state based mechanisms for the airing of grievances and remedy of business-related human rights abuses or harms. *Id.* at 30-31.

[44] LAIDLAW, *supra* note 6, at 91-92. *See also* Justine Nolan, *The Corporate Responsibility to Respect Human Rights: Soft Law or Not Law?*, *in* HUMAN RIGHTS OBLIGATIONS OF BUSINESS: BEYOND THE CORPORATE RESPONSIBILITY TO RESPECT 138-61 (Surya Deva & David Bilchitz eds., 2013); Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, 38 YALE J. ON REGUL. 86 (2020). *See generally* Jen Patja Howell, *The Lawfare Podcast: The Arrival of International Human Rights Law in Content Moderation*, LAWFARE, at 20:00-28:30 (May 27, 2021, 5:01 AM), https://perma.cc/M2UY-HFLW.

[45] David Kaye, *supra* note 8, at 14.

launched a new "Corporate Human Rights Policy" where it committed to respecting human rights according to the UNGPs.[46]

As U.N. Special Rapporteur, David Kaye proffered a framework of standards for content moderation based on the UNGPs.[47] First, platforms must respect human rights "by default."[48] IHRL is to be incorporated directly into internal "platform law." Platforms must make policy commitments aligning with IHRL, which should govern all decisions and policymaking. Second, where platforms restrict user expression, this limitation must be justified under the principles of legality, legitimacy, necessity, and proportionality.[49] Platform rules must be clear and specific, and any action taken must be in accordance with these rules. Platforms must show how enforcement actions taken correspond to "narrowly tailoring restrictions" and such actions must be the least intrusive option available at the time.[50] Any restrictive action must be proportionate to the burden on the user's speech rights, and platforms must be transparent about the factors they take into account in determining restrictive action.[51] This mirrors the requirements established by Article 19 of the ICCPR, which provides for the right to freedom of expression with only restrictions as provided by law and necessary for the rights of others, public safety, or national security.[52] Kaye's framework provides a clear procedure for the balancing of rights, and ensures any limits on freedom of expression are based on specific enumerated circumstances that do not include private interest or profit. IHRL is therefore a way to "re-align the private incentives of platform governance with the broader interest."[53]

---

[46] *Corporate Human Rights Policy*, *supra* note 13; Miranda Sissons, *Our Commitment to Human Rights*, META (Mar. 16, 2021), https://perma.cc/MJ67-ZLR5.
[47] David Kaye, *supra* note 8, at 15.
[48] *Id.* at 16.
[49] *Id.* at 15-16.
[50] *Id.* at 16.
[51] *Id. See also* U.N. Hum. Rts. Comm., *General Comment No. 34: Freedoms of Opinion and Expression*, ¶ 7, U.N. Doc. CCPR/C/GC/4 (Sept. 12, 2011), https://perma.cc/QZ9Z-MYAP.
[52] International Covenant on Civil and Political Rights, arts. 19(2), 19(3), Dec. 16, 1966, 999 U.N.T.S. 171, https://perma.cc/LWP7-ZTJ2. *See* David Kaye (Special Rapateur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Op. and Expression*, ¶ 45, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) ("Companies should incorporate directly into their terms of service and 'community standards' relevant principles of human rights law that ensure content-related actions will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression.").
[53] Sander, *supra* note 14, at 942. Various scholars have outlined the benefits, difficulties and (non)viability of a human rights-based approach to online speech governance. I have chosen to focus on the implementation of the framework and will not address these arguments. *See generally* Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26-70 (2018); evelyn douek, *The Limits of International*

2. Oversight Board as IHRL Exemplar

The Board has emphatically adopted this IHRL framework, largely of its own volition. Klonick, one of the first to conceptualize content moderation as private governance of online speech, presciently noted that, although the Board's Charter does not reflect a "broad or robust adoption of human rights," the Board could have powerful benefits if it "creates a meaningful method for users to enforce such rights."[54] This is precisely what the Board has done. There was a narrow opening created by the language of the Charter. Article 2 dictates that the Board "pay particular attention to the impact of removing content in light of human rights norms protecting free speech." [55] The introduction recognizes freedom of expression as a "fundamental right."[56] Based on these references, the Board has created a reasoning method centered around Kaye's IHRL framework and ICCPR Article 19. It now considers adjudication on whether Meta's decisions "fall within the zone of what the U.N. Guiding Principles require" as its "principal task."[57]

The Board's framework for reviewing decisions, first set out in their second decision,[58] considers the validity of actions taken by Meta according to Meta's Community Standards, Values, and "Relevant Human Rights Standards." In relation to human rights standards, the Board first notes that the UNGPs "establish a voluntary framework for human rights responsibilities of private businesses," and second, considering these UNGPs, lists the relevant standards to be considered (including Article 19).[59] In applying these standards, the Board considers first whether the content was "subject to a mandatory restriction under international human rights law" (e.g., Article 20(2) regarding incitement to discrimination). The Board then judges Meta's actions according to the ICCPR Article 19(3) framework for restricting speech, considering the requirements of legality, legitimate

---

Law in Content Moderation, 6 U.C. IRVINE J. INT'L TRANSNAT'L & COMPAR. L. 37 (2021); Michael Lwin, *Applying International Human Rights Law for Use by Facebook*, 38 YALE J. ON REGUL. 53 (2020); LAIDLAW, *supra* note 6; Benesch, *supra* note 44; Callamard, *supra* note 5, at 214-218.

[54] Klonick, *supra* note 2, at 2478.

[55] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 2, § 2.

[56] *Id.* at Introduction.

[57] OVERSIGHT BD., META, CASE DECISION 2020-003-FB-UA (Jan. 28, 2021) [hereinafter NAGORNO-KARABAKH CASE], https://perma.cc/588C-P9R2.

[58] OVERSIGHT BD., META, CASE DECISION 2020-002-FB-UA (Jan. 28, 2021) [hereinafter MYANMAR HATE SPEECH CASE], https://perma.cc/GCX2-PRA4.

[59] *Id.* For empirical analysis of human rights documents and standards considered by the Board, see OVERSIGHT BOARD TRANSPARENCY REPORTS - Q4 2020, Q1 & Q2 2021, at 35-36, 58-60 (2021), https://perma.cc/5PCV-GNTJ.

aim, necessity, and proportionality.[60] The Board has consistently adopted this structure in all its decisions to date.

This framework is the same as the framework promulgated by Kaye based on the UNGPs and ICCPR: judging Meta's speech decision-making and policies according to the principles of legality, necessity, and proportionality.[61] Kaye's framework is emblematic and at the very center of this new concept of IHRL-based speech governance.[62] The Board echoes Kaye's assertion that, although companies are not nation-states, given their impact on speech and expression, the UNGPs duly provide a framework for applying human rights standards to platforms.[63] Indeed, by referencing the UNHRC endorsement of the UNGPs, the Board appears to acknowledge the UNGP's normative weight.[64] In a more recent decision, the Board explicitly rejected applying only First Amendment rights as the central standard for its adjudicative framework or "defer[ring] to American law,"[65] instead favoring international standards, which the Board notes are informed in part by the First Amendment. This reflects an emphatic and explicit adoption of IHRL standards as their chosen framework.

Although the Charter references freedom of expression and echoes rhetoric of proportionality, the Board's choice of international human rights norms, particularly over a strict First Amendment approach, was by no means an obvious choice. Meta's explicit acceptance of the UNGPs in its first publicly released corporate policy (March 2021) post-dates the Board's first decisions, in which it first used the UNGPs as the basis for this human rights-centric framework.[66] Reference to freedom of expression as a "fundamental right" was preambulatory—the Charter does not specifically grant this right to users.[67] The Charter referred merely to "human rights norms" without referring to international standards or specific norms, and when referring to rights, it in fact uses First Amendment-specific language ("free speech" rather than "freedom of expression"), which could have lent itself to the adoption of these US norms as the central normative framework by the Board. This may have even been the logical choice given

---

[60] MYANMAR HATE SPEECH CASE, *supra* note 58.
[61] David Kaye, *supra* note 8, ¶¶ 46-52.
[62] *See* douek, *supra* note 53, at 43.
[63] David Kaye, *supra* note 8, ¶¶ 41, 45; OVERSIGHT BD., META, CASE DECISION 2021-001-FB-FBR (May 5, 2021) [hereinafter TRUMP CASE], https://perma.cc/6M4M-L6H7. *See also* NAGORNO-KARABAKH CASE, *supra* note 57.
[64] MYANMAR HATE SPEECH CASE, *supra* note 58. *See* U.N. Hum. Rts. Council, *Elaboration of an International Legally Binding Instrument on Transnational Corporations and Other Business Enterprises with Respect to Human Rights*, U.N. Doc. A/HRC/Res/26/9 (July 14, 2014), https://perma.cc/6TGS-GWEV.
[65] TRUMP CASE, *supra* note 63.
[66] *See* Access Now, *After Nearly Two Decades in the Dark, Facebook Releases Its Human Rights Policy*, ACCESS NOW (Mar. 16, 2021, 5:31 PM), https://perma.cc/RHL9-K5CQ.
[67] Klonick, *supra* note 2, at 2478.

the longstanding cultural and ideological commitment to First Amendment norms among Meta's policymakers and executives, as elucidated by Klonick.[68] The Charter notes that "any limits [on expression] should be based on specific values,"[69] although Meta's values seem to be based only nominally in IHRL. The Charter notes the Board should "pay particular attention" to human rights impacts but does not specify a particular methodology. The Board has arguably gone beyond this standard to a much more substantive adoption of human rights framework than envisioned by the Charter.

3. Meta's "Systemic Balancing" and IHRL

Unlike the Board, Meta's speech governance is not guided by an IHRL framework. Meta has transitioned away from its original "post-as-trumps" approach: its general reluctance to interfere with user content under a First Amendment-based "classic libertarian ethos"[70] of limited interference with individual speech.[71]

Various controversies, including Zuckerberg's decisions not to remove Holocaust deniers and right-wing conspiracy theorists, created widespread public outrage.[72] Meta's self-presentation as a neutral speech intermediary came to be viewed as a "myth."[73] This social and cultural reckoning precipitated an internal reckoning which contributed to the adoption of new decision-making methods and new rhetoric around decisions.[74] Meta's terms of service, Community Standards, and the Board's Charter are all key governing documents establishing an allocation and balancing of rights and

---

[68] Klonick, *supra* note 12, at 1621.

[69] META, OVERSIGHT BOARD CHARTER, *supra* note 27. *See also* KETTEMANN & SCHULZ, *supra* note 13, at 19.

[70] Jonathan L. Zittrain, *Three Eras of Digital Governance*, *in* TOWARDS A GLOBAL FRAMEWORK FOR CYBER PEACE AND DIGITAL COOPERATION: AN AGENDA FOR THE 2020S, at 115, 115 (Wolfgang Kleinwachter et al. eds., 2019).

[71] douek, *supra* note 17, at 777-78; Klonick, *supra* note 12, at 1663.

[72] Kara Swisher, *Zuckerberg: The Recode Interview*, VOX (Oct. 8, 2018, 2:21 PM EDT), https://perma.cc/WQD8-UBXU; Ezra Klein, *The Controversy Over Mark Zuckerberg's Comments on Holocaust Denial, Explained*, VOX (Apr. 2, 2018, 6:00 AM EDT), https://perma.cc/HBU9-VLPC; Ryan Mac & Craig Silverman, *"Mark Changed The Rules": How Facebook Went Easy On Alex Jones And Other Right-Wing Figures*, BUZZFEED NEWS (updated Feb. 22, 2021, 10:14 AM), https://perma.cc/RWN7-2Y2F; FRENKEL & KANG, *supra* note 23, at ch. 10.

[73] douek, *supra* note 17, at 777. GILLESPIE, *supra* note 36, at 40, 208. *See also* Anupam Chander & Vivek Krishnamurthy, *The Myth of Platform Neutrality*, 2 GEO. L. TECH. REV. 400, 403-409 (2018).

[74] Mary Anne Franks, *The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?*, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://perma.cc/MHR9-WRUW.

responsibilities of both users and platforms.[75] Amendments to these key documents reflect Meta's transition towards a "systemic balancing" approach to speech governance, under which "rules are written to encompass multiple interests."[76] Meta's updated Values[77] called "Voice," which it vaguely (but non-explicitly) tied to freedom of expression,[78] as Meta's paramount value. However, Voice "should be limited for reasons of authenticity, safety, privacy, and dignity."[79] Meta's Community Standards also now invoke balancing in the context of permitting content which is newsworthy or in the public interest.[80] The Charter also echoes this, noting that limits on speech should be based on Meta's Values.[81] These references show "elements of a traditional proportionality test" embodied in Meta's governing documents.[82]

However, this adoption of a systemic balancing framework does not equate to an IHRL-based approach. Meta has not implemented the human rights by default aspect of Kaye's model.[83] Meta looks for guidance in human rights law standards and documents.[84] The extent to which this has been built into formal governance and decision-making procedures is unclear.[85] Meta's new Corporate Human Rights Policy notes the escalation of "significant and challenging matters" to executives,[86] but there is no requirement to consider human rights in this process. IHRL is generally invoked by executives as an ad-hoc, post-facto justificatory tool, rather than

---

[75] *See* Suzor, *The Responsibilities of Platforms*, *supra* note 10; Celeste, *supra* note 10; Suzor, *Digital Constitutionalism*, *supra* note 10.

[76] douek, *supra* note 17, at 763.

[77] These "Values" are contained in the preamble to Facebook's Community Standards, and are the lens through which the Community Standards should be read and enforced. Monika Bickert, *Updating the Values That Inform Our Community Standards*, META (Sept. 12, 2019), https://perma.cc/D6XR-MKU7.

[78] KETTEMANN & SCHULZ, *supra* note 13, at 19.

[79] douek, *supra* note 17, at 765; Bickert, *supra* note 77. *See also* evelyn douek, *Why Facebook's "Values" Update Matters*, LAWFARE (Sept. 16, 2019, 12:05 PM), https://perma.cc/2WU5-6JGY.

[80] *Corporate Human Rights Policy*, *supra* note 13.

[81] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction.

[82] KETTEMANN & SCHULZ, *supra* note 13, at 20.

[83] David Kaye, *supra* note 8, ¶ 45.

[84] Richard Allan, *Hard Questions: Where Do We Draw the Line on Free Expression?*, META (Aug. 9, 2018), https://perma.cc/7REN-VRYK. *See also* Lwin, *supra* note 53, at 59; KETTEMANN & SCHULZ, *supra* note 13, at 20; *Corporate Human Rights Policy*, *supra* note 13.

[85] Lwin, *supra* note 53, at 55.

[86] Specifically, the policy calls for escalation to the VP for Global Affairs and Communications and General Counsel, with input from Chief Diversity Officer, COO, and CEO. *Appealed Content*, *supra* note 13, § 5. Anecdotal evidence also supports this. *See, e.g.*, Elizabeth Dwoskin & Nitasha Tiku, *Facebook Employees Said They Were 'Caught in an Abusive Relationship' with Trump as Internal Debates Raged*, WASH. POST (June 6, 2020), https://perma.cc/BZ77-34S9; Elizabeth Dwoskin & Robert Costa, *Facebook Chief Mark Zuckerberg Reached out to Speaker Pelosi. She Hasn't Called Him Back.*, WASH. POST (June 11, 2019), https://perma.cc/LY7Z-8DSH; FRENKEL & KANG, *supra* note 23, at 234-40.

as a routine part of decision-making processes.[87] An independent civil rights audit found that Zuckerberg "elevated a selective view of free expression" and has consistently prioritized this over other values such as non-discrimination.[88] Executive decision-making therefore does not appear to be characterized by a proceduralized and consistent consideration and balancing of rights as per IHRL, but considered on an ad hoc basis, informed by decision-makers' idiosyncratic normative understandings of free speech.

Moreover, the "vague and noncommittal" language Meta employs in its documents provides no substantive guarantees or rights to users as required by Kaye's framework.[89] The weak rights that are afforded to users are contained in documents can be amended or removed at Meta's will.[90] Meta's Community Standards are not wholly or substantially based on human rights norms or standards.[91] Rhetorical references to human rights in these documents reflect Meta's attempts to selectively and "narrowly tailor" its obligations to "minimize disruption" to Facebook's business model and practices "beyond the intended purpose"; [92] to outwardly project an image of a rights protective governance system and gain such "legitimacy dividends" associated with this inclusion of human rights language and vague connections to proportionality.[93]

B. Meta's Speech Governance as Probabilism

Although Meta adopts rhetoric echoing IHRL, its speech governance is not based on a consideration of human rights standards and user speech rights as the Board has adopted. It is instead based, at least partially, on protecting its brand and maximizing profit. This informs the main guiding precept of Meta's systemic balancing approach: probability.[94]

Instead of using explicit, publicly available rules, Meta governs Facebook users' speech through "implicit" rules built into and expressed through code and algorithms. [95] Such "architectural regulation" [96]

---

[87] Lwin, *supra* note 53, at 59, 60; KETTEMANN & SCHULZ, *supra* note 13.
[88] LAURA W. MURPHY, FACEBOOK'S CIVIL RIGHTS AUDIT — FINAL REPORT 9, 12 (2020), https://perma.cc/GX9F-UDBS.
[89] Klonick, *supra* note 2, at 2478. *But see* META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction ("Freedom of expression is a fundamental human right.").
[90] Klonick, *supra* note 2, at 2478.
[91] Aswad, *supra* note 53, at 40; Rikke Frank Jørgensen & Lumi Zuleta, *Private Governance of Freedom of Expression on Social Media Platforms: EU Content Regulation Through the Lens of Human Rights Standards*, 41 NORDICOM REV. 51, 57 (2020).
[92] LAIDLAW, *supra* note 6, at 242; Sander, *supra* note 14, at 966.
[93] douek, *supra* note 53, at 66.
[94] douek, *supra* note 17, at 766.
[95] Sander, *supra* note 14, at 946.
[96] GILLESPIE, *supra* note 36, at 179.

determines how speech is organized and presented to users. These technologies adopt "probabilistic methods," such as matching and predictive systems, [97] which necessarily involve error such as false positives or negatives, blind spots, and loopholes.[98] Indeed, error and probability is implicit in the idea of machine learning: such "learning" is "contingent upon continuous access to data . . . if machine learning is to deliver continuous improvement, or at least maintain performance, on assigned tasks." [99] "Algorithmic moderation," or "the use of automated techniques to classify content and apply a content moderation outcome to it," makes speech governance a "confluence of likelihoods" corresponding to the operation of a particular (set of) algorithmic filter(s) and categorizations. [100] This creates an "actuarial," probabilistic system of rights adjudication that accepts enforcement and decision-making errors as inevitable.[101]

Probabilism is also a structuring logic of Meta's speech governance because it enables the operation of Facebook's business model. It allows Meta to construct a governable userbase and collect the behavioral data necessary for its targeted advertising. Machine learning, through its "data appetite . . . devolves the process of assembling and representing a polity" to platforms.[102] The "observations, categories and databases" produced by this technology make "societies not just observable but *governable*"; thus, probability becomes a means to create and govern a society or community of online users.[103] With over 1.93 billion daily active users,[104] Facebook as a product would be unviable without the operation of automation and machine learning. Such technology is essential to the enforcement of the rules and standards which make up Facebook's product and enables it to operate at a behemoth scale. As Ananny notes: "scale makes money, and

---

[97] douek, *supra* note 17, at 798.

[98] For specific examples, see *id.* at 795-96; Daphne Keller & Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM 220 (Nathaniel Persily & Joshua A. Tucker eds., 2020). Regarding filtering errors, see also Emma J. Llansó, *No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem*, BIG DATA & SOC'Y 1-6 (2020); DAPHNE KELLER, STAN. CTR. FOR INTERNET & SOC'Y, DOLPHINS IN THE NET: INTERNET CONTENT FILTERS AND THE ADVOCATE GENERAL'S *GLAWISCHNIG-PIESCZEK V. FACEBOOK IRELAND* OPINION (2019).

[99] Marion Fourcade & Fleur Johns, *Loops, Ladders and Links: The Recursivity of Social and Machine Learning*, 49 THEORY & SOC'Y 803, 808 (2020).

[100] douek, *supra* note 17, at 793; Robert Gorwa et al., *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y, Feb. 28, 2020, at 1-15.

[101] Ananny, *supra* note 42; douek, *supra* note 17, at 797.

[102] Fourcade & Johns, *supra* note 99, at 812.

[103] Ananny, *supra* note 42. *See also* Fleur Johns, *Governance by Data*, 17 ANN. REV. L. & SOC. SCI. 53, 54-56, 61-62 (2021).

[104] Press Release, Meta, Meta Reports Fourth Quarter and Full Year 2021 Results (Feb. 2, 2022) (https://perma.cc/VCY7-ZXFT).

probability enables scale."[105] Meta's human moderation practices also operate on this logic of probability and scale. Frontline human moderators, also known as crowdworkers, have very little time to make decisions. Each piece of content receives "just a sliver of human attention."[106] Due to the volume of content, human reviewers are often unable to consider the content in its entirety.[107] Anecdotal evidence suggests moderators spend just 30 seconds reviewing a post,[108] and leaked internal Meta documents acknowledge that "our focus on scaled content-level enforcement in practice means the volume of decisions which need to be made is impossible for human reviewers to keep up with."[109] Context is actively removed from human content moderation processes, largely for expediency.[110] The outsourcing and "offshoring" of content moderation—while "answer[s] to the problem of scale"—distance moderators from the cultural and political contexts of the speech they review.[111]

The "distributed and decentralized" nature of decision-making, whether by humans or automation, means online speech becomes "desiccated" and detached from its original meaning; "[t]he data subsumes and comes to stand in for the users and their objections. Something is lost, and something is added."[112] Speech governance is thus determined by probabilistic logics where the workings and logics of scale are acute and where most of the enforcement and decision-making takes place: frontline human moderators and automated enforcement, collectively wading through millions of posts from Facebook's billions of users.

C. Probability and IHRL: Systemic vs Individualistic

There are therefore dual normative frameworks operating to guide and structure Meta's online speech governance, manifesting at different levels of the Meta governance process and bureaucracy. IHRL manifests at the level of the Board, the apex appellate body in Meta's speech governance system. Probabilism, an outcome of the continuous expansion of its user

---

[105] Ananny, *supra* note 42.
[106] GILLESPIE, *supra* note 36, at 121.
[107] OVERSIGHT BD., META, CASE DECISION 2021-003-FB-UA (Jan. 28, 2021) [hereinafter MODI & INDIAN SIKHS CASE], https://perma.cc/W9QA-9Y5G.
[108] Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019, 8:00 AM), https://perma.cc/AKJ4-JDWG.
[109] An anonymous whistleblower sent a letter to the SEC Office of the Whistleblower. WHISTLEBLOWER AID, FACEBOOK MISLED INVESTORS AND THE PUBLIC ABOUT ITS ROLE PERPETUATING MISINFORMATION AND VIOLENT EXTREMISM RELATING TO THE 2020 ELECTION AND JANUARY 6TH INSURRECTION, at 6, https://perma.cc/U6UY-RG8W.
[110] GILLESPIE, *supra* note 36, at 122.
[111] *Id.* at 135-137.
[112] *Id.* at 138.

base and user content, operates in the decision-making practices of lower-level bureaucratic actors: automated and human content moderators, where most speech decisions and enforcement occur. However, these frameworks operate on different normative bases and produce arguably conflicting outcomes.

The focus of probabilistic speech governance is systemic, not individualistic. Decisions about speech are made at "system design and tool development" levels and center not the individual speech or speaker, but the occurrence of error that is acceptable to users.[113] Balancing of interests occurs "*ex ante*, at the moment of system design and tool development," where error rates and occurrences are built into and accounted for within this system. [114] Probabilistic governance is thus "intrinsically *systemic* rather than *individualistic*."[115] In contrast, IHRL is a framework for individual rights adjudication, [116] and the Board's binding jurisdiction is only for specific complaints and thus relates to individual users only.[117] This is not to say that the Board does not or has not considered systemic issues, or that Kaye's IHRL-as-content-moderation framework does not address human rights issues in terms of system design and operation.[118] However, the scope of its jurisdiction for review is undoubtedly focused on the adjudication of individual cases, as is its IHRL methodology. While the Board's decisions and recommendations may be applied to analogous situations, this is not binding. Given the scale of Meta's content moderation exercise, the Board will only be able to review a miniscule fraction of content published.[119]

douek describes the individual-centric nature of IHRL-based speech adjudication as "impractical."[120] The objective in probabilistic governance is to "increase the probability that most decisions will be right most of the time and when the system errs, it does so in a preferred direction."[121] Error is "rationalized" when an algorithm or software is created or designed such that the occurrences of such errors remain within so-called "acceptable" limits.[122] This perspective values consistency over absolute protection of

---

[113] douek, *supra* note 17, at 789.
[114] *Id.* at 797.
[115] *Id.*
[116] Aswad, *supra* note 53, at 40, 41, 57.
[117] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 4.
[118] *See* David Kaye, *supra* note 8, at 12-14.
[119] Klonick, *supra* note 2, at 2490-91; douek, *supra* note 3, at 15. *See also* evelyn douek, *The Administrative State of Content Moderation*, YOUTUBE, at 3:45 (Oct. 26, 2021), https://perma.cc/6FC9-AFZ9, https://www.youtube.com/watch?v=RGdeV4KChWE (last visited Apr. 24, 2022).
[120] douek, *supra* note 17, at 790.
[121] *Id.* at 791.
[122] *Id.* at 768.

speech rights.[123] As "probable enforcement is reality,"[124] accepting the inevitability of error and rationalizing this at the system design level is, according to douek, the "*only* possibility between two extremes of severely limiting speech or letting all the posts flow."[125]

Yet a human rights approach understands that, as Balkin argues, "error costs are borne by the speaker, not the filtering system, and the burden is on the speaker to have the block altered or removed."[126] An "error" is not an abstract notion, it is a concrete, lived experience, usually endured by specific (marginalized) groups.[127] In the context of the Board's adoption of IHRL as its methodology, an error may represent not only a breach of internal platform law, but also an abrogation of an individual's rights under international human rights law. As an erroneous limitation on user speech, an enforcement error against a user is prima facie an illegitimate, disproportionate, and/or unnecessary restriction on speech under Article 19(3). Error can also compromise other human rights commitments under the UNGPs, for instance user appeal and due process rights.[128] Error can entail significant human rights impacts which disproportionately affect certain users, particularly with the use of AI and algorithmic enforcement.[129] In accepting and building error into platform design, probabilism therefore accepts a certain rate of individual rights abrogation as a matter not only of inevitability but as an established aspect of Meta's governance system. As error constitutes a potentially unjustified limitation of individual speech rights, probabilistic and IHRL frameworks create normatively conflicting outcomes: one system's acceptable error is another's unacceptable rights abrogation.

III. COLLIDING FORCES: THE RESULTS AND ROLE OF THE OVERSIGHT BOARD

The discord between probabilistic and IHRL frameworks reveals the existence of multiple normative speech governance structures operating at different levels of Meta's speech governance system and bureaucracy. However, this conflict also reflects the broader discord between Meta and the Board's respective approaches to speech governance, and the limits of the Board's ability to mitigate the risks of private speech governance. Meta's probabilistic framework contributes to some of the key problems

[123] Sander, *supra* note 14, at 948.
[124] douek, *supra* note 17, at 798.
[125] *Id.*
[126] Balkin, *Old School/New School Speech Regulation*, *supra* note 15, at 2318.
[127] *See* Ananny, *supra* note 42.
[128] This will be discussed below, especially Part III.B.2.
[129] This will also be discussed below at Part III.A.3 and III.B.3.

with private regulation: the opacity and lack of accountability in the design and operation of Meta's speech governance systems, influenced by private interests and incentives, and an absence of enforceable rights for users. The Board's intended operation, and its object in adopting an IHRL approach, was to mitigate these problems. However, policy and decision-making power is concentrated in the hands of Meta's executives, driven by a desire to fiercely protect and promote the profitability of Facebook as Meta's product. Decisions as to the operation and design of the probabilistic system—the explicit and implicit rules, standards, and technologies—are made almost exclusively at this executive level. This creates a dysfunctional dynamic that is reflected in the interaction between Meta and the Board. The Board has repeatedly asserted its role in improving the substance and enforcement of Meta's policies, largely through its IHRL framework. It strives to introduce minimum standards of transparency, accountability, and remedy as per the UNGPs through its advisory jurisdiction and public reason-giving. Yet, Meta consistently strives to circumscribe this role, and the potential benefits the Board could provide, in order to maintain the status quo of Meta's speech governance system. This status quo, however, bolsters inherently exclusionary, hierarchical speech dynamics that silence and exclude certain speech and speakers, while amplifying others.

A. Interplay Between Meta and the Board: Dynamics of Assertion and Restriction

The Board has, through its interpretative powers and its IHRL framework, expanded its advisory jurisdiction and created an obligatory feedback loop. This has led to some improvement in policy and some level of accountability for Meta's implementation of the Board's recommendations. However, the concentration of decision-making within Meta's executive leadership, informed by the corporate profit logic, restricts the Board's ability to provide the benefits of its advisory jurisdiction and weak-form review. Certain restrictions are explicitly constructed by Meta, like the circumscription of the Board's policy and algorithmic review jurisdiction. Others have emerged through interaction between the Board and Meta, within and after the Board's decisions, including Meta's (mis)translation of the Board's recommendations and refusal to respond to the Board's questions.

1. Expanded Policy Jurisdiction

The Board has narrowly articulated powers. It can instruct Meta to allow or remove content or reverse or uphold a designation in an individual

case.[130] With this binding authority, the Board can thus provide accountability for individual content enforcement decisions. However, policy issues raised, including those related to the system operation or design of Meta's platforms, necessarily fall under the non-binding advisory jurisdiction.[131]

The Board has, largely of its own accord, created a broad scope for its policy recommendations. Although its policy jurisdiction appears relatively narrow, and seemingly secondary to its other powers,[132] the Board interpreted it very broadly and has invoked it in nearly every case to make extensive policy recommendations, including relating to Facebook's system design and policies. This included recommending an "internal audit procedure" to analyze automated enforcement error rates,[133] publishing transparency reports with specific data, breakdowns, and metrics on content removals,[134] and conducting a "comprehensive review" of Facebook's contribution to the Capitol riots and "narratives of election fraud."[135] Meta itself has noted that "the size and scope of the board's recommendations go beyond the policy guidance we anticipated when we set up the board, and several require multi-month or multi-year investments."[136]

The Board has created this opening for itself largely through its human rights framework. Recommendations generally correspond to issues or gaps raised as part of the Board's IHRL framework and based on IHRL concepts. Although the IHRL framework is employed in the context of individual users and grievances, the Board has used this framework as the basis for its provision of much broader, system-based recommendations. The IHRL concepts most frequently cited are proportionality, legality, necessity, and remedial rights.

Proportionality

---

[130] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 4.
[131] *Id*. at art. 1, § 4, cl 4.
[132] It is not in the enumerated authorities but noted as a separate authority. *Id.* at art. 1 § 4 ("*In addition*, the board can provide policy guidance, specific to a case decision or upon Facebook's request, on Facebook's content policies. The board will have no authority or powers beyond those expressly defined by this charter." (emphasis added)).
[133] OVERSIGHT BD., META, CASE DECISION 2020-004-IG-UA (Jan. 28, 2021) [hereinafter BREAST CANCER SYMPTOMS AND NUDITY CASE], https://perma.cc/T4UV-SZXX.
[134] OVERSIGHT BD., META, CASE DECISION 2020-006-FB-FBR (Jan. 28, 2021) [hereinafter HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE], https://perma.cc/BK3M-4Z5U.
[135] TRUMP CASE, *supra* note 6363.
[136] FACEBOOK, FACEBOOK Q1 2021 QUARTERLY UPDATE ON THE OVERSIGHT BOARD (2021), https://perma.cc/RCT6-R5GU.

The Board invoked proportionality in the *Trump Case* at both individual and policy levels.[137] The Board found that Trump's indefinite suspension from Facebook was an "arbitrary penalty," and Meta should re-examine the penalty to ensure it is proportionate to the violation of its rules. In terms of policy, the Board criticized Meta for its application of a "vague, standardless penalty," noting that it was not the Board's role to adjudicate on such penalties, but Meta's role to "create necessary and proportionate penalties" corresponding to specific Meta rules, which can then be applied according to criteria such as the gravity of the violation and risk of repeat violations. Suspension of accounts should be proportionate to the risk posed by the violation.

A second example emphasizing proportionality is the *Navalny Protests Case*.[138] There, the Board decided that while the removal of a post from a Navalny supporter calling another user a "cowardly bot" accorded with the Bulling and Harassment Community Standard, it was a disproportionate restriction on free expression and contrary to Meta's values.[139] The Board noted that Meta's "blunt and decontextualized approach" to enforcement of its Community Standards can "disproportionately restrict freedom of expression." Given the context of freedom of expression in Russia, specifically government-led misinformation campaigns and online suppression of opposition supporters, the post should not have been removed for a relatively minor infringement of the Standards.

Legality

The Board has also invoked the principle of legality, ensuring rules are clear and accessible in multiple cases. In the *Modi & Indian Sikhs Case*, the Board criticized Meta for the lack of accessibility of its Community Standards.[140] Rules concerning the consequences of violating Community Standards are dispersed and not synthesized for users to understand account restrictions. Community Standards are not translated into Punjabi and therefore are not available to users or content moderators working in that language. The Board made corresponding recommendations. Legality has also been invoked in the enforcement context, where the Board has repeatedly recommended that Meta ensure users are notified of the specific reasons for enforcement action taken against them, including the violation of the specific Community Standard and the nature of the

[137] TRUMP CASE, *supra* note 6363; HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE, *supra* note 134134.
[138] OVERSIGHT BD., META, CASE DECISION 2021-004-FB-UA (May 26, 2021) [hereinafter NAVALNY PROTESTS CASE], https://perma.cc/P2WL-RKTS.
[139] *Id.*
[140] MODI & INDIAN SIKHS CASE, *supra* note 107107.

penalty.[141] In the *Trump Case*, the Board explicitly recommended that Meta provide greater transparency around their strikes and penalty system, and share sufficiently detailed notification of such enforcement action to users.

Necessity

In the *Hydroxychloroquine, Azithromycin, and Covid-19 Case*, the Board set out the requirement for the principle of necessity that Meta must show removal was the "least intrusive means to address the legitimate public interest objective" it pursued in removing such content.[142] In this case, removing a post that contained false information on COVID-19 treatments was not the least intrusive means of protecting public health, because Meta has other tools to address such content, like reducing the distribution of such content and providing additional information about a post. In its recommendations, the Board developed a list of criteria that would help the company determine what action is the least intrusive measure to adopt and encourage the adoption of alternative measures, including technological and algorithmic ones, such as "introducing additional friction to a post" by preventing sharing, or "down-ranking" such content to prevent visibility.

Appeal and Remedy Rights

The Board also invoked notions of procedural and appellate rights for users in both the *Breast Cancer Symptoms Case* and the *Modi & Indian Sikhs Case*.[143] Both cases concerned removal of content where the user had difficulties appealing the decision: in the former, there was removal by automation where later human review was not available due to COVID-19 resourcing issues; in the latter, there was content removal where an appeal was not available at all due to such COVID-19 issues. In both cases, the Board recommended that human review be a fundamental part of Meta's appeals process, that Meta address COVID-19 capacity constraints to ensure users have access to review, and that Meta provide transparent rules and processes for appealing decisions.

---

[141] OVERSIGHT BD., META, CASE DECISION 2020-005-FB-UA (Jan. 28, 2021) [hereinafter NAZI QUOTE CASE], https://perma.cc/8CA8-5JV2; NAGORNO-KARABAKH CASE, *supra* note 57; BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133; OVERSIGHT BD., META, CASE DECISION 2021-002-FB-UA (Apr. 13, 2021) [hereinafter DEPICTION OF ZWARTE PIET CASE], https://perma.cc/3GZF-TKR3. For Meta's response, see META, Q2 + Q3 2021 QUARTERLY UPDATE ON THE OVERSIGHT BOARD 16 (2021), https://perma.cc/YU9N-BTKX.

[142] HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE, *supra* note 134.

[143] *See* BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133; MODI & INDIAN SIKHS CASE, *supra* note 107.

Moreover, the Board has inscribed in its Bylaws that Meta must respond to all Board decisions publicly and, where policy guidance is issued, provide a document detailing any "follow-on actions" within sixty days.[144] Meta must provide "regular public updates" on its progress,[145] and the Board will continuously assess Meta's implementation of its responses.[146] Meta has subsequently committed to producing quarterly reports of its progress.[147]

These dynamics have created a feedback loop that has proved relatively successful in compelling Meta to implement certain recommendations and provide clarity on how it creates and implements its policies. Meta policy changes responding directly to Board recommendations include: changes to guidelines on permitted female nudity,[148] definitions of key terms used in the Dangerous Individuals and Organizations policy,[149] a new "newsworthiness" policy,[150] the introduction of a satire exception in the Hate Speech policy,[151] and a (forthcoming) translation of Community Guidelines into Punjabi.[152] Meta has also clarified how it uses automated detection and enforcement systems,[153] and its plans to improve those machine learning models.[154]

The Board has therefore expanded its policy advisory jurisdiction beyond what Meta originally imagined. It has crafted a space for itself within Meta's existing system of content moderation, designed to provide sustained and meaningful improvement of policymaking and enforcement through the creation of an obligatory feedback loop and by using their individual content jurisdiction to produce wide-ranging policy recommendations founded upon human rights standards. However, there

---

[144] META, OVERSIGHT BOARD BYLAWS, *supra* note 33, at art. 2, § 2.3.2.
[145] *Id.*
[146] *See* Oversight Board (@OversightBoard), TWITTER (Feb. 25, 2021, 10:32 AM), https://twitter.com/OversightBoard/status/1365006813570232320, https://perma.cc/6GCY-F2X6 ("In the coming months, we will be assessing how Facebook implements these responses, both to ensure the company takes action and to learn lessons for future decisions."). However, its mandate to enforce this remains unclear.
[147] The first report was released in July 2021. FACEBOOK, *supra* note 136.
[148] *Id.* at 11.
[149] *Id.* at 21.
[150] In response to the TRUMP CASE, Meta published this article: *Approach to Newsworthy Content*, META TRANSPARENCY CTR. (Jan. 19, 2022), https://perma.cc/TSN7-T5MV. *See* Mike Isaac, *Facebook Plans to End Hands-Off Approach to Politicians' Posts*, N.Y. TIMES (Sept. 24, 2021), https://perma.cc/W2VR-7KHE.
[151] Kim Lyons, *Facebook to Update its Community Standards to Clarify What it Considers Satire*, VERGE (June 19, 2021, 10:47 AM), https://perma.cc/ATZ5-ZHKG.
[152] FACEBOOK, *supra* note 136, at 25.
[153] *Id.* at 9, 14.
[154] *Id.* at 9-10.

remain some key barriers to the Board's ability to influence and improve policy and policymaking.

2. Weak-Form Review as Weak-Form Accountability

The Board should not be seen as a wholly altruistic creation, or a significant devolution of decision-making power from the company's executives. All "normative change processes" are escalated to senior executives,[155] who have "full visibility" on all policy changes.[156] The mechanics of this process are unknown.[157] douek has drawn the analogy between Meta's establishment of the Board and an authoritarian regime's divesting of a certain amount of authority to an independent judicial body as a means to legitimize their otherwise "expansive unilateral power" by providing a nominal check on this power.[158] Circumscribing the authority and jurisdiction of such bodies—as Meta has done by providing, for instance, limited scope for binding decision-making—is a mechanism to undermine the effectiveness of such an independent review mechanism.[159]

Endowing the Board with non-binding policy jurisdiction is similar to weak-form judicial review.[160] There are benefits to this model for speech governance. douek, drawing on Dixon, notes such benefits, including "counter[ing] blockages (such as blind spots and inertia) in the 'legislative process'."[161] Indeed, as noted above, the Board has proved relatively successful in improving certain policies. However, weak-form review provides only indirect accountability to users. As Dixon notes, the public deliberation and reasoning provided by weak-form review can provide the public attention and advocacy sufficient to "create pressure on legislators to respond to the demands of certain voters that they revisit an issue in line with evolving democratic majority understandings."[162] Accountability is therefore not a direct result of weak-form review; rather, it is accountability through public "reason-giving,"[163] in this case in both the Board's decisions and Meta's responses. Where users are informed about policy and design problems through the Board's decisions, they can respond accordingly and

---

[155] *Corporate Human Rights Policy*, *supra* note 13, § 5; Lwin, *supra* note 53, at 60.
[156] KETTEMANN & SCHULZ, *supra* note 13, at 28; Lwin, *supra* note 53, at 29.
[157] KETTEMANN & SCHULZ, *supra* note 13, at 28; Lwin, *supra* note 53, at 60.
[158] douek, *supra* note 3, at 17.
[159] *Id.* at 42.
[160] Klonick, *supra* note 2, at 2464; douek, *supra* note 3, at 54.
[161] douek, *supra* note 3, at 54. douek draws on Rosalind Dixon, *The Core Case for Weak-Form Judicial Review*, 38 CARDOZO L. REV. 2193 (2017).
[162] Dixon, *supra* note 161, at 2217.
[163] douek, *supra* note 3, at 57; *see also* Klonick, *supra* note 2, at 2464. This is also noted in the Charter. META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction.

apply pressure. However, Klonick's initial critique of Meta's accountability deficit remains. Meta may respond to threats of user exit, advertiser exit or public pressure, but these are only indirect accountability avenues for users.[164] The Board makes this process of identifying areas for policy change easier, bringing public awareness to key issues and the changes that should be made. Still, this remains an indirect, and therefore fragile and "amorphous", method of holding platforms to account.[165]

The logic underlying Dixon's argument as to the benefits of public reasoning in democratic contexts may not operate in the same way in the consumer context, where Meta is trying to prevent user and advertiser exit, rather than secure votes. Meta is a business, not a democracy "does not rely on popular will in setting its rules."[166] Users remain vulnerable to Meta's "arbitrary will."[167] As Kaye notes, platforms disclose "the least amount of information" and make the minimum changes required to prevent such user exit,[168] but they will not necessarily respond to the level or extent recommended by the Board or expected by users. Meta's responses reflect this practice, and anecdotal evidence suggests this "Team Selective Disclosure" approach is preferred amongst current Meta executives.[169] Under its Charter, the Board can request but not compel information from Meta, a lacuna in authority that has proven disempowering for the Board in its decision-making, as will be discussed below.

These design choices (weak-form review and limited jurisdictional scope) reflect this impulse to limit the divesting of power to the minimum required to secure legitimacy for Meta's content moderation regime whilst maximizing the policymaking and system design choices that remain within the corporate decision-making structures, which are not necessarily informed by human rights considerations, nor are they transparent. Moreover, Meta's corporate structures provide merely indirect and insufficient forms of accountability for issues of policy and system design.

3. Algorithmic Transparency

Meta has also restricted the Board's review scope by creating a significant gap in the Board's jurisdiction: review of Meta's algorithm design

---

[164] Klonick, *supra* note 12, at 1666.
[165] *Id.*
[166] douek, *supra* note 3, at 75.
[167] Klonick, *supra* note 2, at 2478.
[168] David Kaye, *supra* note 8.
[169] Kevin Roose, *Inside Facebook's Data Wars*, N.Y. TIMES (Oct. 4, 2021), https://perma.cc/SLM2-A5VR. *See generally* FRENKEL & KANG, *supra* note 23, at ch. 7.

and algorithmic ranking decisions.[170] Without specific jurisdiction to review algorithmic design and operation, these issues fall under the Board's non-binding policy jurisdiction. The Board has made recommendations related to algorithmic issues, including a push for increased transparency around the employment of algorithmic downranking[171] and around algorithmic design in relation to amplification of certain posts.[172] Meta has largely refused to disclose this information. Instead, it referred to other non-algorithmic tools (such as fact checking, notifications, and the COVID-19 info center[173]) and to separate transparency initiatives that may or may not have authority or information to consider algorithmic design.[174] Although binding policy jurisdiction is not necessarily a viable alternative, this combination of non-binding jurisdiction and a lack of jurisdiction to review the operation of Meta's algorithms means the Board has limited ability to compel Meta to disclose information about the operation of its automated enforcement and algorithmic moderation or improve these practices.

This compounds the current lack of transparency in the operation of Meta's algorithmic moderation and the lack of public reasoning as to the error rates underlying the operation of Meta's probabilistic speech governance system. The process of "rationalizing" error within Meta's content moderation has not yet been done openly and with public consultation and debate.[175] Public conversation may be shifting to more "nuanced" ideas about enforcement errors and ensuring "unavoidable error costs are not disproportionately assigned,"[176] but this is a conversation in its infancy. As Ananny puts it, most users are not privy to the most basic processes of content moderation, "much less the shifting statistical ground on which such judgments stand."[177] A public acceptance of platform error rates is also impeded by a lack of algorithmic transparency.[178] Although algorithmic transparency is not a simple task,[179]

---

[170] douek, *supra* note 3, at 40.
[171] HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE, *supra* note 134.
[172] *Id.*; TRUMP CASE, *supra* note 63.
[173] FACEBOOK, *supra* note 136, at 27.
[174] *See* TRUMP CASE, *supra* note 63. This will be discussed further in Part III.A.5 below.
[175] *See* douek, *supra* note 17, at 762, 808-10.
[176] *Id.* at 766.
[177] Ananny, *supra* note 42; *see also* Gorwa et al., *supra* note 100.
[178] *See* Keller & Leerssen, *supra* note 98; Paddy Leerssen, *The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems*, 11 EUR. J. L. & TECH. 1 (2020); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2016).
[179] Daphne Keller, *Some Humility About Transparency*, STAN. CTR. FOR INTERNET & SOC'Y BLOG (Mar. 19, 2021, 3:09 AM), https://perma.cc/N7ZZ-2LZP. However, there is a plethora of new research into algorithmic impact assessments and quantifying algorithmic harms.

the current discourse, which is "embarrassed" about probabilistic speech governance, renders decisions about acceptable error rates opaque, rather than such occurrences being "transparently acknowledged and defended."[180]

Moreover, racial and gender biases are an intrinsic and unacknowledged part of algorithms and therefore algorithmic moderation.[181] Public dialogue about error rates is not sufficient; we must also interrogate the "sociotechnical dynamics" that contribute to "uneven distributions of probability." [182] With neither Meta's cooperation nor specific jurisdiction to review algorithms that govern platforms and the design of the automated content moderation systems, the Board's ability to provide the public reasoning benefits of its weak-form review in this area are limited. Yet, this process of public reasoning is clearly required; douek's argument for accepting probabilism as the normative basis for speech governance relies upon this process of public rationalization of error rates, and therefore also upon Meta's disclosure of these error rates and the broader dynamics and operation of its algorithms and automated moderation. [183] Not only are these transparency benefits lost by the limitation of the Board's jurisdiction, but Meta is able to maintain the speech governance status quo: very limited and selective transparency about the operation of their algorithmic moderation systems and the impacts of this operation on individual speakers.

4. Translational Discord

These institutional limitations on the Board's scope of review are mirrored by similar dynamics emerging through the process of interaction and dialogue between the Board and Meta. First, Meta's responses to the Board's recommendations, as part of the obligatory feedback loop, are

---

See generally EMANUEL MOSS ET AL., ASSEMBLING ACCOUNTABILITY: ALGORITHMIC IMPACT ASSESSMENT FOR THE PUBLIC INTEREST (2021), https://perma.cc/27RC-CQUF; ROBYN CAPLAN ET AL., ALGORITHMIC ACCOUNTABILITY: A PRIMER (2018), https://perma.cc/7XTW-8XP5. *But see* Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J. L. & TECH. 117 (2021).

[180] douek, *supra* note 17, at 824.

[181] *See generally* Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, PROC. 57TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 1668 (2019); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Megan Garcia, *Racist in the Machine: The Disturbing Implications of Algorithmic Bias*, 33 WORLD POL'Y J. 111 (2016); U.N. Secretary-General, *Report on Artificial Intelligence Technologies and Implications for Freedom of Expression and the Information Environment*, U.N. Doc. A/73/348 (Aug. 29, 2018); DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 24-25 (2019); Sareeta Amrute & Emiliano Treré, *Episode 5: Data & Racial Capitalism*, DATA & SOC'Y (June 14, 2021), https://perma.cc/C5LH-D7TD.

[182] Ananny, *supra* note 42; *see also* Fourcade & Johns, *supra* note 99.

[183] douek, *supra* note 17, at 824.

often overly rhetorical and vague and, at worst, represent Meta's lack of good faith in considering and implementing the Board's recommendations. This reflects Meta's desire to maintain control over the level and types of transparency in which it engages.

Meta's responses are littered with vague, non-specific explanations and commitments. The Board recommended notification of the specific Community Standards where enforcement action is taken.[184] Meta responded that it works "to ensure a consistent level of detail is provided when content is removed,"[185] without specifics about such detail within the notification. Indeterminate language, such as "we will continue to monitor our enforcement and appeals systems,"[186] and "we will continue experimentation to understand how we can more clearly explain our systems to people"[187] means many commitments are neither specific nor measurable. The Board's transparency reporting noted that Meta fully "answered" 130 of 156 question.[188] However, the report did not meaningfully analyze the content of these responses and the extent to which Meta substantively answered the Board's questions.

Meta also has a propensity to mistranslate the Board's recommendations and overstate the extent to which its responses in fact implement these recommendations. Meta has attempted to restrict recommendations to a specific case, even where the Board clearly intended the recommendation to be a broader policy recommendation, for instance regarding notification of enforcement actions.[189] The *COVID-19 Misinformation Case* is an illustrative example of Meta's mistranslations. The Board recommended a transparency report on the enforcement of Meta's Community Standards related to COVID-19 misinformation, requesting specific data and breakdowns to be included.[190] Meta responded that it was already sharing COVID-19 data and would continue to do so, essentially ignoring the extensive list of specific data and analysis that the Board recommended be published.[191] Meta invoked the upcoming "Transparency Center," which it promised would "add more detail" about

---

[184] NAZI QUOTE CASE, *supra* note 141141141.
[185] *Id.*
[186] BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133.
[187] FACEBOOK, *supra* note 136, at 16.
[188] OVERSIGHT BOARD TRANSPARENCY REPORTS - Q4 2020, Q1 & Q2 2021, *supra* note 59, at 38-39, 61-62.
[189] NAZI QUOTE CASE, *supra* note 141. This issue was noted in several subsequent cases: NAGORNO-KARABAKH CASE, *supra* note 57; BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133; DEPICTION OF ZWARTE PIET CASE, *supra* note 141141141141141.
[190] HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE, *supra* note 134.
[191] *Oversight Board Selects Case on Hydroxychloroquine, Azithromycin, and COVID-19*, META (Feb. 25, 2021, 10:00 AM), https://perma.cc/LM9A-AMA7.

COVID-related enforcement (without specifying what this detail would be), as implementing the Board's recommendation.[192] Yet, as douek notes, this was essentially a "rejection," rather than an implementation, of the Board's recommendation.[193]

A similar issue again arose with Meta's response to the Board's recommendation in the *Breast Cancer Symptoms and Nudity Case*. In that case, an Instagram user's post showing pictures of breast cancer symptoms with corresponding descriptions was removed for violation of Meta's Community Standards on adult nudity and sexual activity by a "machine-learning classifier," and Meta said human moderation was not possible at the time due to COVID-19.[194] The Board recommended that Meta ensure that users can appeal automated decisions about removal of their content for violation of Meta's Community Standards to a human content reviewer. Meta responded that automation is an "important tool in re-reviewing content decisions" and it will "continue to monitor [its] enforcement and appeals systems to ensure that there's an appropriate level of manual review and will make adjustments where needed."[195] A further update noted that "the appeal will be reviewed by a content reviewer, except in cases where we have capacity constraints, such as those related to COVID-19," and the recommendation was marked as "fully implemented."[196] This is arguably antithetical to the Board's recommendation. Where there are "capacity constraints," which could arguably be the case for the whole content moderation system given the scale of the operation, human review will not be available. The unavailability of human review was the exact complaint the Board was preoccupied with; providing a justification as to why and when such review is not available is not "fully implementing" the recommendation.

Meta's response to the *Trump Case* was similar. The *Trump Case* concerned whether two posts by former US President Donald Trump on January 6, 2021 directed at Capitol riot protestors were against Community Standards prohibiting praise or support of people involved in violence. The Board's recommendation for clarification of Meta's newsworthiness policy was at least a partial success, as Meta later published its enforcement approach for this policy (factors for consideration include: country-specific circumstances; nature of the speech; country's political structure; whether

---

[192] *Id.*; FACEBOOK, *supra* note 136, at 24.
[193] douek, *Facebook Responds to the First Set of Decisions*, *supra* note 35.
[194] BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133. This was also noted in OVERSIGHT BD., META, CASE DECISION 2021-005-FB-UA (May 20, 2021) [hereinafter ARMENIAN GENOCIDE CASE], https://perma.cc/VG8V-6NJQ.
[195] *Case on Breast Cancer Symptoms and Nudity*, META TRANSPARENCY CTR.), https://perma.cc/AZG3-PS7B (last updated Jan. 22, 2022).
[196] FACEBOOK, *supra* note 136, at 14.

content presents risk of harm; nature of the speaker).[197] Still, one may query whether this provides sufficient detail to create a set of substantive "criteria" for enforcement as described by Meta.[198]

The Board also called for greater transparency around Meta's "cross-checking" policy, which provides more levels of review to prevent "false-positive enforcement errors" for "high profile accounts."[199] Meta initially gave no substantive response while also arguing that error rates associated with such processes could not possibly be tracked,[200] which douek not only argued was incorrect, but also amounted to a failure of due process.[201] It was only after public pressure, and pressure from the Board itself, that Meta submitted a Policy Advisory Opinion request to the Board to review its cross-checking policy.[202] Despite the Board's recommendation in the *Trump Case* that all users be subject to the same rules on Meta's platforms, rather than instituting separate policies and enforcement exceptions for political leaders, it remains Meta's policy not to fact-check posts from politicians.[203]

In reference to the Capitol riots, the Board also recommended a "comprehensive review" and "open reflection on the design and policy choices that Facebook has made that may enable its platform to be abused." [204] Meta denied responsibility [205] but noted that it "regularly review[s] . . . policies and processes in response to real world events" and had implemented the recommendation by "expand[ing] our research initiatives to understand the effect that Facebook and Instagram have on elections."[206] It is unclear what this "research initiative" will involve, what

---

[197] *Approach to Newsworthy Content*, META TRANSPARENCY CTR., https://perma.cc/7FA5-LMXA (last updated Jan. 19, 2022).

[198] META, Q2 + Q3 2021 QUARTERLY UPDATE ON THE OVERSIGHT BOARD , *supra* note 141141, at 12.

[199] *See Reviewing High-Impact Content Accurately via Our Cross-Check System*, META TRANSPARENCY CTR., https://perma.cc/552P-6YFK (last updated Jan. 19, 2022).

[200] *Case on Former President Trump's Suspension from Facebook*, META TRANSPARENCY CTR., https://perma.cc/JX5V-NWWB (last updated Jan. 19, 2022).

[201] douek, *Facebook's Responses in the Trump Case Are Better Than a Kick in the Teeth, but Not Much*, *supra* note 35; *see also* Jen Patja Howell, *The Empire (Facebook) Strikes Back (at the Oversight Board's Trump Decision)*, LAWFARE (June 10, 2021, 5:01 AM), https://perma.cc/FZ4Q-VXR4.

[202] Oversight Board (@OversightBoard), TWITTER (Sep. 28, 2021, 7:45 AM), https://twitter.com/OversightBoard/status/1442862972926189575, https://perma.cc/V7L4-FMHA.

[203] *Fact-Checking Policies on Facebook*, META BUS. HELP CTR., https://perma.cc/Z3WP-L3NQ.

[204] TRUMP CASE, *supra* note 63.

[205] *Case on Former President Trump's Suspension from Facebook*, *supra* note 200, at Recommendation 14 (implementing in part) ("The responsibility for January 6, 2021 lies with the insurrectionists and those who encouraged them, whose words and actions have no place on Facebook.").

[206] *Id.*

scope (or independence) the researchers will have in their inquiries, what data they will be able to access, and indeed whether design and algorithmic choices and policies will be considered, as envisioned by the Board.[207] A further update notes that Meta has conducted consultations with "more than 20 external stakeholders" without additional detail as to this consultation process and its outcomes.[208]

In both the *Trump* and *COVID-19* cases, Meta responded to the Board by pointing to existing initiatives over which Meta has total control, including the information it discloses. Largely without tailoring these initiatives to the substance of the Board's specific recommendations, Meta presented them as a direct implementation of the Board's recommendations. This dissonance between the Board's recommendations and Meta's responses undermines douek's argument that weak-form policy review is sufficient because of the "high reputational costs for Facebook in disregarding a decision of the FOB . . . [or] any substantial undermining of its authority."[209] Rather than outrightly disregarding recommendations or denying the Board's authority, Meta has mistranslated (and therefore only partially implemented), or sometimes completely ignored, the Board's recommendations. Meta attempts to receive the legitimacy benefits of purporting to implement the recommendations of this independent body while in fact making very little (or no) changes to policy and disclosing as little as possible.

5. Dodging Questions

Meta has proved very reticent to answer the Board's inquiries. The Board has the power to request information,[210] but Meta is under no obligation to adhere to this request and may decline where it determines that the information "is not reasonably required for decision-making, is not technically feasible to provide . . . or cannot or should not be provided because of legal, privacy, safety or data protection restrictions or concerns."[211] The Board cannot compel Meta to answer.

The Board requested information about the amplification of Trump's posts by the "platform's design decisions, including algorithms, policies, procedures and technical features," and any contribution of such design

---

[207] Facebook has not been particularly supportive of independent researchers recently. *See* James Vincent, *Facebook Bans Academics Who Researched Ad Transparency and Misinformation on Facebook*, VERGE (Aug. 4, 2021, 7:08 AM), https://perma.cc/D5X7-K48K; Laura Edelson & Damon McCoy, *We Research Misinformation on Facebook. It Just Disabled Our Accounts.*, N.Y. TIMES (Aug. 10, 2021), https://perma.cc/E22Y-EE6N.

[208] META, *supra* note 141, at 24.

[209] douek, *supra* note 3, at 56.

[210] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 4.1.

[211] META, OVERSIGHT BOARD BYLAWS, *supra* note 33, at art. 2, § 2.2.2.

decisions to the Capitol attacks.[212] Meta refused to answer, giving no justification. Meta thus refused to engage in the Board's attempt to provide a forum for public reason-giving and transparency regarding the broader systemic and design features that amplified Trump's inflammatory speech with significant "offline" political consequences. Meta referred to the forthcoming "research initiative" as a vague substitute for its refusal to answer. This again shows Meta's desire to disclose information on its own terms, rather than through transparency efforts by the Board.

Meta also refused, based on the bylaw exception, to respond to the Board's question as to whether Meta removed a Facebook post regarding the Indian government's treatment of those involved in national farmers' protests upon request by Indian authorities.[213] Meta determined that the requested information was irrelevant. Yet, it was potentially very relevant to the Board's decision-making, particularly from a human rights perspective. As the Board noted, the context of the farmers' protests was very politically sensitive, and the removal of the video not only undermined the freedom of expression of this specific minority group, but it also (inadvertently or otherwise) supported the Indian government's agenda to do so.[214] There were potentially legal reasons for Meta not to divulge this information, with the introduction of a new law requiring platforms to remove content where requested by the government.[215] However, responding to state pressure is where IHRL is supposed to have one of its greatest impacts for online speech governance, especially such drastic and blatant state pressure as this new law represents. IHRL "enables forceful normative responses against undue State restrictions" and provides the normative basis for companies to resist "government demands for excessive content removals."[216] The Board provides an avenue for Meta to publicly reason and justify its decision to resist (or accept where appropriate) such requests. By refusing to answer (legal justifications not withstanding), Meta refused to engage in this process of public reasoning and thereby undermined one very tangible benefit of the Board's IHRL

[212] TRUMP CASE, *supra* note 63.
[213] MODI & INDIAN SIKHS CASE, *supra* note 107.
[214] This was also noted in the NAVALNY PROTESTS CASE, *supra* note 138.
[215] Under a law passed in India in 2021, it is unlawful for platforms not to comply with Indian government requests for content removal and account closures, and to disclose when such removal has occurred. *See India's Internet Law Adds to Fears Over Online Speech, Privacy*, AL JAZEERA (July 15, 2021), https://perma.cc/TV6E-SBMU; Billy Perrigo, *India's New Internet Rules Are a Step Toward 'Digital Authoritarianism,' Activists Say. Here's What They Will Mean,* TIME (Mar. 11, 2021), https://perma.cc/YJQ8-84TV; Sheikh Saaliq, *India Introduces New Rules to Regulate Online Content*, ASSOCIATED PRESS NEWS (Feb. 25, 2021), https://perma.cc/PMR3-7ZWF.
[216] David Kaye, *supra* note 8, at 14.

framework. Meta also denied the Board the opportunity to hold Meta accountable for its interactions with governments and its potential complicity in government projects of censorship and restriction of minority groups' freedom of expression.

Where Meta refuses to cooperate, the benefits of the Board and its weak-form review are limited to public awareness. Knowing that a specific question was asked and that Meta refused to answer, the public can deduce what it can from this interaction, informed by the Board's own assessments or recommendations. Again, accountability is indirect, resting on public pressure to which Meta is not guaranteed to (sufficiently) respond. Meta maintains its control over policy changes and transparency, disclosing very little (or nothing). This creates a dynamic of very little oversight and accountability for such platform design and policy questions, which is particularly concerning in light of the potentially very damaging impacts for user speech rights.

B. Exclusionary Speech Practices: Institutional and Systemic Barriers to User Speech Rights

Adopting similar methods, Meta has also restricted the Board's ability to act as anything more than a symbolic appeal avenue for Facebook users. Restricting the Board's purview, this time to "important" cases with an element of political controversy, and limiting its size inhibits Facebook users' ability to access the Board to enforce their freedom of expression rights, otherwise unprotected by Meta, and to remedy "inevitable" enforcement errors as part of its probabilistic speech governance. This reflects a broader trend of exclusion and hierarchy embedded in the logic and operation of Meta's advertising business model and the machine learning underpinning the platform. These factors actively inhibit certain users from participating in online speech, while amplifying the speech of others, perpetuating traditional social, economic, and political hierarchies.

1. Board as Appeal and Grievance Mechanism

The UNGPs and Kaye's framework mandate the development of "scalable" and "operational-level" appeal and remedial mechanisms for users that "operate consistently with human rights standards."[217] The Board was designed to provide users with an effective appeal avenue,[218] reflected in the Charter. Echoing Pillar III of the UNGPs, the Board is to

[217] *Id.* at 18; U.N. Working Grp. on Bus. and Hum. Rts, *supra* note 43, at 31-35. *See generally* Pillar III.
[218] This featured heavily in the initial internal and public discussions. *See* Klonick, *supra* note 2, at 2450-51.

"provide an accessible opportunity for people to request its review and be heard"[219] and "increase access to remedy" for users and "relevant right holders."[220] One of the principal benefits of the Board is therefore the "participatory empowerment" it provides users through this grievance process.[221]

The Board's decisions show that it perceives itself in a similar way and is willing to interpret the Charter in light of, and consistent with, this role. The Board asserted its jurisdiction to hear appeals where Meta reversed its decision and restored content, *after* the Board's case selection but *prior to* its decision.[222] It justified this on the basis that "the Board's process offers users an opportunity to be heard and to receive a full explanation for why their content was wrongly removed."[223] Meta's ability to exclude cases from the Board's ambit simply by reversing Meta's decision would limit the benefits of transparency and public reason-giving provided by the Board, reducing the Board's ability to consider the systemic issues underlying an individual complaint. Not only is the Board itself an appeal mechanism, but, according to an IHRL framework, it raises deficiencies in user participation and appeal opportunities within Meta's moderation practices and makes corresponding recommendations. The Board has raised Meta's failure to notify users of enforcement action against them and their specific breach of publicly-available rules as failing to meet minimum procedural fairness requirements and limiting users' ability to appeal content removals.[224] The unavailability of human review undermines users' right to appeal and remedy.[225] Failure to consider context precipitates a "blunt and decontextualized approach [that] can disproportionately restrict freedom of expression."[226] The Board therefore perceives itself as a key appeal and remedial mechanism for users, not only in hearing appeals but in its ability to make recommendations to improve key appeal and due process deficiencies in Meta's internal governance processes.

2. Benefits Interrupted: Institutional Barriers to Accessing the

---

[219] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction.
[220] *Corporate Human Rights Policy*, *supra* note 13, at Commitment 3. Note that douek disagrees. *See* douek, *supra* note 3, at 6.
[221] Klonick, *supra* note 2, at 2489, 2499.
[222] Their jurisdiction to do so under the Charter was disputed by Facebook. *See* BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133.
[223] *Id.*
[224] ARMENIAN GENOCIDE CASE, *supra* note 194194.
[225] MODI & INDIAN SIKHS CASE, *supra* note 107.
[226] NAVALNY PROTESTS CASE, *supra* note 139; ARMENIAN GENOCIDE CASE, *supra* note 194194.

Board

Just as with the circumscription of the Board's policy review jurisdiction, Meta has limited the ability of the Board to act as a meaningful accountability mechanism by limiting both its purview, to only the more politically sensitive cases, and its size, such that it is dwarfed by the sheer scale of Meta's behemoth speech governance system.

The draft charter noted that the Board will "review Facebook's most challenging content decisions—focusing on important and disputed cases."[227] douek argues that this framing allows the Board to act as a "shield for [any] controversy that can attend divisive political decisions," allowing Meta to "outsource value judgments."[228] This drafting did not appear in the final Charter, however the ethos remains. The Board is to "oversee important matters of expression,"[229] selecting cases "that raise important issues pertaining to respect for freedom of expression," which are "of critical importance to public discourse."[230] All decisions to date have considered freedom of expression issues.[231] They also have the anticipated added degree of political interest or controversy.[232] The participatory benefits for the majority of Facebook users whose complaints or speech may not be politically sensitive, including those subject to enforcement errors, are minimal; their individual case is unlikely to be chosen and heard by the Board without meeting these two threshold factors.

Moreover, the size of the Board means that it will be able to hear very few cases and is thus unable to provide an "accessible" appeal body to

---

[227] META, DRAFT CHARTER: AN OVERSIGHT BOARD FOR CONTENT DECISIONS (2019), https://perma.cc/C9KT-PDNT; Nick Clegg, *Charting a Course for an Oversight Board for Content Decisions*, META (Jan. 28, 2019), https://perma.cc/QKF7-MFV4.

[228] douek, *supra* note 3, at 25. douek again employs the authoritarian regime analogy here.

[229] META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction. There are very few explicit limitations for those applying for an appeal. *See* META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 2, § 1; META, OVERSIGHT BOARD BYLAWS, *supra* note 33, at art. 2, § 1.2; *see also Appealing Content Decisions on Facebook or Instagram*, OVERSIGHT BD., https://perma.cc/CJC5-DNTK.

[230] OVERSIGHT BD., OVERARCHING CRITERIA FOR CASE SELECTION, https://perma.cc/J6JL-4R5F.

[231] *An Empirical Look at the Facebook Oversight Board*, LAWFARE, https://perma.cc/4WSN-9J5E.

[232] *See, e.g.*, TRUMP CASE, *supra* note 63; DEPICTION OF ZWARTE PIET CASE, *supra* note 141 (posting of images of a traditional blackface cartoon); MODI & INDIAN SIKHS CASE, *supra* note 107 (videos concerning Indian farmers protests); HYDROXYCHLOROQUINE, AZITHROMYCIN AND COVID-19 CASE, *supra* note 134 (misinformation about COVID-19 treatments); NAZI QUOTE CASE, *supra* note 141 (post quoting Goebbels); NAGORNO-KARABAKH CASE, *supra* note 57 (posting of hate speech against Armenians in the context of the Nagorno-Karabakh conflict); NAVALNY PROTESTS CASE, *supra* note 139 (regarding Russian Opposition protests); most recently, OVERSIGHT BD., META, CASE DECISION 2021-006-IG-UA (July 8, 2021) [hereinafter PKK FOUNDER ABDULLAH ÖCALAN CASE], https://perma.cc/QQ5M-SY74 (overturning original removal of post advocating for end of imprisonment of founder of the Kurdistan Workers' Party).

users.[233] With only twenty Board members and five members per decision panel,[234] the Board can only hear a handful of cases a year, despite receiving 524,000 user appeals between October 2020 and June 2021,[235] and twenty-six Meta referrals.[236] Given the enormous scale of Meta's content moderation system, the Board "cannot be expected to offer this kind of procedural recourse or error correction in anything but the smallest fraction of these cases."[237] This will remain the case even with the potential increase to forty members.[238] Unlike the scalable, accessible appeal body advocated for by Kaye and the UNGPs, and as presented by Meta, the Board provides at best a symbolic, exceptional appeal avenue for users. Without access to the Board, users are largely unable to enforce their freedom of expression rights because these rights are not guaranteed to users under Meta's governing documents and are not proceduralized into decision-making. Moreover, individuals with an "average" enforcement complaint without an additional political or high-profile aspect—such enforcement errors being inevitable and regularly occurring according to probabilistic content moderation systems—essentially have no independent mechanism of appeal, being unlikely to fulfil the criteria for the Board's selection and unlikely to be heard above the millions of other complaints, in any case. The Board's ability to mitigate the problems of this probabilistic speech governance by providing users with an avenue to remedy inevitable enforcement errors is therefore significantly restricted.

3. Business Models, Platform Design and Exclusionary Speech Governance

Meta not only circumscribes users' ability to exercise and enforce their freedom of expression rights by curtailing the role and scope of the Board, but this circumscription is in fact part of the very fabric of Facebook's operation and design. Balkin's understanding of the "digital age" as engendering a "pervasive social conflict brought about by technological change," embodied in conflicting social values and communicated through law and by extension the interpretation of free speech in theoretical and

---

[233] Facebook framed the Board this way in its Charter. META, OVERSIGHT BOARD CHARTER, *supra* note 27, at Introduction.
[234] META, OVERSIGHT BOARD BYLAWS, *supra* note 33, at art. 1, § 3.1.3.
[235] OVERSIGHT BOARD TRANSPARENCY REPORTS - Q4 2020, Q1 & Q2 2021, *supra* note 59, at 3. *See* Rebecca Heilweil, *You Can Finally Ask Facebook's Oversight Board to Remove Bad Posts. Here's How.*, VOX (Apr. 13, 2021, 12:20 PM), https://perma.cc/K4A8-MFMC.
[236] In Q1 2021 (ending March 31st, 2021), the Board only heard three of twenty-six Facebook referrals. FACEBOOK, *supra* note 136, at 4.
[237] douek, *supra* note 3, at 5-6.
[238] There is scope in the Charter for this increase to forty members. *See* META, OVERSIGHT BOARD CHARTER, *supra* note 27, at art. 1, § 1.

practical terms, has proved prescient in this age of private speech governance by online platforms.[239] Platforms as "machine-learning-powered corporations" have a new-found authority to create the terms of inclusion and exclusion, as they have become the important "mediators" between users and governing bodies.[240] The process and logics of platform engagement, fueled by and fueling the data hunger essential to machine learning, are structured by traditional "vectors of domination": economic, social and cultural capital, class, race and gender.[241]

Meta's speech governance ecosystem is inextricably tied to the business model of its platforms, which relies on increasing profit by maximizing user engagement: the more time users spend on Facebook, the more behavioral data the platform collects, which Meta monetizes through targeted advertising.[242] Continuous and increasing availability of user datasets enables the refining of Facebook's machine learning models, allowing Meta to promote increased user engagement through targeted, personalized feeds and advertising.[243] Indeed, machine learning, guided by its inevitable "data hunger," has been engineered to collect and analyze user behavior embedded within user data to further increase engagement.[244]

As Fourcade and Johns note, the once "inclusionary promise of machine learning has shifted toward more familiar sociological terrain": economic, social and cultural capital determine what is considered as "value-generating" speech online.[245] Those in privileged groups and with sufficient capital dominate and benefit from the operation of Facebook's machine learning. Trump, for instance, "weaponized" Facebook in the lead

---

[239] Balkin, *supra* note 24, at 14 ("We face, in other words, what Marx would have called a contradiction in social relations produced by technological innovation. By 'contradiction,' I don't mean a logical contradiction, but rather an important and pervasive social conflict brought about by technological change, a conflict that gets fought out in culture, in politics, and, perhaps equally importantly, in law. The social contradiction of the digital age is that the new information technologies simultaneously create new forms of freedom and cultural participation on the one hand, and, on the other hand, new opportunities for profits and property accumulation that can only be achieved through shutting down or circumscribing the exercise of that freedom and participation. The social conflict produced by technological change is both a conflict of interests and a conflict of values. It produces opposed ideas of what freedom of speech means. The social contradictions of the digital age lead to opposing views about the scope and purposes of the free speech principle.").

[240] Fourcade & Johns, *supra* note 99, at 813.

[241] *Id.* at 815-16.

[242] Sander, *supra* note 14, at 953; JACK M. BALKIN, FIXING SOCIAL MEDIA'S GRAND BARGAIN 3 (2018) ; FRENKEL & KANG, *supra* note 23, at ch. 3.

[243] Fourcade & Johns, *supra* note 99, at 822; Johns, *supra* note 103, at § 4.4. *See, e.g.*, BREAST CANCER SYMPTOMS AND NUDITY CASE, *supra* note 133. In that case, the post was detected and removed by "machine-learning classifier trained to identify nudity in photos."

[244] Fourcade & Johns, *supra* note 99, at 808.

[245] *Id.* at 816.

up to the 2016 elections by algorithmically amplifying his speech through the purchasing of Facebook ads, flooding the platform with posts and advertisements and microtargeting specific audiences, with the advantage of Meta's policy not to fact-check politicians' speech and political advertisements.[246] Whistleblower documents produced by former Meta employee Frances Haugen in 2021 show that Meta's assertions that it enforces its policies and standards equally and consistently, and that the Cross-Check system provides only an "additional review,"[247] were inconsistent with its practice of effectively exempting such "XChecked" pages and profiles from enforcement and integrity action, which complaint documents describe as "whitelisting" certain "privileged users," enabling blanket violations of Meta's terms of service.[248] Those with existing political and economic power benefit from, and can exploit the operation of, Meta's algorithms through these traditional markers of privilege and capital.[249]

Moreover, controversial or politically divisive speech is amplified by Meta's algorithms. Meta's data scientists found that "bad for the world" content, that which is inflammatory and politically controversial, including false or misleading information, keeps users on the platform for longer, thus securing Meta (and its advertisers) more data.[250] The result is amplification of politically divisive, controversial speech, often aligned with the far-right, conspiracy groups and sensationalist misinformation more broadly.[251] Whistleblower documents show that, prior to the January 6th Capitol riots, Meta recognized that "our core product mechanics . . . are a significant part of why" such "hate speech, divisive political speech and misinformation . . . flourish on the platform."[252]

---

[246] *Fact Checking Policies on Facebook*, *supra* note 203.

[247] WHISTLEBLOWER AID, FACEBOOK MISLED INVESTORS AND THE PUBLIC ABOUT EQUAL ENFORCEMENT OF ITS TERMS GIVEN THAT HIGH-PROFILE USERS ARE "WHITELISTED" UNDER ITS "XCHECK" PROGRAM 1-4 (2021).

[248] *Id.* at 6-8; *see* Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.*, WALL ST. J. (Sept. 13, 2021), https://perma.cc/V43B-WQYK.

[249] *See generally* Joseph Thai, *Facebook's Speech Code and Policies: How They Suppress Speech and Distort Democratic Deliberation*, 69 AM. U. L. REV. 1641-88 (2020); ÁNGEL DÍAZ & LAURA HECHT-FELELLA, DOUBLE STANDARDS IN SOCIAL MEDIA CONTENT MODERATION 13 (2021), https://perma.cc/Z4WK-ZG2G.

[250] FRENKEL & KANG, *supra* note 23, at 185, 286. *See also* DIPAYAN GHOSH, TERMS OF DISSERVICE: HOW SILICON VALLEY IS DESTRUCTIVE BY DESIGN (2020); David Lauer, *Facebook's Ethical Failures Are Not Accidental; They Are Part of the Business Model*, 1 AI & ETHICS 395 (2021); Julia Carrie Wong, *"Good for the world"? Facebook Emails Reveal What Really Drives the Site*, GUARDIAN (Dec. 5, 2018, 8:13 PM), https://perma.cc/5ZCY-7KNB.

[251] *See* Karen Hao, *How Facebook Got Addicted to Spreading Misinformation*, MIT TECH. REV. (Mar. 11, 2021), https://perma.cc/KBU2-2T5K.

[252] WHISTLEBLOWER AID, *supra* note 109, at 10; *see* Keach Hagey & Jeff Horwitz, *Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.*, WALL ST. J. (Sept. 15, 2021, 9:26 AM), https://perma.cc/5RHX-QFXL.

Meanwhile, typically marginalized or disadvantaged groups—women, racial and ethnic minorities, low socio-economic groups—are more frequently subjects of content removals, deplatforming, hate speech, and online harassment.[253] Internal Meta studies found that Meta identifies and removes only (and "optimistically") two to five percent of hate speech content, contrary to Meta and Zuckerberg's statements that Meta proactively finds and removes over ninety percent of hate speech content.[254] Meta's role in violence against Rohingya Muslims in Myanmar is a harrowing example.[255] This disparity in enforcement reproduces and creates new types and manifestations of inequality and oppression,[256] perpetuating social and economic hierarchies that have traditionally excluded and silenced the speech of certain groups and individuals.[257]

The dynamics of online speech are not simply an algorithmic or technological problem; humans are always in the broader design and operation of the machine learning "loop."[258] Machine learning promotes social interactions and behaviors and favors "certain companies, their shareholders and executives, while compounding conditions of social dependency and economic precarity for most other people."[259] Probabilistic speech governance, built on machine learning and inextricably tied to Facebook's business model, reinforces the perpetuation of these social and economic hierarchies and the exclusion/amplification of certain speech. Probabilistic speech governance therefore operates to the advantage of some and the disadvantage of others, rather than being a neutral system of impartial error and chance.

---

[253] DÍAZ & HECHT-FELELLA, *supra* note 249, at 10-12; Sap et al., *supra* note 181; Thomas Davidson et al., *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, *in* PROCEEDINGS OF THE THIRD WORKSHOP ON ABUSIVE LANGUAGE ONLINE 25-35 (2019). *See also* Karen Hao, *Facebook's Ad-Serving Algorithm Discriminates by Gender and Race*, MIT TECH. REV. (Apr. 5, 2019), https://perma.cc/4NFT-62PK; Bani Sapra, *How Facebook's Content Moderation Failed Palestinians*, WIRED (July 27, 2021), https://perma.cc/9EA7-HB7B.

[254] WHISTLEBLOWER AID, *supra* note 109, at 2, 6-7.

[255] Christina Fink, *Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar*, 71 J. INT'L AFF. 43-52 (2018); Neriah Yue, *The "Weaponization" of Facebook in Myanmar: A Case for Corporate Criminal Liability Notes*, 71 HASTINGS L.J. 813-844 (2019); Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. TIMES (Nov. 6, 2018), https://perma.cc/3HAN-84WX; David Kaye, *supra* note 8, at ¶ 41; FRENKEL & KANG, *supra* note 23, at 180-181, chs. 9, 10.

[256] *See* Fourcade & Johns, *supra* note 99, at 819.

[257] DÍAZ & HECHT-FELELLA, *supra* note 249. *See generally* Christian Fuchs, *Capitalism, Patriarchy, Slavery, and Racism in the Age of Digital Capitalism and Digital Labour*, 44 CRITICAL SOCIO. 677-702 (2018). For an interesting Australian perspective, see Ariadna Matamoros-Fernández, *Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube*, 20 INFO. COMMC'N & SOC'Y 930-46 (2017).

[258] Fourcade & Johns, *supra* note 99, at 806.

[259] *Id.* at 824.

IV. CONCLUSION

Meta has significant incentives to entrench its position and maintain the status quo, despite regulators in the US and EU circling.[260] Meta posted a doubling of profits and revenue in the first quarter of 2021.[261] However, recent revelations, particularly the "Facebook Files,"[262] have compelled a new level of scrutiny on Meta and its platforms. These documents essentially confirmed that Meta and its executives knew about the various online and real-world harms caused by the operation and design of Facebook and Meta's other platforms—the proliferation of hate speech, conspiracy theories, and far-right content. Moreover, Meta used this information to further increase user engagement, and therefore profit.[263] The change of the company's name from Facebook to Meta was in many ways a symbolic attempt to escape the backlash of these revelations and prevent user exit, instead compelling users to look towards their new "Metaverse" and continue to share moments of their lives through new Meta technology.[264]

Yet, it is clear that Meta will need more than a name change to ward off critics and regulators alike. These recent revelations have shown that the problems caused by and within Facebook's operation in fact arise from the business model and ethics of the platform and its governance. Private and economic interests are a significant, if not driving, force of the operation and design of speech governance frameworks. Platforms do not govern in ways that purely, or even in large part, further social good, despite

---

[260] Zolan Kanno-Youngs & Cecilia Kang, *They're Killing People: Biden Denounces Social Media for Virus Disinformation*, N.Y. TIMES (July 19, 2021), https://perma.cc/4YDH-9V24; Press Release, Fed. Trade Comm'n, FTC Sues Facebook for Illegal Monopolization (Dec. 9, 2020), https://perma.cc/299T-Q3JK; Alana Abramson & Vera Bergengruen, *Joe Biden's Fight With Facebook Is Just Beginning*, TIME (July 20, 2021, 7:00 AM), https://perma.cc/7AY4-ZC6J; Cat Zakrzewski, *European Union's Top Antitrust Enforcer Calls for Greater Global Alignment on Tech Regulation*, WASH. POST (July 12, 2021, 12:43 PM), https://perma.cc/NUF2-AV76; Cat Zakrzewski et. al., *Facebook Whistleblower Frances Haugen Tells Lawmakers That Meaningful Reform is Necessary 'For Our Common Good'*, WASH. POST (Oct. 5, 2021, 9:04 PM), https://perma.cc/YU8E-YXQE; Sam Schechner & Stu Woo, *Facebook Whistleblower Frances Haugen Calls for New Tech Laws in Europe*, WALL ST. J. (Oct. 25, 2021, 1:34 PM), https://perma.cc/7XQF-PL5L.

[261] Mike Isaac, *Facebook Nearly Doubles Its Profit and Revenue Rises 48 Percent, as Tech Booms*, N.Y. TIMES (May 5, 2021), https://perma.cc/GA7F-U9LA.

[262] *The Facebook Files*, WALL ST. J. (Oct. 1, 2021), https://perma.cc/ZH5C-CU4C.

[263] Ryan Mac & Cecilia Kang, *Whistle-Blower Says Facebook 'Chooses Profits Over Safety'*, N.Y. TIMES (Oct. 3, 2021), https://perma.cc/95SC-9CK7.

[264] Casey Newton, *Mark Zuckerberg is Betting Facebook's Future on the Metaverse*, VERGE (July 22, 2021), https://perma.cc/BG3A-BRPS; Tom Wheeler, *Seeing Past the Cool: Facebook's New Smart Glasses*, BROOKINGS INST. (Sept. 21, 2021), https://perma.cc/9SMC-QFY7.

platforms now being undeniably essential to, and influential in, democratic processes.

For many "optimists," [265] the Oversight Board was to provide a counterweight to these private, economic interests, providing independent oversight over Meta's decision-making and policies with the protection of individual user's speech rights as its core aim. It has, in many respects, exceeded the expectations of both Meta and many skeptics. It has adopted an IHRL-based adjudication framework, which allows the Board to judge Meta's actions and decisions according not only to the standards Facebook sets for itself, but also international human rights standards. This recognizes the normative weight of the UNGPs as a form of "soft law" as well as the need for Meta's speech governance to be tethered to universal standards, similar to those to which nation-states, with their like ability to censor and restrict speech, are also subject. It allows users to enforce their freedom of expression rights against Meta, as Meta itself provides no such rights to users in its internal platform law.

The Board's adoption of IHRL counterbalances, and in some respects conflicts with, Meta's probabilistic speech governance framework—born from Facebook's business model, which relies on maximizing user engagement and the exploitation of user data—and by the machine learning upon which the platform operates, which requires continuous and increasing sources of data. Whereas error in this system is not only inevitable but accepted (although not always publicly) at a system design level, error in an IHRL-sense can constitute an abrogation of individual speaker's right to expression. However, the Board's earliest decisions have proven that the Board can speak Meta's probabilistic language. It recognizes where automated enforcement has failed, where context has not been considered, and where scale and error has disproportionately and negatively affected certain minority or vulnerable groups. It conveys this to users and the public in a way we can understand, digest, and potentially act upon. The Board has also proven itself willing and able to impute to itself as broad a scope and authority as possible, through both its interpretation of the Charter and its IHRL framework. Through this, the Board attempts to temper the operation of private interests and incentives in Meta's speech governance, introducing a consideration of public interest into this broader system.

Yet, these earliest decisions have also shown that the Board's ability to improve Meta's speech governance in terms of policy and design, and substantively mitigate the concerns private regulation of speech raises, is limited in a number of ways. Meta has limited the scope of the Board's

[265] Klonick, *supra* note 2, at 2491-92.

operation and authority. This is a tactic aimed at maintaining concentration of power amongst the highest echelons of Meta's leadership, while reaping the legitimacy benefits of establishing an independent review body. This has largely allowed Meta to maintain the status quo of its current probabilistic speech governance system, characterized by opacity in its algorithmic design and operation, a lack of transparency and disclosure, and a lack of direct accountability to users, who must continue to rely on indirect, weak methods of accountability and rights enforcement. The operation of probabilism feeds into an exclusionary, hierarchical speech governance system. Along with Meta's advertising-based business model, Meta (re)enforces and (re)creates social, economic, and political hierarchies that threaten the speech rights of certain individuals and groups, particularly those who already struggle to be heard in politics and society, and for whom the democratizing, participatory promise of the internet held so much hope.

The Board has recognized "the power of [its] recommendations lies not in [Meta's] initial responses, but in the action the company takes."[266] The Board has therefore framed its role as "*steering* Facebook towards greater transparency," declaring that "[o]ver time, we believe that the combined impact of our recommendations will push Facebook to be more transparent and benefit users."[267] This will be both the key challenge and potential for the Board: to work with the tools it has been given by Meta in order to hold Meta accountable for, and promote greater transparency within, the operation and design of Meta's platforms. The Board operates on an individual level in the hope that this will lead to broader change that remedies similar types of errors and systemic problems. The Board has a growing public voice and presence that it can, and should, use to draw attention to these issues. The Board is a critical step in both compelling Meta to change its policies and operations, and in compelling Facebook users, civil society, and government to demand greater transparency and accountability from Meta.

---

[266] Oversight Board (@OversightBoard), TWITTER (Oct. 26, 2021, 4:30 AM), https://twitter.com/OversightBoard/status/1452960786155024386, https://perma.cc/6F9A-3YCT ("We know the power of our recommendations lies not in Facebook's initial responses, but in the action the company takes. That's why we recently set up a team to assess how our recommendations are implemented and to ensure FB is held accountable.").

[267] *Oversight Board Demands More Transparency from Facebook*, OVERSIGHT BD. (Oct. 2021), https://perma.cc/R54S-NAAR.