
Platform Oversight in Practice: How Language Shapes Procedure, Not Outcome, in Meta's Oversight Board

Endalkachew H. Chala

As private platforms increasingly govern public discourse, legitimacy hinges not only on moderation outcomes but on the institutional processes through which disputes are selected, reviewed, and adjudicated. This study examines the representational structure of platform governance through a case-level analysis of 147 decisions issued by Meta's Oversight Board between 2020 and 2025, coding geographic distribution, linguistic representation, policy domain, adjudicative outcomes, and procedural pathways. Oversight is globally distributed but internally stratified: representation spans diverse national contexts yet concentrates in a few countries, and non-English cases cluster disproportionately in the Global South. Policy attention favors high-stakes expressive domains. Adjudicative outcomes do not differ significantly by language or geopolitical classification; instead, language shapes procedural routing, as English-language cases are more often resolved through summary reversal and non-English cases proceed disproportionately to full panel review. Institutional design, more than identity characteristics, thus structures how disputes are handled. Because the Board's archive captures only the subset of disputes entering formal review, these patterns reflect institutional attention and procedural filtering rather than underlying dispute distributions. Legitimacy in corporate digital oversight therefore cannot be assessed through outcome parity or demographic representation alone; case selection and procedural routing are central.

1 Introduction

In many Global South contexts, Meta's platforms—Facebook, Instagram, WhatsApp, and Threads—are frequently described as synonymous with the internet (Nothias 2020), reflecting uneven broadband infrastructure, mobile-first connectivity, and zero-rated

access programs that embed these services in everyday communicative life (Singh et al. 2017). They function not merely as social networks but as foundational communication infrastructures mediating news access, political discourse, and civic participation (Nosthoff and Maschewski 2024).

This infrastructural centrality carries substantial consequences. United Nations investigators concluded that Facebook played a “determining role” in disseminating incitement against the Rohingya in Myanmar (UNHRC 2019), and comparable failures of contextual, local-language moderation were documented during Ethiopia’s Tigray War (Ayalew 2025), around WhatsApp-driven mobilization in the 2019 Indian general election (Kazemi et al. 2022), in Brazil’s messaging ecosystem (Evangelista and Bruno 2019), and across the Asia-Pacific more broadly (Sinpeng et al. 2021). The livestreamed Christchurch mosque attacks in New Zealand exposed parallel deficiencies in detection and response even in well-resourced, English-language contexts, as platforms struggled to stem the video’s viral re-circulation across North America and Europe (Klonick 2019). Such episodes reframed moderation failures as legitimacy deficits in private platform governance (Nahmias and Perel 2021).

Meta established the Oversight Board in response to this criticism. Announced in 2018 and operational since 2020, it is an independent quasi-judicial body empowered to review select moderation decisions and issue binding rulings (Douek 2019). Its creation shifted debate from technical moderation toward institutional legitimacy and the quasi-judicialization of transnational corporate governance (Douek 2024), as platforms exercising authority without democratic delegation sought to construct legitimacy through independence, transparency, and global representation (Suzor, Van Geelen, and Myers West 2018; Price and Price 2023).

Scholars remain divided on whether this innovation meaningfully constrains platform power. Some read the Board as an experiment in private constitutionalism; Douek (2024) cautions that procedural sophistication risks becoming a performance of legitimacy absent substantive constraint, while Helfer and Land (2023) reframe it as a de facto international human-rights tribunal. Early rulings nonetheless revealed an active adjudicator: the Board overturned a substantial share of Meta’s actions and intervened in politically charged disputes—from the conditional upholding of Donald Trump’s suspension (Oversight Board 2021) to scrutiny of enforcement exemptions for elite users (Oversight Board 2022)—prompting characterizations of it as a quasi-constitutional “Supreme Court” of platform governance (Klonick 2020).

The Board operates as a post hoc appellate institution, reviewing decisions only after they have passed through automated detection, human moderation, and internal appeals. Because frontline moderation is proprietary, governed by non-disclosure agreements, and largely inaccessible to researchers (Roberts 2019; Rogers 2025), these processes usually cannot be observed directly. The Board is a rare exception: its published decisions render high-level moderation publicly documented, formally reasoned, and systematic.

cally codable, offering a quasi-observational record of institutional process rather than one inferred through normative debate.

Using the Board as a research site offers three analytical advantages. First, it yields a systematically codable dataset in which decisions specify policy domain, geographic context, platform surface, procedural pathway, and adjudicative outcome. Second, as a capacity-limited appellate reviewer, the Board selects only a small share of appeals, making proportionality an empirical question—persistent over- or under-representation of topics or regions becomes a measurable institutional feature rather than a speculative concern. Third, location metadata make regional attention and outcome patterns empirically tractable, even though pathways to high-level review vary across regions by language, civil-society capacity, and media visibility.

Although developed around Meta, this approach extends in principle to other institutionalized oversight mechanisms. The current landscape, however, is uneven. Most comparable bodies are advisory rather than adjudicatory: TikTok convenes regional Safety Advisory Councils and a United States Content Advisory Council (TikTok 2024), Spotify established a Safety Advisory Council in 2022 (Spotify 2022), and Twitter maintained a Trust and Safety Council until its dissolution later that year (O’Brien and Ortutay 2022). Such councils counsel platforms on policy design but neither rule on individual cases nor publish reasoned decisions (Kettemann and Schulz 2023). Meta’s Oversight Board is a rare exception: it issues binding rulings on specific disputes and publishes formally reasoned decisions, generating a bounded governance record amenable to systematic comparative institutional analysis (Gorwa 2019; Klonick 2020). This relative institutional uniqueness is precisely what makes the Board an unusually observable site for studying how platform governance is structured. Accordingly, the study asks:

- **RQ1:** What is the geographic and linguistic distribution of cases reviewed at the highest level of platform governance?
- **RQ2:** How do geopolitical and linguistic contexts influence the procedural mechanisms through which moderation disputes escalate to institutional review (e.g., appeal routes, referral pathways, and case-type classifications)?
- **RQ3:** To what extent are adjudicative outcomes associated with geopolitical region or language, and does institutional review exhibit outcome neutrality across contexts?
- **RQ4:** How do distributions across policy domains reflect institutional prioritization within platform governance?

Collectively, these questions move beyond descriptive mapping to ask how institutional authority is distributed, exercised, and justified. Geographic representation, linguistic inclusion, procedural escalation, adjudicative outcome, and policy prioritization are not merely operational features; they are indicators of how legitimacy is constructed in sys-

tems that regulate global speech. The next section develops the conceptual framework that grounds this reading: institutional legitimacy as structured attention.

2 Conceptual Framework: Institutional Legitimacy as Structured Attention

Platform governance is best understood not simply as policy architecture but as institutional practice embedded in global information infrastructures, where legitimacy becomes a question of institutional performance rather than procedural design alone. To operationalize this view, the study advances an Institutional Attention and Legitimacy Framework that conceptualizes oversight bodies as systems allocating interpretive attention across disputes, communities, and policy domains. Platforms now exercise regulatory functions historically associated with public institutions—rulemaking, adjudication, and enforcement—yet without democratic delegation (Gillespie 2018; Suzor 2019), generating legitimacy deficits (Balkin 2018; Douek 2024). In response, they build institutional architectures that construct legitimacy through visible commitments to independence, transparency, and representativeness; Meta’s Oversight Board exemplifies this legitimacy-building approach through structured review, multi-stakeholder participation, and publicly reasoned decisions (Price and Price 2023; Suchman 1995).

Governance power operates not only through formal procedure but through institutional attention: the capacity to determine which disputes become visible, reviewable, and normatively consequential. This framing draws on two traditions. Scholarship on agenda control treats selection itself as power, structuring which issues gain salience and which are organized out of contention (Schattschneider 1960; Bachrach and Baratz 1962). The attention-based view of organizations, in turn, conceptualizes institutions as systems that allocate finite attention across competing demands (Ocasio 1997). This attention is further shaped by governance infrastructures and global inequality: content regulation depends on sociotechnical systems operating at scale (Gorwa, Binns, and Katzenbach 2020; Roberts 2019), while linguistic capacity, infrastructural access, and geopolitical prominence shape whose grievances become institutionally legible (Benjamin 2019; Noble 2018; Milan and Treré 2019). The Board’s docket thus functions as an institutional record of how governance authority is distributed across regions, languages, procedures, and policy domains.

Four empirical dimensions operationalize institutional attention, following the axes of distribution just described (region, language, procedure, and policy domain):

- **Geopolitical and linguistic representation:** the distribution of institutional visibility across regions and language contexts.
- **Procedural routing:** the pathways through which disputes enter and move through institutional review.

- **Adjudicative outcomes:** the presence of neutrality or asymmetry in how disputes are resolved.
- **Policy-domain concentration:** the allocation of institutional attention across substantive issue areas.

Together, these dimensions render institutional legitimacy empirically examinable: an observable pattern of structured attention rather than a purely normative aspiration.

3 Literature Review

3.1 Platform Governance and Legitimacy

This framework builds on several strands of scholarship spanning platform governance, content moderation, and institutional legitimacy. The institutional turn in content moderation has generated interdisciplinary work that treats moderation not as a discrete technical task but as a site of governance, where law, human rights, and communication intersect and where private corporations exercise quasi-sovereign authority over global speech. The Oversight Board sits squarely within this turn: far from peripheral to content moderation, it constitutes its highest formal tier of review, an appellate body that issues reasoned decisions on the moderation choices platforms make at scale. Gillespie (2018) characterizes platforms as “custodians” with discretionary power to define acceptable expression, while Gorwa (2022, 2024) situates moderation within transnational policy-making that reshapes relations among states, corporations, and civil society. Legal scholarship frames this authority as a legitimacy problem: platforms adjudicate and enforce without democratic delegation (Balkin 2018; Suzor 2019), and oversight mechanisms such as the Board attempt to approximate adjudicatory norms of independence, transparency, and reason-giving (Klonick 2020; Price and Price 2023).

3.2 Moderation Infrastructure, Political Economy, and Institutional Attention

Where the first strand foregrounds the macro-institutional question of legitimacy, a parallel literature turns to moderation’s operational substrate—the algorithmic classifiers and media-matching systems that embed policy judgments within proprietary architectures (Gorwa, Binns, and Katzenbach 2020), the often-exploitative human labor sustaining moderation at scale (Gray and Suri 2019; Roberts 2019), and the political-economic pressures that balance expressive commitments against commercial imperatives (Napoli 2019; Mulligan and Bamberger 2021). These accounts also note the limits oversight faces: restricted access to proprietary algorithms and internal enforcement data constrains independent evaluation of systemic bias (Pasquale 2015). For present purposes, the salient point is that these infrastructural and economic arrangements shape institutional attention—which disputes become visible, reviewable, and consequential—making the allocation of oversight attention across regions, languages, and policy domains an

empirical question.

3.3 Meta's Oversight Board: Scholarship and Empirical Gap

The most prominent instance of institutionalized oversight is Meta's Oversight Board, an independent body empowered to adjudicate appeals and issue binding rulings across Meta's platforms. Scholars conceptualize it as a novel form of transnational private adjudication that translates international human-rights principles into platform governance (Douek 2022; Klonick 2020; Suzor 2019), with legal analyses comparing its reasoning to international tribunals interpreting freedom-of-expression standards (Helfer and Land 2023). Price and Price (2023) read it as a legitimacy-building architecture that compensates for the democratic deficits of private governance. Wong and Floridi (2023) offer a systematic institutional assessment, identifying transparency gains and the Board's assertiveness alongside limited jurisdiction, Meta's discretion over whether rulings set precedent beyond the specific case, and insufficient representational diversity. Such concerns play out across a docket that ranges from regionally specific disputes to the Board's most visible interventions—from a case concerning Ethiopia to its review of a former head of state's suspension and of preferential enforcement for elite users (Oversight Board 2021, 2022).

Other work situates the Board within digital constitutionalism and private ordering (Douek 2021; Klonick 2020; Suzor 2019), while Douek (2024) argues more critically that institutional incentives favor procedural legitimacy over substantive accountability, such that the Board often avoids the most difficult cases it was designed to confront. Yet this scholarship is overwhelmingly doctrinal, conceptual, and case-based. Its central claims—about human-rights translation, legitimacy, digital constitutionalism, and institutional incentive—rest on legal analysis, close readings of individual decisions, or normative argument rather than on systematic measurement. What remains absent is a quantitative account of the Board's docket as a whole: how its cases distribute across geographic, linguistic, procedural, and substantive policy dimensions. This study supplies that account. By systematically quantifying a large, near-comprehensive corpus of the Board's published decisions, it provides empirical grounding for the conceptual and legal arguments reviewed above, allowing claims about representation, procedural routing, and adjudicative outcomes to be contextualized and situated against the actual composition of the record rather than inferred from exemplary cases. Because frontline moderation remains opaque, the Board's published decisions offer a rare, systematically documented window onto high-level adjudication. The next section outlines the research design, data, and analytical procedures.

4 Research Design

Quantitative research on platform governance faces a persistent measurement problem: moderation outputs are observable, but the institutional pipeline through which disputes are escalated, reviewed, and adjudicated rarely is, forcing reliance on indirect proxies such as enforcement statistics and platform transparency reports (Gillespie 2018; Rogers 2025). Access has narrowed further as platform APIs become restricted or commercially gated (Blakey 2024; Bruns 2019; Freelon 2018), and transparency reports remain aggregated and selectively curated (Suzor 2019). These constraints also raise a representativeness question: what enters the institutional oversight record, and whether that record distributes attention proportionally across topics, geographies, and communities.

The Board offers a tractable site for these questions. As a post hoc appellate body, its published decisions form a bounded, public corpus of the terminal layer of institutional review, each case carrying structured metadata and formal reasoning suited to analyzing proportionality and representativeness.

To operationalize this approach, the study adopts a case-based quantitative design that treats the Oversight Board as a structured sampling mechanism within platform governance. Rather than using content categories or enforcement outputs as proxies, the analysis draws on the Board’s official archive as the unit of analysis. This design measures platform governance at the level of formal institutional adjudication, aligning empirical observation with the Board’s delegated authority as a quasi-judicial oversight body in transnational content regulation. The approach aligns with computational communication research that reconstructs governance datasets through systematic extraction and analysis of publicly available digital institutional records (Marres and Weltevrede 2013).

4.1 Data Collection and Corpus Construction

The study constructs a corpus of Oversight Board case decisions and policy advisory opinions published in the **/decision/** directory of the Board’s official website. Rather than sampling, the objective was to assemble as comprehensive a set as the published archive allowed at the time of data collection. Because the Board publishes on a rolling basis, the resulting corpus reflects the decisions retrievable at that point rather than an exhaustive record of every action the Board has taken.

The resulting corpus constitutes a structured institutional dataset derived from the Board’s official decision archive, which serves as the formal public record of its adjudicatory and advisory activity. The dataset covers decisions published between **January 2020 and December 2025**, corresponding to the Board’s first five years of operation. Because decisions are released on a rolling basis and archival materials are periodically updated, the dataset represents a temporally bounded but evolving institutional record

rather than a closed historical repository.

Each entry provides standardized case-level documentation, including decision type (e.g., standard decision, summary decision, policy advisory opinion), platform surface, implicated policy domains and topical areas, and geographic context where available. Records also document procedural histories, including Meta’s original enforcement action, issues presented for review, and the scope of the Board’s mandate. Decisions include formal institutional reasoning, key findings, outcome determinations, and recommendations directed to Meta concerning policy revision, enforcement practice, and transparency standards. Many entries include supplementary materials such as policy updates, advisory requests, full advisory texts, and appendices of public comments.

To identify the decision pages available, the website’s robots.txt file was programmatically accessed to retrieve official sitemap URLs. These sitemaps were recursively parsed to extract all URLs matching the Board’s decision-page structure, yielding the set of publicly accessible decision pages indexed at the time of collection.

Because the Oversight Board does not provide a public API for structured metadata access, the dataset was constructed through iterative web scraping combined with manual inspection of page architecture (Bruns 2019; Marres and Weltevrede 2013). Initial extraction relied on structured JavaScript objects embedded within decision pages. Although this approach successfully retrieved full decision texts, it did not consistently expose structured metadata fields such as publication date, policy classification, geographic location, or adjudicative outcomes.

Subsequent DOM-based extraction captured labeled metadata fields (e.g., Platform, Topic, Location, Date), but introduced noise from navigation and footer elements. Manual inspection revealed two dominant page templates:

- (1) a multi-case bundle template containing repeated labeled case blocks, and
- (2) an earlier single-case template with differently structured metadata sections.

A hybrid parsing strategy was therefore implemented: the scraper first attempts bundle-based extraction and, if unsuccessful, falls back to a single-case routine.

Data cleaning procedures included removal of template artifacts, metadata normalization, categorical harmonization, and date standardization to ISO format. Outcomes and decision types were standardized across templates, and case identifiers were extracted where available.

Figure 1 summarizes the corpus-construction process. Sitemap parsing initially identified 228 candidate decision records. Ten were removed—four whose pages could not be resolved and six that recorded no adjudicative outcome—leaving 218 entries. A further 71 records corresponded to duplicate case blocks appearing within multi-case bundle pages; consolidating these yielded the final analytic corpus of **147 distinct decision-level cases** published between 2020 and 2025.

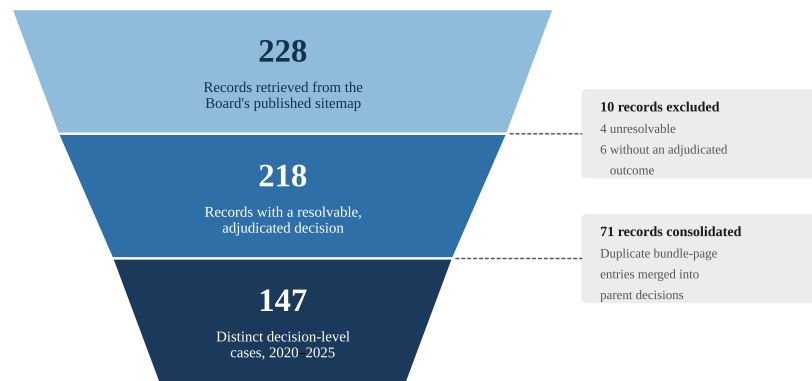


Figure 1: Construction of the analytic corpus, from 228 retrieved records to 147 distinct decision-level cases (2020–2025).

All scraping procedures followed ethical digital research practices, including rate limiting, minimal server load, and respect for access controls. Because Board decisions are official public records that the institution publishes for transparency and accountability, and because collection began from the site's robots.txt and published sitemap rather than circumventing any access control, web scraping constitutes an appropriate and permissible method for assembling the corpus.

4.2 Variable Construction

Each decision-level observation includes structured metadata and derived variables for institutional, geopolitical, procedural, and substantive dimensions of oversight.

Primary country labels were extracted from decision metadata. Countries were classified using a Global North/Global South framework informed by World Bank income groupings and geopolitical research conventions (Dados and Connell 2012).

Language context was operationalized as a binary indicator distinguishing English-only cases from those involving non-English content. Language identification relied on structured metadata and textual cues within decision summaries. Cases whose language could not be determined from these sources were coded as English for the binary indicator—a conservative choice that minimizes the estimated share of non-English content, whose sensitivity to alternative codings is examined in the results.

Procedural pathway was coded as either summary decision—where Meta reversed its decision following Board engagement—or full Board review involving panel deliberation and formal adjudicatory reasoning.

Policy domains were coded using the Community Standards categories specified on decision pages, while topical areas were recorded using issue tags provided in the metadata. Adjudicative outcomes were standardized into three categories: overturn, uphold, and mixed. Content modality (e.g., video, image, meme, audio, text) was coded through rule-based keyword detection applied to decision summaries and manually verified. Evidence excerpts were retained to ensure auditability and reproducibility. Additional indicators captured the presence of institutional recommendations and supplementary advisory materials.

4.3 Quantitative Analysis Strategy

Analyses employed descriptive statistics and inferential tests aligned with proportionality and representativeness research. Geographic representation was assessed through cross-tabulations and proportional comparisons across Global North, Global South, and mixed-region classifications.

Associations between language context and geopolitical classification were tested using chi-square tests of independence, with effect sizes estimated using Cramér's V (Agresti 2013). Outcome equity was evaluated using chi-square tests examining whether adjudicative outcomes varied across language and geopolitical categories. Overturn rates were compared across groups.

Procedural pathway differences were examined using chi-square tests assessing associations between language, region, and review type. The overturn analysis was restricted to full Board reviews; summary decisions were excluded because, by definition, they record cases in which Meta reversed its action, which would render overturn tautological.

Because the corpus comprises the decision-level cases published to the Board's public archive for 2020–2025, assembled systematically rather than sampled, the 147 cases constitute a large, near-comprehensive collection rather than a probability sample. The Board's cumulative output is larger—its December 2025 five-year review documents more than 200 published decisions since inception (Oversight Board 2025)—and the present corpus reflects the distinct decision-level cases retrievable from the Board's published sitemap at the time of collection, after consolidating multi-case bundle pages. Inference therefore concerns the precision of within-corpus comparisons rather than generalization to an external population. For categorical comparisons, expected cell frequencies were examined, and Fisher's exact test was substituted for the chi-square approximation wherever an expected count fell below five. The subset available for modeling overturn among fully reviewed cases is modest ($n = 72$), which constrains statistical power; associations in that model are accordingly interpreted as suggestive rather than confirmatory.

Policy domain concentration was evaluated using proportional distributions and structural concentration metrics. Market-style concentration was measured using the

Herfindahl–Hirschman Index (HHI), and inequality across policy categories was assessed using the Gini coefficient.

4.4 Limitations

The dataset includes only published Oversight Board decisions and does not capture the full volume of user appeals submitted for review. Because appeal volume is itself geographically skewed—heavily concentrated in North America and Europe and comparatively sparse from Sub-Saharan Africa, the Asia-Pacific, and Oceania (Oversight Board 2025)—the docket’s geographic distribution reflects both appellant demand and the Board’s case selection, and should not be read as a pure measure of either. Topic and location metadata reflect institutional labeling and may not fully capture contextual nuance. A small number of pages contained template irregularities that limited metadata extraction. Despite these constraints, the dataset offers a systematically assembled, near-comprehensive corpus of decision-level Oversight Board adjudication suitable for the representational analyses that follow.

4.5 Positionality

I approach this study as an Ethiopian-born, multilingual scholar whose research and professional experience span content moderation, trust and safety, platform governance, and AI-mediated public communication, including work as a moderation consultant and on a National Science Foundation–funded project on AI-assisted online deliberation. This applied and academic background attunes me to both the institutional goals of platform governance and its practical, linguistic, and contextual challenges. Observing how Meta’s platforms operate in contexts such as Ethiopia—as vital communication infrastructures, yet with limited linguistic resources, local expertise, and accountability—sensitizes me to how formal geographic inclusion can coexist with deeper inequalities in linguistic accessibility, procedural visibility, and institutional power.

I conducted this study independently and externally to Meta and the Board, with no access to internal deliberations, unpublished appeals, or proprietary systems; the analysis rests entirely on the public record, read through a critical perspective that treats case selection and procedural routing as exercises of power. This standpoint may make me especially attentive to Global North–Global South and low-resource-language inequalities, but I did not assume they would produce unequal outcomes. To limit prior expectations, I used explicit, evidence-retained coding, tested alternative language codings, and reported null alongside significant findings—one of which, the absence of outcome differences by language or region, redirected my attention from final rulings to procedural routing. I therefore treat positionality as a standpoint to be disclosed and examined, not eliminated.

4.6 Methods Transparency Summary

This study reconstructs a corpus from the official decision archive of the Oversight Board, enabling direct observation of institutional adjudication rather than relying on indirect moderation proxies. Systematic sitemap discovery, hybrid scraping protocols, template validation, and metadata normalization supported broad coverage and replicability. Structured variables capture institutional, geopolitical, linguistic, procedural, and outcome dimensions of oversight. By measuring governance at the level of formal institutional decisions, the study strengthens representational validity, procedural transparency, and institutional accountability in platform governance research.

4.7 Data Availability and Replication

The dataset underlying this study will be made publicly available upon publication. A curated replication package—including the cleaned decision-level dataset, a variable codebook, data-construction documentation, and analysis scripts—will be deposited in an open research repository (such as the Open Science Framework) to ensure long-term preservation and citation stability. In the interim, these materials are available from the author on reasonable request. All materials are provided to support transparency, reproducibility, and cumulative research on platform governance and institutional oversight.

5 Representational Structure of Oversight Board Adjudication

5.1 Geographic Scope and Country Concentration

Geographic representation reveals a globally distributed yet internally stratified oversight docket. As shown in Figure 2, cases span a wide range of national contexts, but institutional attention is concentrated in a limited set of countries. At the country level the United States remains the single most frequent origin, but grouping cases by world region shows a relatively even global spread—North America (23), Europe (22), Latin America and the Caribbean (20), the Middle East and North Africa (18), and Sub-Saharan Africa (17)—with a further 21 cases spanning multiple regions. This relatively even regional spread is itself notable, since the appeals from which the Board selects are far less evenly distributed: several lower-volume regions, such as Sub-Saharan Africa, nonetheless appear prominently in the docket, whereas areas from which few appeals arise—such as Oceania and East Asia—remain thinly represented (Oversight Board 2025). The distribution therefore reflects the Board’s active case selection rather than a direct mirror of appeal volume. By contrast, many countries appear only once or twice, indicating broad jurisdictional reach but comparatively shallow representation across much of the platform’s global user base. Across the corpus, 65 countries are referenced 198 times in total, but coverage is thin: the median country appears only twice, and 41 of the 65 (63%)

are referenced just once or twice, while the five most-referenced jurisdictions—led by the United States with 31 mentions—account for roughly a third of all references. The pattern is thus one of wide but shallow geographic reach rather than sustained engagement with most national contexts.

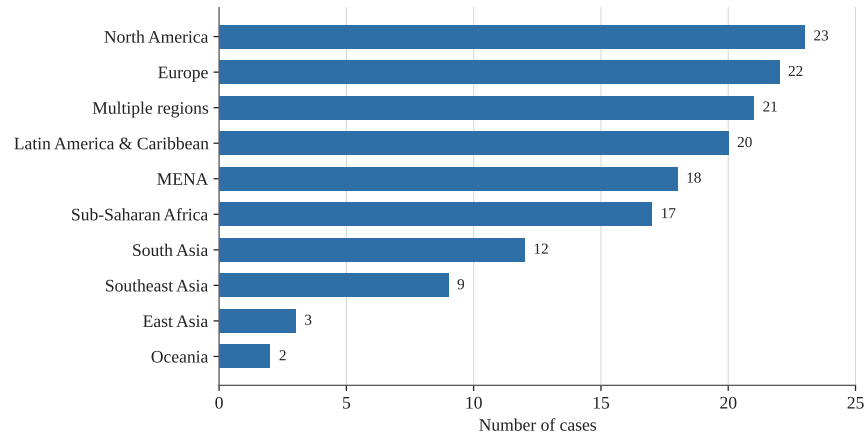


Figure 2: Distribution of Oversight Board cases by world region (N = 147).

When aggregated into geopolitical groupings, Global South contexts account for the largest share of the docket (n = 80), ahead of Global North contexts (n = 45), with a further 22 cases spanning both. That the Global South predominates—even though appeals originate disproportionately from the Global North—underscores the selective character of the Board’s docket. All three groupings are substantively represented, yet institutional attention remains unevenly distributed within each. The Board’s jurisdiction is therefore global in scope but hierarchically structured in its allocation of institutional visibility.

These patterns are consistent with institutional critiques emphasizing representational asymmetries in the Board’s composition and governance structure. Wong and Floridi (2023) note that the concentration of membership and leadership roles in North America and Europe may constrain sensitivity to region-specific moderation challenges—including in the very Global South contexts that dominate the docket yet remain thinly represented on the Board itself. While docket composition cannot be reduced to institutional demographics alone, the observed distribution aligns with concerns that governance architecture shapes which disputes become institutionally visible and prioritized for review.

5.2 Linguistic Diversity and Geopolitical Structure

Linguistic distribution within the Oversight Board docket further illustrates the relationship between diversity and concentration in formal oversight. As shown in Figure 3, most decisions do not state the content language (75 of 147 cases), a limitation I treat transparently rather than resolve by assumption. Among cases with an identified language,

English is the most frequent (18), followed by Arabic (11), Spanish (9), and a long tail of languages appearing only once or twice, with a further 17 cases multilingual. Because English accounts for only a small share of the full docket, and because undetermined cases originate disproportionately from non-English-speaking jurisdictions, a location-informed estimate places non-English or multilingual content at roughly 71% of the corpus (sensitivity range 37–88%, depending on how undetermined cases are coded). Formal oversight is thus linguistically diverse, even as English retains a privileged position among the cases the Board explicitly language-labels.

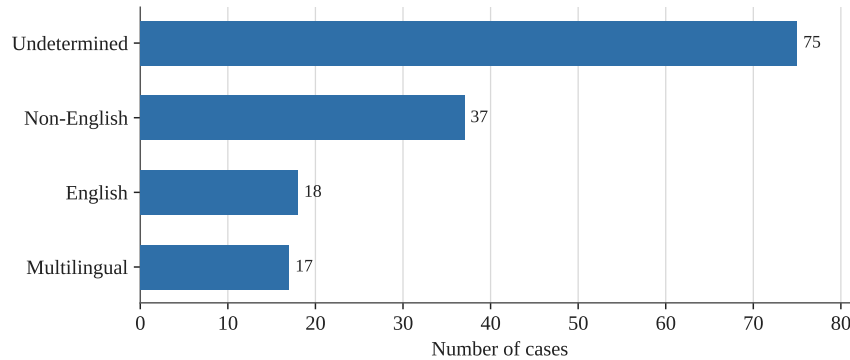


Figure 3: Content language of Oversight Board cases (N = 147).

The presence of multiple languages indicates that the Board’s archive extends beyond Anglophone disputes and reflects genuine linguistic diversity within its adjudicatory record. The distribution nonetheless remains uneven: English is the most frequently identified single language, and because most cases carry no language label, the docket’s visible profile skews toward English even as non-English and multilingual content likely constitute its majority. Institutional visibility within formal oversight thus appears disproportionately anchored to English-language content.

This pattern reflects broader structural inequalities in global content moderation systems. Research shows that moderation infrastructures operate unevenly across linguistic environments, with automated detection systems and enforcement tools heavily reliant on training data concentrated in English and other high-resource languages (De Gregorio and Stremlau 2023; Nicholas and Bhatia 2023). As a result, moderation systems often struggle to detect harmful content or contextual nuance in many languages spoken predominantly in the Global South (Joshi et al. 2020).

Empirical research on moderation in low-resource languages further documents these disparities. Comparative studies of Maghrebi Arabic dialects, Kiswahili, Tamil, and Quechua show that languages with limited computational resources experience persistent gaps in moderation capacity, enforcement consistency, and appeals processes (Elsawah, Bhatia, and Thakur 2025). These challenges are compounded by the outsourcing of moderation labor and the frequent assignment of moderators to linguistic contexts with which they have limited dialectal or cultural familiarity.

Linguistic inequalities also intersect with broader patterns of participation in platform oversight. Wong and Floridi (2023) report that more than two-thirds of appeals submitted to the Oversight Board in 2021 originated from the Global North, while several Global South regions accounted for only a small fraction of submissions. Limited awareness of the appeals process, infrastructural barriers, and linguistic accessibility may therefore restrict the ability of many users to bring disputes into the institutional review pipeline.

The present dataset extends the temporal scope of this observation. Although the number of published decisions has increased since the period examined by Wong and Floridi, the overall structure of linguistic representation remains broadly similar. The docket includes cases from multiple language environments, yet institutional attention continues to concentrate among a few high-visibility languages, even as the docket as a whole spans many. In this sense, the expanded dataset suggests persistence rather than resolution of earlier representational asymmetries in the Board's adjudicatory record.

These descriptive patterns provide a basis for examining whether linguistic context is associated with other dimensions of institutional review, including procedural pathways and adjudicative outcomes.

5.3 Linguistic Context and Geopolitical Distribution

The relationship between language and geopolitical context further clarifies the distribution observed above. Throughout these analyses, language context is the conservative binary indicator described earlier, in which the 75 language-undetermined cases are coded as English; because this folds predominantly non-English cases into the English category, any language differences it reveals are, if anything, understated. A chi-square test indicates a statistically significant association between language context and geopolitical classification, $\chi^2(2) = 15.74$, $p < .001$, Cramér's $V = .33$. The result suggests that linguistic context and geopolitical classification are related within the Board's docket, with the effect size indicating a moderate association.

As illustrated in Figure 4, English-language cases dominate Global North contexts (39 English vs. 6 non-English), whereas Global South cases exhibit a more balanced distribution (41 English vs. 39 non-English); mixed-region cases fall in between (13 English vs. 9 non-English).

This pattern indicates that English-language disputes appear disproportionately in oversight cases originating from Global North contexts, while Global South cases involve a more linguistically diverse set of disputes. Importantly, this concentration reflects not only a North–South divide but also the prominence of English within Global North cases specifically. Among the Global North cases whose language is identified, English appears far more often than the other European languages present, such as French and Italian. The linguistic concentration observed in the docket therefore reflects the institutional

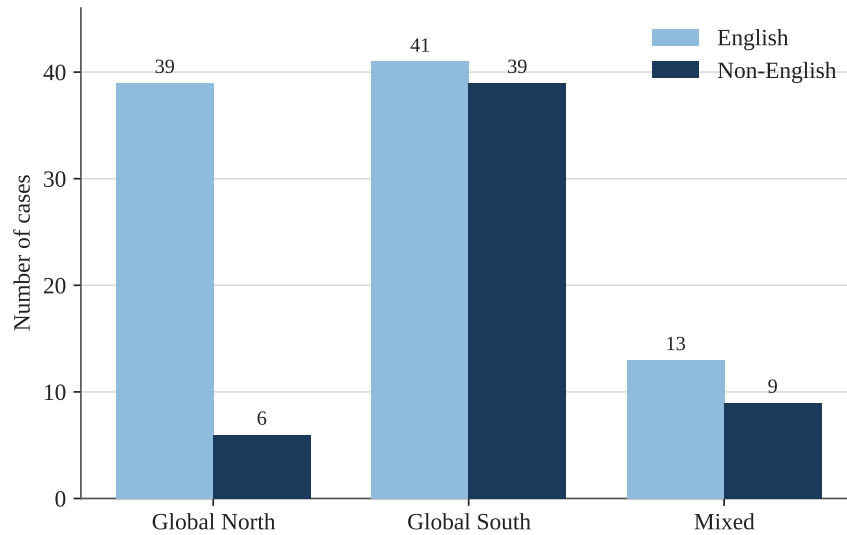


Figure 4: Language context by geopolitical classification.

prominence of English as the primary language through which platform disputes become formally visible within oversight processes.

This pattern aligns with broader research on multilingual moderation systems. Studies show that moderation outcomes depend not only on identifying harmful language but also on interpreting community rules, cultural context, and linguistic nuance across diverse online environments. Ye et al. (2023), for example, demonstrate that moderation decisions in multilingual settings often hinge on differences in community rule systems rather than the presence of offensive language alone. Moderation infrastructures must therefore interpret linguistic meaning, contextual cues, and normative expectations that vary across languages and communities.

Because moderation capacity is uneven across linguistic environments, the linguistic composition of oversight cases may reflect broader structural dynamics within platform governance. If reporting mechanisms, enforcement pipelines, and appeals pathways operate unevenly across languages, the disputes that ultimately reach formal oversight will likely carry these asymmetries forward. Language thus operates not only as a descriptive attribute of cases but also as a structural dimension shaping how institutional attention is distributed within platform governance.

5.4 Topical Concentration in Case Selection

The distribution of Oversight Board decisions across Community Standards reveals uneven institutional attention across issue domains. Concentration metrics indicate substantial concentration across Community Standards ($\text{HHI} \approx .18$) alongside pronounced inequality ($\text{Gini} \approx .60$). This pattern suggests that although no single category accounts for a majority of the docket, institutional attention is heavily concentrated in a small

number of high-stakes moderation domains: the three most frequent standards alone account for roughly 70% of labeled cases.

Relative to an equal-distribution baseline, high-stakes expressive domains—particularly Hate Speech, Dangerous Individuals and Organizations, and Violence and Incitement—appear disproportionately in the Board’s decisions. By contrast, Community Standards such as Regulated Goods, Fraud and Deception, and Bullying and Harassment appear only rarely.

These patterns indicate that the Board’s docket is structured around politically salient and normatively contested speech disputes. Case selection therefore reflects an adjudicative logic oriented toward disputes with broader expressive and public-interest implications rather than proportionally mirroring the overall distribution of platform enforcement actions. This selective attention reinforces the view that oversight institutions function less as comprehensive review mechanisms than as agenda-setting venues that define which categories of speech become sites of global governance debate.

5.5 Language, Policy Domains, and Issue Types

The topical concentration documented above does not appear to be driven by language-specific issue types. As illustrated in Figure 5, the distribution of the most frequent Community Standards categories is broadly similar across English and non-English cases. High-salience categories such as Hate Speech, Dangerous Individuals and Organizations, and Violence and Incitement dominate the docket in both linguistic contexts, while lower-frequency categories remain marginal across languages.

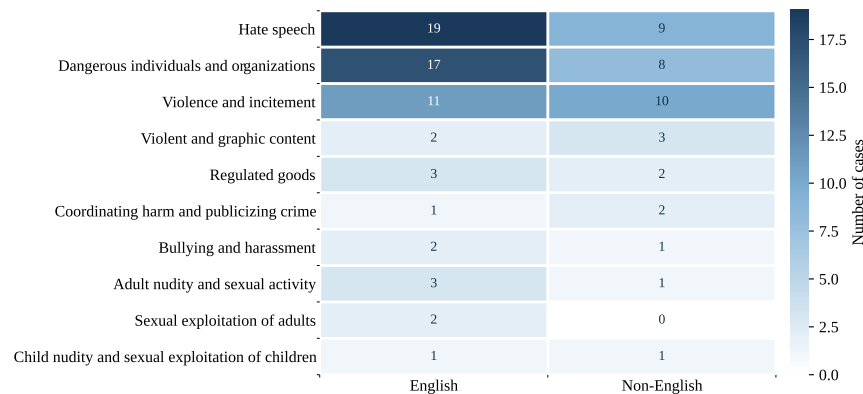


Figure 5: Oversight Board cases by Community Standard and language context (top ten standards).

Linguistic representation therefore does not correspond to differentiated issue domains. Instead, the same set of politically salient policy categories structures oversight cases across both linguistic contexts.

That the same high-stakes categories dominate across linguistic contexts indicates that the Board’s agenda is organized around policy domain rather than language: particular

kinds of speech disputes are prioritized for review regardless of the language in which they arise. The Board's selective attention is thus keyed to substantive policy salience rather than to the linguistic composition of the underlying disputes.

5.6 Language and Adjudicative Outcomes

Linguistic context likewise does not appear to shape adjudicative outcomes, once summary decisions are set aside. Because a summary decision is, by design, a case in which Meta reversed its own enforcement error after the Board flagged it, such decisions are almost uniformly recorded as reversals (60 of the 61 in the corpus); including them would conflate Meta's self-correction with the Board's adjudication. Restricting the comparison to the 72 fully reviewed cases with a definitive overturn-or-uphold outcome, overturn rates are broadly comparable across linguistic contexts—64% for English-language cases and 73% for non-English cases—and the difference is not statistically significant, $\chi^2(1) = 0.61$, $p = .434$, Cramér's $V = .09$ (Figure 6). A parallel comparison by geopolitical classification yields the same result: overturn rates do not differ significantly between Global South and other cases, $\chi^2(1) = 0.03$, $p = .872$. Both null findings are confirmed by Fisher's exact test and hold under the alternative language codings described earlier.

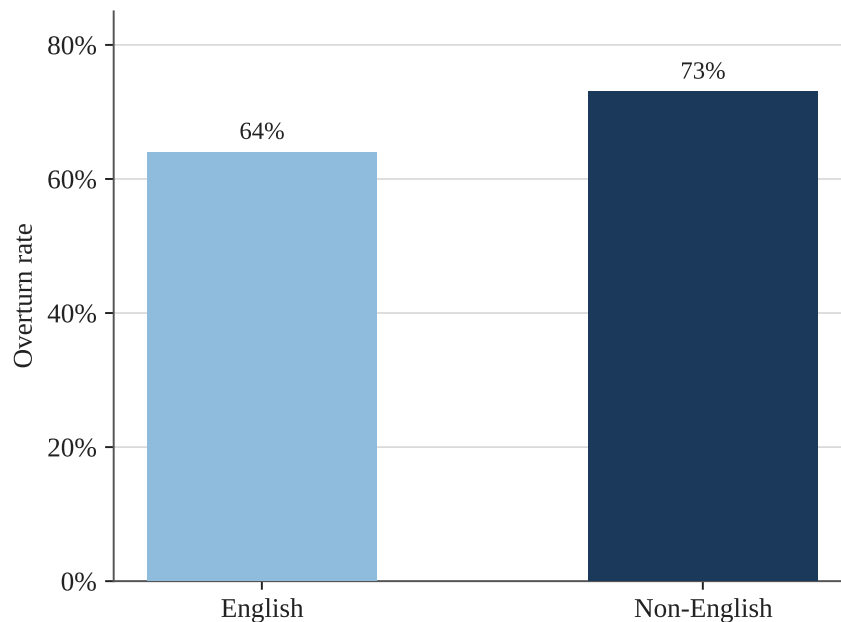


Figure 6: Overturn rate by language context among fully reviewed cases.

Once a dispute receives full panel review, its outcome does not systematically vary between English and non-English contexts.

Although language distribution within the docket reflects broader geopolitical and infrastructural dynamics, these differences do not translate into divergent adjudicative

outcomes. In other words, while the composition of cases entering oversight exhibits linguistic structuring, the Board’s final decisions appear broadly consistent across linguistic contexts. This result suggests that linguistic asymmetries may shape which cases reach formal oversight, but not how those cases are ultimately adjudicated.

5.7 Procedural Pathways as Institutional Differentiators

Although adjudicative outcomes do not vary by language, procedural routing does. A chi-square test indicates a statistically significant association between language context and procedural pathway (full review vs. summary decision), $\chi^2(1) = 6.62$, $p = .010$, Cramér’s $V = .21$. This result suggests a moderate relationship between linguistic context and how cases move through the Board’s review structure.

As illustrated in Figure 7, English-language cases are roughly evenly distributed between full review and summary decisions, whereas non-English cases are substantially less likely to proceed through summary pathways and more frequently receive full panel review.

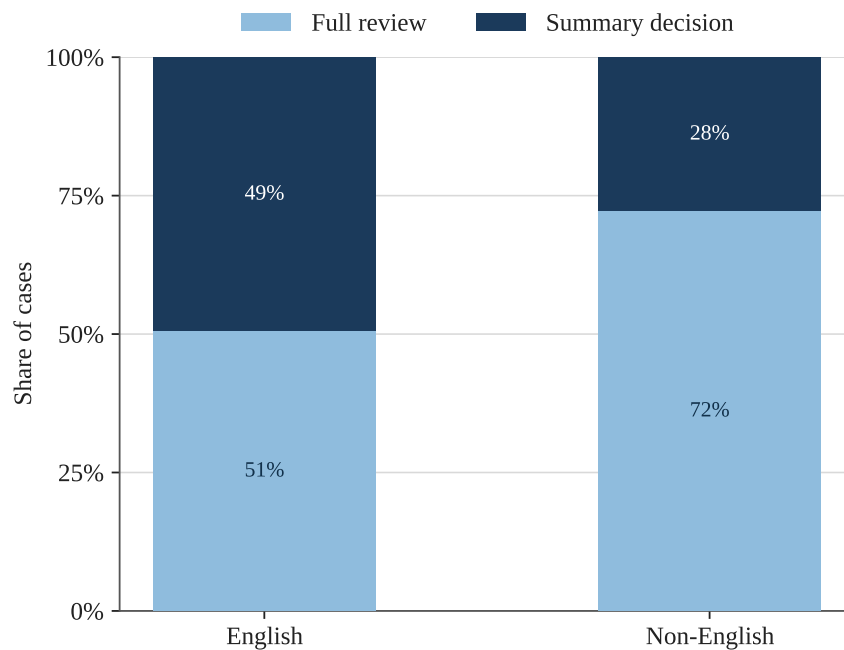


Figure 7: Procedural pathway by language context.

Because a summary decision is defined as an instance in which Meta reversed its action following Board engagement, and such reversals are recorded as overturns, procedural pathway cannot be treated as an independent predictor of overturn without circularity; the pattern is therefore reported descriptively. Summary decisions account for 61 of 147 cases (41.5%), nearly all of which are reversals (60 of 61), while the remaining 86 cases (58.5%) proceed to full Board review. As reported in Section 5.6, this null holds when attention is restricted to full-review cases: neither language nor Global South

classification is significantly associated with overturn, a result robust across alternative language codings and confirmed by Fisher's exact test. Among genuinely reviewed cases, adjudicative outcomes thus appear largely independent of the identity characteristics examined here.

These findings suggest that the Board's institutional architecture—particularly its procedural design—plays a more consequential role in structuring adjudicative patterns than linguistic or geopolitical identity. In other words, differences in how cases are processed appear to matter more than differences in where those cases originate. This pattern reinforces the view that platform oversight operates through structured procedural filters that shape which disputes receive deeper adjudication and how frequently platform decisions are reversed.

5.8 Integrated Institutional Pattern

The results reveal a layered institutional structure. Geographic representation is globally dispersed but internally concentrated. Linguistic diversity is geopolitically structured, with non-English cases disproportionately situated within Global South contexts. Topical concentration reflects prioritization of high-stakes expressive disputes. Policy domains are not linguistically segregated, and adjudicative outcomes do not differ by language or geopolitical classification. Procedural routing, however, varies significantly by language: English-language disputes are disproportionately resolved through summary reversal, whereas non-English disputes more often undergo full panel review.

The Board's representational structure therefore cannot be reduced to a simple Global North bias or linguistic marginalization—indeed, Global South contexts form the largest share of the docket. Instead, disparities emerge primarily at the level of institutional process rather than final decision. Procedural architecture—not geography or language alone—appears to shape how representation is distributed and experienced within platform governance.

This layered pattern further suggests that potential inequalities in platform governance may arise less from adjudicative outcomes themselves than from the institutional pathways through which disputes enter and move through oversight. In this sense, oversight institutions structure governance through successive stages of case selection, procedural routing, and adjudication, each of which shapes how disputes become visible and how institutional attention is ultimately allocated.

6 Discussion

This study examined the representational structure of formal platform governance through a case-level analysis of decisions issued by the Board. Rather than relying on moderation outputs or aggregate transparency reports, the study treated the Board's

official archive as an institutional record of adjudication. This approach makes visible how disputes become formal cases, how issues are prioritized, and how institutional procedures structure outcomes.

The findings indicate that representational patterns in platform oversight are structured rather than random. Geographic representation is globally dispersed but internally stratified. Linguistic diversity is present yet geopolitically patterned. Policy attention is concentrated in high-stakes expressive domains. Most notably, among fully reviewed cases adjudicative outcomes do not vary significantly by language or geopolitical classification; where language leaves its clearest mark is in procedural routing rather than in outcomes. Representation, therefore, appears to be shaped more by institutional architecture than by identity categories alone.

These results extend platform governance scholarship by reframing legitimacy as an institutional design problem rather than solely a distributional one. Existing research often evaluates fairness through outcome disparities—who benefits, who is disadvantaged, and which groups appear marginalized (Gillespie 2018; Douek 2022). The present findings suggest that legitimacy also depends on how cases enter oversight, how they are routed through institutional procedures, and which disputes receive full adjudication. Institutional process, not only substantive outcome, becomes central to governance equity (Balkin 2018; Klonick 2020).

Conceptually, the Oversight Board produces a form of quasi-observational institutional data. Like the docket of the Supreme Court of the United States, the archive records the subset of disputes that reach formal review rather than the full universe of underlying conflicts. The visible record therefore captures institutional attention—what issues are elevated, which contexts become legible, and how representation is constructed through procedural inclusion.

Observable patterns in the docket do not necessarily mirror the full distribution of platform disputes; they reflect the filtering mechanisms of institutional design. If formal oversight disproportionately addresses particular topics, regions, or procedural routes, the record signals structural priorities embedded within governance systems. Absence may therefore be as analytically meaningful as presence.

The findings also complicate simplistic narratives of bias. The absence of outcome disparities across language and geopolitical categories is consistent with procedural neutrality in final decisions, though the small number of fully adjudicated cases warrants caution. Yet disparities emerge in institutional routing, indicating that representation operates through process rather than verdict. Equity in platform governance thus cannot be judged by outcome parity alone; procedural access matters too.

From a legitimacy standpoint, this layered structure aligns with theories of institutional accountability in platform governance. Legitimacy depends not only on fairness of outcomes but also on transparency of procedures, inclusivity of representation, and

consistency of institutional reasoning (Balkin 2018; Kaye 2019). Institutional analyses of the Oversight Board similarly emphasize its role as a procedural governance mechanism designed to increase transparency and accountability within private content moderation systems (Wong and Floridi 2023; Price and Price 2023). The Board's archive provides rare empirical access to these dimensions within private digital governance.

The Oversight Board thus functions as a hybrid institutional actor: part appellate body, part policy laboratory, and part transparency mechanism. The resulting archive offers a rare window into how platform rules are enforced in practice, and how institutional design shapes the distribution of attention and authority.

Although this analysis centers on the Board, its implications extend to the wider field of platform oversight mechanisms—internal appeals systems, trusted-flagger programs, and external bodies such as TikTok's advisory councils or the co-regulatory structures emerging under the European Union's Digital Services Act. Wherever such institutions publish adjudicative records, the same interpretive caution applies: the docket reflects institutional attention rather than the underlying distribution of harm. Reading these records this way directs audit attention upstream—from outcomes toward case selection and routing (cf. Gillespie 2018; Douek 2022)—and raises comparative questions for future work: whether similar routing asymmetries recur in other oversight bodies, and whether they narrow as multilingual review capacity improves.

6.1 Conclusion

This study analyzed the representational structure of platform oversight through a systematic examination of the Board's decisions from 2020 to 2025. By treating the Board's archive as an institutional dataset, the analysis revealed how governance disputes become formal cases and how institutional procedures shape their adjudication. The findings show that geographic and linguistic representation within the docket is globally distributed yet internally stratified, while policy attention concentrates on high-stakes expressive disputes. Although adjudicative outcomes among fully reviewed cases do not vary significantly across linguistic or geopolitical contexts, procedural routing does: language shapes how disputes move through the Board's review structure more than whether they are ultimately reversed.

These results suggest that inequalities in platform governance may emerge less from final decisions than from the institutional pathways through which disputes become visible and reviewable.

The Board's decision archive provides a rare empirical record of platform rule enforcement. As such oversight systems proliferate across platforms and regulatory regimes, understanding how institutional design structures representation and accountability will be central to their legitimacy.

References

- Agresti, Alan. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.
- Ayalew, Yohannes Eneyew. 2025. "Linguistic Blind Spots in Platform Content Moderation in Ethiopia: The Case of Facebook." *Modern Languages Open* 2025 (1). <https://doi.org/10.3828/mlo.v0i0.468>.
- Bachrach, Peter, and Morton S. Baratz. 1962. "Two Faces of Power." *American Political Science Review* 56 (4): 947–52. <https://doi.org/10.2307/1952796>.
- Balkin, Jack M. 2018. "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *UC Davis Law Review* 51 (3): 1149–210.
- Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity.
- Blakey, Elizabeth. 2024. "The Day Data Transparency Died: How Twitter/X Cut Off Access for Social Research." *Contexts* 23 (2): 30–35. <https://doi.org/10.1177/15365042241252125>.
- Bruns, Axel. 2019. "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research." *Information, Communication & Society* 22 (11): 1544–66. <https://doi.org/10.1080/1369118X.2019.1637447>.
- Dados, Nour, and Raewyn Connell. 2012. "The Global South." *Contexts* 11 (1): 12–13. <https://doi.org/10.1177/1536504212436479>.
- De Gregorio, Giovanni, and Nicole Stremlau. 2023. "Inequalities and Content Moderation." *Global Policy* 14 (5): 870–79. <https://doi.org/10.1111/1758-5899.13243>.
- Douek, Evelyn. 2019. "Facebook's 'Oversight Board': Move Fast with Stable Infrastructure and Humility." *North Carolina Journal of Law & Technology* 21 (1).
- . 2021. "Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability." *Columbia Law Review* 121 (3): 759–834.
- . 2022. "Content Moderation as Systems Thinking." *Harvard Law Review* 136 (2): 526–607. <https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf>.
- . 2024. "The Meta Oversight Board and the Empty Promise of Legitimacy." *Harvard Journal of Law & Technology* 37 (2): 373–445.
- Elsawah, Mona, Aliya Bhatia, and Dhanaraj Thakur. 2025. *Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages*. Center for Democracy & Technology. <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/>.

- Evangelista, Rafael, and Fernanda Bruno. 2019. "WhatsApp and Political Instability in Brazil: Targeted Messages and Political Radicalisation." *Internet Policy Review* 8 (4): 1–23. <https://doi.org/10.14763/2019.4.1434>.
- Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–68. <https://doi.org/10.1080/10584609.2018.1477506>.
- Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press. <https://doi.org/10.12987/9780300235029>.
- Gorwa, Robert. 2019. "What Is Platform Governance?" *Information, Communication & Society* 22 (6): 854–71. <https://doi.org/10.1080/1369118X.2019.1573914>.
- . 2022. *Who Are the Stakeholders in Platform Governance?* SocArXiv. <https://doi.org/10.31235/osf.io/ayx8h>.
- . 2024. *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. New York: Oxford University Press. <https://doi.org/10.1093/osq/9780197692851.001.0001>.
- Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1). <https://doi.org/10.1177/2053951719897945>.
- Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.
- Helfer, Laurence R., and Molly K. Land. 2023. "The Meta Oversight Board's Human Rights Future." *Cardozo Law Review* 44 (6): 2233–300.
- Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–93. <https://doi.org/10.18653/v1/2020.acl-main.560>.
- Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.
- Kazemi, Ashkan, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A. Hale. 2022. "Research Note: Tiplines to Uncover Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian General Election on WhatsApp." *Harvard Kennedy School Misinformation Review*, <https://doi.org/10.37016/mr-2020-91>.
- Kettemann, Matthias C., and Wolfgang Schulz. 2023. *Platform://Democracy: Perspectives on Platform Power, Public Values and the Potential of Social Media Councils*. Leibniz Institute for Media Research, Hans-Bredow-Institut.

- Klonick, Kate. 2019. "Inside the Team at Facebook That Dealt with the Christchurch Shooting." *The New Yorker* (April 25, 2019). <https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting>.
- . 2020. "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression." *Yale Law Journal* 129:2418–99. <https://yalelawjournal.org/feature/the-facebook-oversight-board>.
- Marres, Noortje, and Esther Weltevrede. 2013. "Scraping the Social? Issues in Live Social Research." *Journal of Cultural Economy* 6 (3): 313–35. <https://doi.org/10.1080/17530350.2013.772070>.
- Milan, Stefania, and Emiliano Treré. 2019. "Big Data from the South(s): Beyond Data Universalism." *Television & New Media* 20 (4): 319–35. <https://doi.org/10.1177/1527476419837739>.
- Mulligan, Deirdre K., and Kenneth A. Bamberger. 2021. "Allocating Responsibility in Content Moderation: A Functional Framework." *Berkeley Technology Law Journal* 36 (3): 1091–172. <https://doi.org/10.15779/Z383B5W872>.
- Nahmias, Yifat, and Maayan Perel. 2021. "The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations." *Harvard Journal on Legislation* 58 (1): 145–94. https://journals.law.harvard.edu/jol/wp-content/uploads/sites/86/2021/02/105_Nahmias.pdf.
- Napoli, Philip M. 2019. *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. New York: Columbia University Press. <https://doi.org/10.7312/napo18454>.
- Nicholas, Gabriel, and Aliya Bhatia. 2023. "Toward Better Automated Content Moderation in Low-Resource Languages." *Journal of Online Trust and Safety* 2 (1). <https://doi.org/10.54501/jots.v2i1.150>.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press. <https://doi.org/10.18574/nyu/9781479833641.001.0001>.
- Nosthoff, Anna-Verena, and Felix Maschewski. 2024. "The Platform Economy's Infrastructural Transformation of the Public Sphere: Facebook and Cambridge Analytica Revisited." *Philosophy & Social Criticism* 50 (1): 178–99. <https://doi.org/10.1177/01914537231203536>.
- Nothias, Toussaint. 2020. "Access Granted: Facebook's Free Basics in Africa." *Media, Culture & Society* 42 (3): 329–48. <https://doi.org/10.1177/0163443719890530>.
- O'Brien, Matt, and Barbara Ortutay. 2022. "Musk's Twitter Disbands Its Trust and Safety Advisory Group." Associated Press, December 12, 2022. Accessed June 1, 2025. <https://apnews.com/article/elon-musk-twitter-trust-safety-council-disbanded-675db53734fc005d4b9c58f3c7a74886>.

- Ocasio, William. 1997. "Towards an Attention-Based View of the Firm." *Strategic Management Journal* 18 (S1): 187–206.
- Oversight Board. 2021. "Case Decision 2021-001-FB-FBR: Former President Trump's Suspension," May 5, 2021. Accessed June 1, 2025. <https://www.oversightboard.com/decision/FB-691QAMHJ/>.
- . 2022. "Policy Advisory Opinion on Meta's Cross-Check Program," December 6, 2022. Accessed June 1, 2025. <https://www.oversightboard.com/decision/pao-nr730ofi/>.
- . 2025. "From Bold Experiment to Essential Institution," December 4, 2025. <https://www.oversightboard.com/wp-content/uploads/2025/12/Oversight-Board-From-Bold-Experiment-To-Essential-Institution-December-2025-2.pdf>.
- Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press. <https://doi.org/10.4159/harvard.9780674736061>.
- Price, Monroe E., and Joshua M. Price. 2023. "Building Legitimacy in the Absence of the State: Reflections on the Facebook Oversight Board." *International Journal of Communication* 17:3315–25.
- Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press. <https://doi.org/10.12987/9780300245318>.
- Rogers, Richard. 2025. "On Methods for Content Moderation Research." In *Content Moderation across Social Media Platforms*, edited by Richard Rogers, 4–22. Amsterdam: Digital Methods Initiative, University of Amsterdam. https://www.digitalmethods.net/pub/Dmi/ContentModerationStudy/Content_Moderation_UvA_23June_2025_opti.pdf.
- Schattschneider, E. E. 1960. *The Semisovereign People: A Realist's View of Democracy in America*. New York: Holt, Rinehart / Winston.
- Singh, Siddharth, Vedant Nanda, Rijurekha Sen, Sohaib Ahmad, Satadal Sengupta, Amreesh Phokeer, Zaid Ahmed Farooq, et al. 2017. "An Empirical Analysis of Facebook's Free Basics." In *Proceedings of the 2017 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, 37–38. <https://doi.org/10.1145/3078505.3078554>.
- Sinpeng, Aim, Fiona Martin, Katharine Gelber, and Kirril Shields. 2021. *Facebook: Regulating Hate Speech in the Asia Pacific*. University of Sydney and University of Queensland. <https://doi.org/10.25910/J09V-SQ57>.
- Spotify. 2022. "Introducing the Spotify Safety Advisory Council." Spotify Newsroom, June 13, 2022. Accessed June 1, 2025. <https://newsroom.spotify.com/2022-06-13/introducing-the-spotify-safety-advisory-council/>.

- Suchman, Mark C. 1995. "Managing Legitimacy: Strategic and Institutional Approaches." *Academy of Management Review* 20 (3): 571–610. <https://doi.org/10.5465/amr.1995.9508080331>.
- Suzor, Nicolas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge University Press.
- Suzor, Nicolas P., Tess Van Geelen, and Sarah Myers West. 2018. "Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda." *International Communication Gazette* 80 (4): 385–400. <https://doi.org/10.1177/1748048518757142>.
- TikTok. 2024. "Engaging Our Advisory Councils." TikTok Transparency Center. Accessed June 1, 2025. <https://www.tiktok.com/transparency/en/advisory-councils>.
- United Nations Human Rights Council. 2019. "Independent International Fact-Finding Mission on Myanmar." OHCHR. <https://www.ohchr.org/en/hr-bodies/hrc/myanmar-ffm/index>.
- Wong, David, and Luciano Floridi. 2023. "Meta's Oversight Board: A Review and Critical Assessment." *Minds and Machines* 33 (2): 261–84. <https://doi.org/10.1007/s11023-022-09613-x>.
- Ye, Meng, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. "Multilingual Content Moderation: A Case Study on Reddit." In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3828–44. <https://doi.org/10.18653/v1/2023.eacl-main.276>.

Authors

Endalkachew H. Chala, PhD is an independent researcher based in the Twin Cities, Minnesota; a Community Fellow at the Center for an Informed Public, University of Washington; a faculty affiliate at Arba Minch University, Ethiopia; and formerly an assistant professor at Hamline University. His research examines platform governance, content moderation, and multilingual online discourse. ORCID: 0000-0001-6210-6706. Email: endalk2006@gmail.com.

Acknowledgments

Not applicable.

Data Availability Statement

The corpus of Oversight Board decisions analyzed in this study was constructed from the Board's publicly available decision archive. Replication data, codebook, and analysis scripts will be deposited in a public repository upon publication and are available from the author on reasonable request in the interim.

Funding Statement

The study received no external funding.

Disclosure Statement

The author reports no conflict of interest. The author has no financial, professional, or personal relationship with Meta Platforms, Inc., the Oversight Board, or any other organization discussed in this article that could be perceived to have influenced the research or its findings.

Ethical Standards

This study relied exclusively on publicly available institutional records (published Oversight Board decisions) and did not involve human subjects; it was therefore exempt from institutional review board review.

Keywords

Content moderation; platform governance; Meta Oversight Board; multilingual moderation.

Appendices

Appendix A: Country Classification by Region and Global North/South

Country mentions were extracted from each decision's location metadata and classified by world region and by Global North/South status. Across the corpus, 118 of 198 country-level mentions (60%) reference Global South contexts. Table A1 summarizes mentions by region; Tables A2 and A3 list the full classification for all 65 countries, split by Global North/South status.

Table A1: Case mentions by world region. Counts are country-level mentions extracted from decision location metadata, not distinct cases.

World region	Countries (n)	Case mentions
MENA	11	43
Europe	16	42
North America	2	33
Latin America & Caribbean	10	21
Sub-Saharan Africa	10	21
South Asia	5	17
Southeast Asia	7	13
East Asia	3	5
Oceania	1	3

Table A2: Country classification: Global North. The 15 Global North countries appearing in decision location metadata, in alphabetical order.

Country	World region	Mentions
Australia	Oceania	3
Belgium	Europe	1
Canada	North America	2
France	Europe	8
Germany	Europe	5
Greece	Europe	2
Israel	MENA	11
Italy	Europe	1
Japan	East Asia	2
Netherlands	Europe	2
South Korea	East Asia	1
Spain	Europe	2
Sweden	Europe	2
United Kingdom	Europe	7
United States	North America	31

Table A3: Country classification: Global South. The 50 Global South countries appearing in decision location metadata, in alphabetical order.

Country	World region	Mentions
Afghanistan	South Asia	3
Albania	Europe	1
Argentina	Latin America & Caribbean	2
Armenia	MENA	2
Azerbaijan	MENA	2
Bangladesh	South Asia	1
Brazil	Latin America & Caribbean	5
Cambodia	Southeast Asia	1
China	East Asia	2
Colombia	Latin America & Caribbean	3
Croatia	Europe	1
Cuba	Latin America & Caribbean	1
Ecuador	Latin America & Caribbean	1
Egypt	MENA	2
Eritrea	Sub-Saharan Africa	1

Table A3: Country classification: Global South. *(continued)*

Country	World region	Mentions
Ethiopia	Sub-Saharan Africa	3
Guinea-Bissau	Sub-Saharan Africa	1
Guyana	Latin America & Caribbean	1
Haiti	Latin America & Caribbean	1
India	South Asia	6
Indonesia	Southeast Asia	1
Iran	MENA	3
Iraq	MENA	2
Kenya	Sub-Saharan Africa	2
Latvia	Europe	1
Lebanon	MENA	2
Malaysia	Southeast Asia	2
Mexico	Latin America & Caribbean	4
Morocco	MENA	1
Myanmar	Southeast Asia	3
Niger	Sub-Saharan Africa	1
Nigeria	Sub-Saharan Africa	3
Pakistan	South Asia	5
Palestinian Territories	MENA	10
Peru	Latin America & Caribbean	2
Philippines	Southeast Asia	3
Poland	Europe	3
Russia	Europe	3
Senegal	Sub-Saharan Africa	2
Serbia	Europe	1
Singapore	Southeast Asia	1
Somalia	Sub-Saharan Africa	3
South Africa	Sub-Saharan Africa	2
Sri Lanka	South Asia	2
Sudan	Sub-Saharan Africa	3
Syria	MENA	4
Thailand	Southeast Asia	2
Turkey	MENA	4
Ukraine	Europe	2
Venezuela	Latin America & Caribbean	1