Media and Communication
2026 • Volume 14 • Article 12653
https://doi.org/10.17645/mac.12653

cogitatio

ARTICLE

Open Access Journal

From Detection to Counterspeech: Auditing AI Moderation and Fact-Checking Practices in Ethiopia's Multilingual Online Sphere

Endalkachew H. Chala [1,2]

[1] Independent Researcher, USA
[2] Center for an Informed Public, University of Washington, USA

Correspondence: Endalkachew H. Chala (endalk2006@gmail.com)

Submitted: 27 April 2026 **Accepted:** 20 July 2026 **Published:** 8 September 2026

Issue: This article is part of the issue "The Role of AI for Counter Speech: Detection, Intervention, and Risks" edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at https://doi.org/10.17645/mac.i500

Abstract

AI-assisted content moderation promises to manage online incivility at scale while sparing human moderators traumatic content. Yet most evidence comes from Western languages and contexts. This study examines how AI hate-speech detection interacts with counterspeech and fact-checking in Ethiopia's polarized Amharic and Afan Oromo spaces. Drawing on infrastructural invisibility and civic labor, it triangulates a computational audit of three classifiers against 838 hand-annotated posts (2020–2025), document analysis of platform self-descriptions and transparency reports, and 20 interviews with Ethiopian fact-checkers, volunteer flaggers, and counter-speakers. The most widely used generic classifier cannot read either language natively and, on English translations, recovers only about a tenth of hate speech; locally-oriented classifiers perform far better on Amharic but collapse on Afan Oromo, leaving it effectively unserved. Across every tool, the dominant error is under-detection, not over-removal, and the same silence recurs in platforms' self-descriptions. Rather than sparing anyone the work, AI's failure displaces it onto an unpaid volunteer ecology that absorbs political and psychological costs platforms benefit from yet refuse to name.

Keywords

Afan Oromo; Amharic; civic labor; counterspeech; Ethiopia; hate speech detection; infrastructural invisibility; low-resource NLP

1. Introduction

AI-enabled content moderation is premised on structural scalability: one classifier, trained once, deployed across markets (Gillespie, 2018; Gorwa et al., 2020). In practice, coverage is uneven, tracking linguistic and institutional power (Joshi et al., 2020; Sap et al., 2019). Ethiopia is a sharp case. Between 2020 and 2022, as the Tigray conflict produced a major humanitarian emergency, civil society, journalists, and researchers documented platforms' failure to moderate inciting content in Amharic and Afan Oromo—Ethiopia's two most widely spoken languages and the primary languages of its online public sphere (Amnesty International, 2023; Azoulay, 2024). Amharic is the federal working language and a long-standing *lingua franca*; Afan Oromo has the most first-language speakers and is increasingly a language of wider communication. Both remain markedly under-resourced in the commercial datasets that train moderation systems. These accounts converged on one pattern: Hate speech spread faster than platform systems identified or acted on it.

Two bodies of scholarship bear on this failure but rarely meet: Research on AI content moderation documents the linguistic limits and biases of automated systems, largely within high-resource languages, while research on counterspeech, fact-checking, and civic labor theorizes human moderation but rarely examines its encounter with automation in conflict-affected, low-resource settings (Gray & Suri, 2019; Roberts, 2019). Ethiopia sits at their intersection: two under-resourced languages, a conflict-riven political context that made AI moderation urgently consequential, and a volunteer ecology whose moderation signal platforms absorb without acknowledgment.

This study examines that intersection through a mixed-methods audit guided by two complementary frames. Infrastructural invisibility (Plantin et al., 2018; Star & Ruhleder, 1996) names the condition in which working systems become legible only at breakdown; civic labor (Gray & Suri, 2019; Roberts, 2019) names the underrecognized human work platforms depend on but do not acknowledge. Together they define the study's objective: to establish what publicly deployable moderation AI does on Ethiopian-language content; how platforms represent, and obscure, that work; and how the human ecology absorbs its failures. The theoretical framework Section 2 below develops these frames; the literature review then derives three research questions (RQs) that operationalize this objective.

The audit finds publicly deployable classifiers failing unevenly and consistently toward under-detection rather than over-removal, with Afan Oromo least served; document analysis finds platforms reproducing the same silence in their self-descriptions; and interviews reveal a volunteer ecology performing the moderation work platform AI does not, at political and economic cost. The study extends infrastructural invisibility and civic labor to a conflict-affected, low-resource setting, showing that civic resilience rests on a politically contested labor that platforms simultaneously depend on and refuse to name.

2. Theoretical Framework

This study analyzes AI-assisted moderation as a single arrangement in which automated systems and human labor share the work, viewed through two complementary frames: infrastructural invisibility and civic labor. The frames view the same situation from opposite ends: One explains why moderation fails unnoticed, the other who absorbs the cost when it does.

2.1. Infrastructural Invisibility

Infrastructures become legible only in breakdown, when systems on which routine action depends fail and expose their operation (Star & Ruhleder, 1996). Plantin et al. (2018) extend this to platforms, which present themselves as open services while functioning as backbones of public communication. The frame names two conditions this analysis documents.

The first is capability absence: A classifier may not operate on a language at all—an absence coded into the infrastructure, visible only under direct examination. A platform deploying such a tool encounters not an error but zero outputs, which aggregate statistics cannot distinguish from an absence of violative material. The failure is thus structurally invisible.

The second is discursive silence: Platforms may describe moderation in global terms while saying nothing about particular languages, and transparency reports reveal only what platforms choose to disclose, so their silences are themselves infrastructural (Ananny & Crawford, 2018). X's 2022 retreat from periodic transparency reporting exemplifies this: Removing a disclosure regime does not merely create a data gap but renders a partly visible infrastructure fully invisible, with implications beyond the non-disclosure period.

The frame also has methodological force. External algorithmic audits force infrastructures into breakdown, submitting inputs that trigger errors and measuring outputs against ground truth (Bandy, 2021; Sandvig et al., 2014). The study's three strands each enact this logic, together producing a fuller picture of the moderation infrastructure than any single register could supply.

2.2. Civic Labor and its Metabolism

Gray and Suri (2019) introduce "ghost work" for human labor behind ostensibly automated operations: paid minimally, fragmented into micro-tasks, and invisible in platform self-presentation. Roberts (2019) documents commercial moderation's psychological and political costs. This study adapts these accounts to a different configuration: In Ethiopia, moderation is sustained not only by paid moderators but by a diffuse civic ecology of volunteer administrators, fact-checkers, and counter-speakers whose work is then absorbed by the platforms, uncredited and unpaid (cf. Gibson, 2023).

The frame requires two refinements here. The first is political heterogeneity. Unlike the relatively coherent labor of Reddit moderators (Matias, 2019; Seering et al., 2019), Ethiopian volunteer flagging is politically contested: Administrators flagging hate and misinformation operate alongside rival groups that mobilize coordinated mass-reporting (Meisner, 2023). Such campaigns sometimes target genuine violations, but at other times they target content that reporters merely dislike or politically oppose. This heterogeneity refines rather than undermines the frame: Platforms metabolize the labor as an undifferentiated signal, feeding reputation systems and retraining pipelines regardless of whether a flag is civic-minded, adversarial, or counterspeech (Crawford & Gillespie, 2016). The labor's political provenance is erased the moment it becomes platform input.

The second refinement concerns compensation. Where ghost work emphasizes economic extraction—workers paid below livable rates—the Ethiopian extraction is infrastructural rather than economic: the labor

reduces the load on platforms' automated and paid-moderation systems and supplies a signal for system improvement, yet goes uncredited. Volunteers absorb the psychological and political risks—exposure to hate content, retaliation, and the affective load of sustained engagement in polarized environments (Pinchevski, 2023; Roberts, 2019; Steiger et al., 2021)—while platforms gain the operational benefit. The frame names this as civic-labor extraction particular to low-resource, conflict-affected settings.

Each frame addresses what the other cannot. Infrastructural invisibility explains why the Ethiopian moderation gap persists even as platforms nominally commit to Ethiopian content; civic labor explains how the gap is partly absorbed in practice—the work platform AI does not do is performed, imperfectly and at cost, by humans who fill the infrastructural void. That absorption is incomplete: Harmful content still circulates, and the volunteer ecology mitigates harm without resolving it. Section 6 returns to this pairing: Civic resilience rests on a configuration unsustainable for volunteers and invisible to the platforms it sustains.

3. Literature Review

With these frames in view, the review situates the study within two largely parallel bodies of scholarship.

3.1. AI Content Moderation and its Linguistic Limits

Automated content moderation is the dominant paradigm for managing user-generated content at scale (Gillespie, 2018; Gorwa et al., 2020). Automation and AI are analytically distinct: Automation executes predefined rules (Endsley & Kaber, 1999; Hitomi, 1994), whereas AI infers patterns and generalizes novel inputs. Contemporary moderation combines both; this study reserves "AI moderation" for the learned-classifier layer. Following Sponholz (2021), the study treats hate speech not as a fixed legal category but as discourse attacking or demeaning groups on protected characteristics, its boundaries contested in public deliberation. These systems sit within a broader governance shift: Platforms have retreated from costly human moderation toward automated, increasingly deregulated arrangements, framing AI as the remedy for problems automation itself produces (Jacobsen, 2025; Katzenbach, 2021). Generic toxicity classifiers, including Google's Perspective API, show biases against non-dominant dialects and marginalized groups (Masud et al., 2024; Sap et al., 2019) and degrade in languages underrepresented in commercial datasets (Joshi et al., 2020). This asymmetry is structural: Investment datasets and models concentrate on English and other high-resource languages. Even smaller European languages, such as Finnish or Hungarian, receive limited commercial classifier coverage (Lees et al., 2022); the Ethiopian case differs less in the fact of scarcity than in its stakes, where moderation failures intersect with active conflict and weak institutional protection. Parallel literature documents systematic over-enforcement—false positives that silence counterspeech and marginalized voices—alongside persistent under-enforcement of genuine harm (Haimson et al., 2021; Oliva et al., 2021), a pattern this study's audit examines in Ethiopian-language content.

A growing body of work develops locally adapted classifiers for lower-resource languages. AfriHate (Muhammad et al., 2025) released multilingual hate-speech datasets and per-language models for 15 African languages, including Amharic and Afan Oromo. Earlier work built corpora predominantly for Amharic (Ayele et al., 2023) and far less for Afan Oromo (Ababu & Woldeyohannis, 2022), a disparity reflecting the differential resourcing of Ethiopian languages (see Demilie & Salau, 2022). Yet these classifiers are evaluated

almost entirely on the benchmarks that produced them (Blodgett et al., 2020). This absence yields the study's first RQ:

> RQ1: How do publicly deployable classifiers perform on real-world Amharic and Afan Oromo content drawn from outside of their training distributions, and what do their errors reveal about under- and over-enforcement?

Detection is only one layer of the moderation infrastructure; how platforms publicly account for it is another. Transparency reports, policy pages, and help-center documentation form the public record of moderation, yet what they disclose about lower-resource languages has received little scrutiny. This article therefore asks:

> RQ2: How do platforms describe their AI moderation of Ethiopian-language content, and what do those descriptions leave invisible?

This study addresses both, and what neither detection nor description accounts for is absorbed by the human actors who work in the gaps they leave—the subject of Section 3.2, and the source of the study's third RQ.

3.2. Counterspeech, Fact-Checking, and the Civic Labor of Platform Maintenance

The labor of identifying and responding to harmful speech anchors a second body of scholarship on counterspeech, fact-checking, and volunteer moderation. That labor anchors a second body of scholarship on counterspeech, fact-checking, and volunteer moderation. Counterspeech—speech that responds to, challenges, or corrects hate speech—anchors research on how online communities resist exclusionary discourse (Benesch, 2014; Buerger, 2021). Fact-checking (Graves, 2016) is less reactive, more institutionally organized, and focused on specific claims rather than patterns of speech. It emerged from professional journalism, with its own accreditation and boundary work (Graves, 2018), and increasingly operates through platform partnerships that shape what is checked (Bélair-Gagnon et al., 2023; Vinhas & Bastos, 2025). This professionalized practice is distinct from the volunteer flagging and counterspeech examined here: It targets verifiable factual claims rather than hate speech (Koliska & Roberts, 2025), so hateful but non-factual content often falls outside its remit—a gap the volunteer ecology partly absorbs. Recent work examines AI's integration into both practices: AI-generated counter-messages (Mun et al., 2023), AI-assisted fact-checking tools (Dierickx et al., 2023), and tools supporting counter-speakers (Mun et al., 2024). A growing literature documents AI uptake within fact-checking organizations and its effects on routines and output (Gonçalves et al., 2024). Crowdsourced approaches such as community notes have been advanced as a scalable alternative, with mixed evidence on their adequacy (Borenstein et al., 2025; Corsi et al., 2024), while pre-bunking—inoculating audiences against manipulation techniques—offers a further complement (Lewandowsky & van der Linden, 2021). The institutional ground is shifting: Meta's 2025 move to discontinue its third-party fact-checking program in favor of community-driven moderation has destabilized the partnership model on which many fact-checkers relied (Cazzamatta, 2026).

A distinct strand examines the moderation labor that sustains platform ecosystems without equivalent recognition. Gray and Suri (2019) and Roberts (2019) document the human work on which automated systems depend, from training-data annotation to edge-case adjudication, and its systematic invisibility in platforms' self-descriptions. Volunteer moderation has been examined in Anglophone contexts (Matias,

2019; Seering et al., 2019), but analogous civic infrastructure in the Global South and low-resource-language contexts remains under-theorized. Crawford and Gillespie (2016) analyze flagging as user-produced infrastructure whose signal platforms metabolize without recognizing flaggers as part of the moderation apparatus. In Ethiopia, Ayalew (2025) documents how Facebook's linguistic blind spots—prioritizing English over major Ethiopian languages—have hindered efforts to curb hate speech and incitement. Extreme-speech scholarship situates this failure in politically charged environments where the categorization of hate, dissent, and coded speech is itself contested (Pohjonen & Gagliardone, 2017)—a contestation classifiers cannot resolve, leaving the interpretive labor of moderation to situated human judgment.

The Ethiopian case has received recent empirical treatment. Azoulay (2024), drawing on interviews and fact-check analysis from HaqCheck and Ethiopia Check, finds fact-checkers "doubly challenged"—by authoritarian constraint, conflict-driven disinformation, donor politics, and unstable international support—describing themselves as cleaning up Facebook "for free." The present study extends this account in two directions: outward, to the volunteer ecology surrounding professional fact-checkers; and inward, through a computational audit of the AI moderation tools those fact-checkers question. Across this scholarship, such labor is constitutive of platform ecosystems, not incidental to them (Crawford & Gillespie, 2016; Gray & Suri, 2019; Matias, 2019; Roberts, 2019). Yet this dynamic, theorized as "hybrid moderation" (Gorwa et al., 2020), remains undertheorized in politically polarized, low-resource settings. This omission motivates the study's third RQ:

> RQ3: How do fact-checkers, flaggers, and counter-speakers in Ethiopia's multilingual sphere use, trust, or resist AI moderation tools, and how do platforms metabolize the civic labor they produce?

The two literatures rarely meet. This study joins them by centering Ethiopia, where automation's linguistic failures and the civic labor that absorbs them coincide.

4. Method

4.1. Mixed-Methods Design

The study triangulates three empirical strands in a mixed-methods design. Each addresses a distinct register at which the platform–moderation–counterspeech nexus becomes legible: technical (computational audit), institutional (document analysis), and practitioner (semi-structured interviews). No single strand suffices; each depends on the others to establish context (Flick, 2018). The computational audit establishes what publicly deployable AI moderation tools do on Ethiopian-language content. The document analysis situates that performance against platform claims about moderation practices. The interviews surface how fact-checkers and counter-speakers experience and work around these tools. Integrative analysis in Section 6 describes the three strands against the theoretical framework. Facebook, X, and TikTok were selected across all three strands because they are among the most widely used social-media platforms in Ethiopia and the principal venues for the political discourse this study examines (Kemp, 2025).

4.2. Strand 1: Computational Audit

As an external algorithmic audit (Bandy, 2021; Sandvig et al., 2014), this strand forces classifiers into the Ethiopian-language context and measures what breaks. The audit evaluates three classifiers against a gold-standard corpus of 951 Amharic and Afan Oromo social-media posts purposively sampled from political-activist pages, influencers, and ordinary users on public Facebook, X, and TikTok (2020–2025). The corpus spans major Ethiopian political events: the Tigray conflict, the 2020 assassination of Hachalu Hundessa, Covid-19, and the conflicts in Amhara and Oromia regions. Two coders independently annotated each item; after reconciliation and the exclusion of cases not mapping onto the binary decision, 838 items were audit-eligible (Amharic $n = 470$; Afan Oromo $n = 368$). Gold labels were binarized (Hate = 1; Counter and Neutral = 0), yielding 597 Hate and 241 not-Hate items, with inter-annotator agreement (three-class $\kappa = .971$).

Three classifiers were evaluated, representing distinct tiers of language specificity: perspective API (generic), a fine-tuned AfroXLMR model trained on the AfriHate dataset (Africa-centric; Muhammad et al., 2025), and Amharic mBERT (language-specific). Perspective API, which supports neither language, was evaluated in native mode and on human English translations; the Africa-centric model was fine-tuned for this study because no deployable checkpoint is released; and Amharic mBERT was applied to the Amharic subset only. Performance was assessed using precision, recall, and $F1$ against the binarized gold standard, disaggregated by language. Over-enforcement was operationalized as false positives on Counter or Neutral items and under-enforcement as false negatives on Hate items.

4.3. Strand 2: Document Analysis

This strand reads silences in platforms' self-description as infrastructural features, reconstructing how Meta's Facebook, along with X and TikTok, describe their AI-enabled moderation of Ethiopian-language content and what they leave invisible. For this strand, the same three platforms were selected because each publishes the transparency and policy documentation that makes self-description analytically accessible. Document analysis is appropriate because public-facing platform materials are both the primary record of self-presentation and a field of strategic silence (Bowen, 2009). Sources comprise platforms' self-descriptions and transparency reports, Meta Oversight Board decisions on AI and automation, and civil-society, academic, and journalistic materials that triangulate platform claims (Amnesty International, 2023; Azoulay, 2024). Analysis uses a thematic coding scheme focused on platform accountability and the visibility of Ethiopian-language moderation.

4.4. Strand 3: Semi-Structured Interviews

This strand elicits from practitioners the friction their encounter with platform infrastructure produces, drawing a stratified purposive sample (Patton, 2015) of Ethiopian fact-checkers, volunteer flaggers, and counter-speakers. Twenty-one interviews were conducted, with one excluded after a recording fault left it partially inaudible and the participant, declining a follow-up, asked that the data be withdrawn, leaving 20 drawn from across the main venues of Ethiopian counter-hate activity—Facebook page administrators, TikTok counterspeech creators, civic X accounts, and fact-checking organizations—identified through snowball sampling (Noy, 2008) from publicly documented seed pages. These interview participants are

distinct from the public accounts whose posts were sampled for the computational audit. The sample was constructed to include politically heterogeneous participants—individuals aligned with opposing ethnic–political positions in Ethiopia's polarized online sphere, including administrators of opposed pages, to surface the rival-flagging dynamic. Interviews were conducted between November 2025 and June 2026. Of the 20 participants, 12 worked primarily as fact-checkers and four as counter-speakers, with the remaining four active in both roles; flagging was an activity shared across the sample rather than a separate specialization. In the findings, participants are identified by pseudonymous first names with a role descriptor (e.g., Gemechu, fact-checker). Because the Ethiopian counter-hate ecology is small and security-conscious, the sample does not claim exhaustive coverage: It was constructed for analytic range across roles, platforms, and political alignment, following a purposive rather than probability logic; the aim is transferability rather than statistical generalization, and the most security-exposed actors are likely underrepresented.

Interviews followed a semi-structured guide covering the use of AI tools, trust in and resistance to platform moderation, the rival-flagging dynamic, and the labor dimensions of the work. Analysis followed reflexive thematic conventions (Braun & Clarke, 2019), with attention to deviant-case analysis, convergence with the computational findings, and practitioner-centered narratives.

The two data-bearing strands were handled under distinct ethics determinations. The interview component constitutes human-subjects research and was conducted under a study protocol reviewed and approved by an independent institutional review board prior to data collection. All participants provided informed consent prior to the interview; participants reviewed their own quotations before publication; politically exposed participants received additional protections appropriate to their risk. Interview audio and transcripts were stored on encrypted, access-restricted media, with identifying details removed from analytic files. The computational corpus, by contrast, comprises public social-media posts and does not constitute human-subjects research; it was anonymized at intake, with usernames and identifying details removed, explicit slurs masked, and post dates randomized within the collection window. Public figures named in political contexts are retained as matters of public record.

5. Results

5.1. Computational Audit (RQ1)

The audit evaluated three classifiers across tiers of language specificity: reporting coverage, performance, and error composition.

5.1.1. Classifier Coverage and Performance

Perspective API rejected all native-mode requests (LANGUAGE_NOT_SUPPORTED_BY_ATTRIBUTE; coverage = .00). On human English translations, its best case, it scored 687 of 838 items (82%), with near-perfect precision but very low recall (precision = 1, recall = .10, $F1$ = .19), catching just 52 of 502 scored Hate items. Both failures reflect the capability absence that is the first condition of infrastructural invisibility.

The Africa-centric tier was a fine-tuned AfroXLMR model (Muhammad et al., 2025; validation macro-*F*1 = .72), which scored every item (coverage = 1.00). On Amharic it was the strongest classifier audited (*F*1 = .75; precision = .91, recall = .64), but on Afan Oromo—for which AfriHate provides no training data—performance collapsed (*F*1 = .29; recall = .17). Because the model is constant across languages and varies only in training-data availability, this within-model drop isolates data scarcity rather than anything intrinsic to Afan Oromo.

Amharic mBERT, evaluated on all 470 Amharic items, achieved *F*1 = .68 (precision = .86, recall = .56), missing 147 of 331 Amharic Hate items (44.4%). Against its reported training-distribution *F*1 of .92, this is a 24-point deployment decrement, consistent with the documented in-domain versus out-of-domain generalization gap (Blodgett et al., 2020).

For Afan Oromo (*n* = 368), no tool provided usable detection: Perspective recovered almost no hate on translations, AfroXLMR caught only 45 of 266 Hate items (*F*1 = .29), and mBERT was inapplicable. The more under-resourced language is least served at every tier, consistent with the resourcing asymmetry in the natural language processing (NLP) literature (Joshi et al., 2020). Table 1 reports these results in full: coverage, precision, recall, and *F*1 for each classifier by language subset.

Table 1. Binary hate-detection performance by classifier and language subset.

Classifier	Language	*n*	Precision	Recall	*F*1	Coverage
Perspective API (native)	All	838	—	—	—	0/838
Perspective API (translation)	All	838	1.00	.10	.19	687/838
Perspective API (translation)	Amharic	470	1.00	.11	.20	409/470
Perspective API (translation)	Afan Oromo	368	1.00	.09	.16	278/368
AfriHate (AfroXLMR, fine-tuned)	All	838	.92	.43	.59	838/838
AfriHate (AfroXLMR, fine-tuned)	Amharic	470	.91	.64	.75	470/470
AfriHate (AfroXLMR, fine-tuned)	Afan Oromo	368	.94	.17	.29	368/368
Amharic mBERT	Amharic	470	.86	.56	.68	470/470
Amharic mBERT	Afan Oromo	368	—	—	—	0/368

Notes: *N* = 838 audit-eligible items (Amharic = 470; Afan Oromo = 368); coverage = proportion of eligible items scored; native = original-language text; translation = human English translations; AfriHate = AfroXLMR fine-tuned on the AfriHate Amharic training split.

5.1.2. Error Composition and Enforcement Patterns

Under-enforcement dominated across all tiers. Perspective missed 450 of 502 scored Hate items (89.6%); AfroXLMR missed 119 of 331 Amharic (36.0%) and 221 of 266 Afan Oromo Hate items (83.1%); and mBERT missed 147 of 331 Amharic (44.4%). Even the strongest configurations let a third to a half of Amharic hate through, while Afan Oromo hate went almost entirely undetected.

Over-enforcement was comparatively rare: precision stayed high (.86–1.00) and perspective produced no false positives. The two local models' modest false positives nonetheless fell disproportionately on counterspeech: of AfriHate's 20 Amharic false positives, 17 (85%) were counterspeech; of mBERT's 29, 20 (69%) were. Though low in absolute terms, this concentration on counterspeech is consistent with prior

findings that classifiers disproportionately flag counterspeech and marginalized voices (Haimson et al., 2021; Oliva et al., 2021; Sap et al., 2019).

5.2. Document Analysis (RQ2)

The document corpus shows the same infrastructural-invisibility pattern at a different register. Meta, X, and TikTok present themselves as governing harmful content through policy, automated detection, and human review. Visible are policy categories, enforcement philosophies, and aggregate metrics; invisible are language-specific infrastructures, the labor sustaining them, and the model-performance signals that would permit external evaluation.

5.2.1. Visible Rules, Less Visible Operations

Platforms are most transparent at normative classification. Meta, X, and TikTok publish detailed taxonomies of prohibited behaviors (Gillespie, 2018; Meta, n.d.-b; TikTok, 2025; X Help Center, n.d.), but these are taxonomic rather than operational: They explain how harmful content is classified while saying little about implementation for Ethiopian-language material. Ananny and Crawford (2018) call this "seeing without knowing"—visibility of outputs while the sociotechnical machinery stays inaccessible to verification. The Ethiopian case sharpens the critique, since moderation here turns on multiple languages, ethnicized references, sarcasm, coded speech, and rapid context shifts during conflict (Pohjonen & Gagliardone, 2017). Platforms invoke a vocabulary of technical competence—proactive detection, machine learning, classifiers, NLP—but rarely explain how these tools are configured or evaluated for Ethiopian languages. The public explanation demonstrates that AI moderation exists without making its adequacy empirically testable.

5.2.2. Crisis-Visible Ethiopia

Ethiopia enters the platform record episodically, not as a stable site of language-specific governance, surfacing most fully at moments of crisis or external scrutiny. The pattern is clearest at Meta, whose Ethiopia statements cluster around the 2021 election, the Tigray conflict, and Oversight Board attention (Meta, 2021; Oversight Board, 2022). The Meareg Amare case—a Tigrayan professor killed after Facebook posts doxxed and targeted him—became the basis for Kenyan litigation; the proceedings and Amnesty's investigation exposed how little Ethiopian-language moderation Meta had, rather than prompting the company to detail it (Amnesty International, 2023). That scrutiny intensified with whistleblower Frances Haugen's 2021 Senate testimony, which described Facebook as "literally fanning ethnic violence" in Ethiopia and reported that its hate-speech classifiers then covered only two of the country's six most-spoken languages (Esberg, 2021; Hao, 2021)—an insider account converging with the gap this study's audit measures directly. The platforms diverge: Meta is comparatively more transparent through Oversight Board review yet opaque about which Ethiopian languages have dedicated classifiers and how they perform under conflict; X offers thinner specificity and a two-year reporting gap between the last Twitter Transparency Report (H2, 2021; this and earlier reports issued under previous ownership have since been removed from X's transparency archive) and the September 2024 X report (X Corp., 2024); TikTok describes intervention typologies beyond removal (TikTok, 2025, 2026) but remains locally nonspecific, lacking an Oversight Board equivalent (Oversight Board, 2024). Flyverbom (2019) calls this a "digital prism": transparency that illuminates some aspects while shadowing others.

5.2.3. Infrastructural Opacity and the Limits of Metrics

The most consequential invisibility lies not in policy but in infrastructure. Public documents rarely reveal classifier coverage across Ethiopian languages, training data, annotation protocols, dialect sensitivity, reviewer expertise, or error patterns under distribution shift. They reference "human review" without disclosing the scale, composition, or linguistic competence of the Ethiopian reviewer workforce, and never acknowledge the volunteer pages, TikTok counterspeech creators, and civic X accounts whose flagging, counterspeech, and fact-checking supply signal platforms metabolize. The metrics platforms do publish (Meta, n.d.-a; TikTok, 2026; X Corp., 2024) render moderation as volume, not effectiveness: they omit Ethiopian-language error rates, English/non-English performance comparisons, and counterspeech misclassification, substituting visibility for accountability (Raji et al., 2020).

5.3. Interview Findings (RQ3)

The interview material shows how practitioners inhabit the moderation environment the audit measures and the document analysis reconstructs. Six themes organize the findings.

5.3.1. Civic Labor as Distributed Moderation Infrastructure

Participants frame their work not as platform use but as ongoing infrastructural maintenance across distributed networks: fact-checking, monitoring, translation, verification, training, and trust repair—a breadth existing platform-labor taxonomies do not capture. Where Crawford and Gillespie (2016) analyze flagging as infrastructure built on uncompensated labor, the present material shows infrastructural labor precedes the flag, in verification and translation work platforms neither commission nor acknowledge. Senait, the organizer of a digital fact-checking campaign, produces public-facing "educational" content, translating platform community guidelines into Amharic and Afan Oromo and walking users through the reporting process—unpaid work that enforces platform rules on the companies' behalf. Dawit, a fact-checker and digital investigator who verifies and counters misleading claims across video, image, and text, describes a single operational frame spanning media monitoring, investigative verification, training, resource curation, and operational security. Participants emphasize that the work is never individual, as Gemechu, a fact-checker at an Ethiopian civil-society organization, put it: "It is never really done alone, even when you look like one person online." Networks of colleagues, journalists, local informants, and dialect consultants contribute to what platforms see as a single account.

5.3.2. AI as Assistant, Not Authority

Participants take a conditional, instrumental stance toward AI, not enthusiastic adoption or categorical rejection. They use AI for triage, translation, drafting, summarization, and research while retaining final interpretive authority, as Ermias, an administrator of a Facebook page that counters misinformation, explained: "I use them mostly as support tools, not decision-makers. They are useful for speed, but I do not trust them with final meaning." Participants edit AI outputs substantially: The generic register of AI drafts would undermine counterspeech credibility in a community where voice and stance are politically legible. Dawit frames investigative verification as a professional discipline; tool use is embedded in method.

5.3.3. Translation Failure and the Politics of Context

Participants describe AI tools as useful for triage but unreliable on the semantic phenomena that give Ethiopian political discourse its force. Tools fail on sarcasm, coded language, political slogans, religious phrases, and emotionally charged language; a small mistranslation can change whether a post looks hateful, neutral, or defensive. One participant showed how Facebook's default translation of an Afan Oromo post rendered it into near-incoherent English—scrambling who acted, against whom, and with what stance—leaving a reader unable to tell whether the original post attacked, defended, or merely reported. This is not an Afan Oromo peculiarity but the visible edge of a resourcing gradient that degrades machine translation and classifiers alike, most sharply where less training data exists. The mechanism is sociolinguistic: In Amharic and Afan Oromo, hateful meaning is frequently carried pragmatically—through implicature, code-switching, proverbial and religious framing, and shifting in-group references—rather than through fixed lexical markers, so classifiers tuned to surface vocabulary miss precisely the context that determines whether an utterance attacks, defends, or merely references a group (Pohjonen & Gagliardone, 2017). The interviews supply the mechanism behind the audit's under-detection and counterspeech false positives: AI extracts surface-lexical signal, while what matters is pragmatic and contextual. Earlier NLP work on Ethiopian languages (Ayele et al., 2023; Demilie & Salau, 2022) identifies classifier failure on code-mixed, politically loaded, and religiously framed content as a persistent obstacle to deployable moderation.

5.3.4. Moderation as Misrecognition and the Politics of Adaptation

Participants describe platform moderation as misrecognition: The system detects lexical signals without grasping the communicative purpose of the post. As Henok, a counter-speaker who runs a multimedia TikTok fact-checking page, put it: "The system seems to detect the sensitive words but not the purpose of the post. You appeal and usually get a generic answer, if any answer comes at all." This corroborates the audit: Amharic mBERT flags counterspeech at a rate approaching its rate on Hate. The downstream consequence is adaptation: Participants reshape their speech around anticipated moderation—paraphrasing rather than quoting hate, avoiding trigger words, and developing platform-specific tactics. Meron, a volunteer fact-checker active on Facebook during the Tigray war, observed: "Sometimes I avoid a phrase even when I need to discuss it, because I know the system may punish the post before it understands the context." Platform governance operates not only through takedowns but upstream, through anticipation shaping what practitioners say, quote, and frame.

5.3.5. Rival Flagging and the Weaponization of Platform Rules

Participants describe reporting systems as contested political terrain, not neutral civic goods, as Getachew, a staff member at a local civil-society organization, noted: "Some reporting is about genuine harm, but some is clearly strategic. People use platform rules as weapons in political conflict." Flagging systems produce signal regardless of provenance, and adversarial groups exploit this indifference. Participants describe a characteristic signature: A post doing well suddenly attracts multiple reports, and platforms do not differentiate adversarial from legitimate flags. The civic labor platforms benefit from is politically heterogeneous, and that heterogeneity is invisible at the aggregate level where decisions are made. Beyond the infrastructural extraction the frame names, rival flagging adds a political dimension: The weaponization of reporting infrastructures is itself a value that platforms absorb without responsibility for producing.

5.3.6. The Costs of Civic Labor

The work imposes affective, political, and economic costs. As Senait described it: "It costs time, sleep, attention, and sometimes peace of mind. You spend so much time around hostility, grief, manipulation, and urgency that it changes your nervous system." Unlike commercial moderators (Roberts, 2019), these fact-checkers and counter-speakers are not hired, vetted, or compensated, and absorb exposure to hate and political risk without mental-health services, rotation schedules, or worker protections. Dawit's emphasis on operational security shows the labor is protective as well as epistemic—self-defense where investigators are themselves vulnerable. Mulugeta, a diaspora-based fact-checker and journalist, described the sustainability problem: "A lot of this work is sustained through a patchwork: other jobs, grants, small institutional support, volunteer effort. It is not stable enough for the level of social value it produces." Azoulay (2024) documents the organizational version: Short-term donor funding pushes Ethiopian fact-checking organizations toward training rather than verification.

That the work is so precariously funded—most participants are volunteers, not paid staff—makes the question of why they persist especially pointed. They answered in terms of contribution rather than reward. As Gemechu put it: "I keep working in this field because I see my work help, improve, and contribute my part." For Ermias, the motivation was protective: "Despite the risk, I keep going, because it is better to do something than to let false information spoil community relationships and destroy what is factual and true."

6. Discussion

6.1. The Infrastructure of Invisibility

As Table 2 summarizes, the three strands converge on a single infrastructural pattern. Read through the infrastructural-invisibility frame (Plantin et al., 2018; Star & Ruhleder, 1996), the institutional and technical findings cohere. The computational audit documents Perspective API's failure to read Ethiopian-language content and Amharic mBERT's deployment-time performance gap; the document analysis shows platforms making global claims about automation while staying silent on Ethiopia-specific operations. For Ethiopian-language content, the moderation apparatus operates invisibly precisely because it operates poorly, and the mechanisms that would render it visible—per-language transparency reporting, granular regulatory obligations, independent accountability bodies—are themselves absent or receding (Gorwa et al., 2020; Katzenbach, 2021). The consequences are concrete: in April 2026, harassment targeting an Ethiopian creator who posted in Afan Oromo—content violating TikTok's own policy—went unflagged by both companies' automated systems, as did the posts she made before she died by suicide; those posts remained on Facebook after TikTok removed them (Chala, 2026). X's post-2022 retreat from periodic transparency reporting is not merely a research inconvenience but an instance of infrastructural re-occlusion: the deliberate removal of visibility that once supported external audit.

Audit alone cannot close this gap: It depends on a floor of platform transparency that, as the X case shows, can be withdrawn. The more durable response is infrastructural rather than methodological—mandated disclosure and standing accountability bodies, not better tools.

Table 2. Synthesis of findings across the three strands.

Strand	Key finding	Theoretical reading	Convergence
Computational audit	Generic tools cannot read either language natively and recover ~10% of hate on translation; local tools reach *F*1 .75 (Amharic) but collapse on Afan Oromo; all under-detect.	Capability absence renders moderation infrastructurally invisible.	Quantifies the breakdown practitioners describe.
Document analysis	Platforms describe Ethiopian-language moderation without making it legible for external evaluation; disclosure is partial and revocable.	Self-description reproduces invisibility; visibility is selective.	Shows the same silence at the institutional register.
Interviews	A volunteer ecology performs the detection, counterspeech, and verification platforms do not, at affective, political, and economic cost.	Civic labor fills the infrastructural void, uncompensated and unrecognized.	Supplies the human mechanism behind the audit gap.

6.2. The Civic Labor That Fills the Infrastructural Void

Where infrastructural invisibility names the absence platform AI produces, civic labor names what fills it (Gibson, 2023; Gray & Suri, 2019; Roberts, 2019). The interviews document a volunteer ecology of hate-speech flaggers, counter-speakers, and fact-checkers who perform much of the moderation platform AI does not. This does not contradict the audit; it completes its implication. If platform AI does not flag hate speech in Amharic and Afan Oromo, someone does—a network structurally under-supported by the platforms their work serves. A fact-checker in Azoulay's (2024, p. 14) study describe the companies as "a total disappointment" that "are failing countries like Ethiopia." The audit supplies the empirical analogue: what they describe qualitatively as failure, it measures as zero native coverage and persistent under-detection even by purpose-built Amharic classifiers, with no usable tool for Afan Oromo.

The two refinements the frame anticipated are borne out. The ecology's political heterogeneity is not incidental: Rival groups flag each other's content in contestation, producing an economy of reports in which what counts as hate is itself an object of struggle. Platforms ingest these flags as undifferentiated signals, benefiting from the ecology's load-bearing capacity without accounting for the contestation that generates it. This labor is unevenly costly: Volunteers absorb the psychological toll of sustained exposure to hate, the political risk of identifiable counterspeech, and the economic cost of uncompensated work, compounded by the automated suppression of counterspeech itself—counter-speakers face not only the hate they answer but the penalization of their answers.

These findings make two contributions. Empirically, they document a low-resource, conflict-affected case in which moderation labor theorized largely for commercial and Western contexts (Gibson, 2023; Gillespie, 2018; Pinchevski, 2023) is performed by an unpaid, politically exposed civic ecology. Theoretically, they show that infrastructural invisibility and civic labor are not separate problems but two faces of one arrangement: Platforms render their Ethiopian-language moderation invisible precisely by relying on a civic labor they neither compensate nor name.

6.3. Limitations

Several limitations bound these findings. First, data availability constrains the study: A more systematic evaluation would draw on platform-internal performance data and Ethiopia-specific transparency disclosures, but these do not exist in usable form—an absence itself part of the phenomenon documented here. The audit therefore rests on an external, purposively sampled corpus rather than exhaustive data. Because the corpus lacks per-row platform-origin or content-type documentation, the audit is language-level rather than platform- or content-type-level, and reports performance for the 2020–2025 window; explicit slurs were masked as an ethics safeguard, likely depressing recall. Second, the findings are specific to the Ethiopian case—its languages, platforms, and conflict dynamics—and should extend to other low-resource or conflict-affected settings only with caution: The mechanisms (capability absence, civic labor, selective disclosure) are likely to recur, but their configuration is context-bound. Third, the interview sample cannot claim exhaustive coverage of a small, security-conscious ecology.

7. Conclusion

AI-assisted moderation in Ethiopian online spaces shows a structural mismatch among what platforms describe, what publicly deployable classifiers can do, and what Ethiopian fact-checkers and volunteer counter-speakers encounter. Through the infrastructural-invisibility frame, it reveals a moderation apparatus whose operation on Ethiopian-language content is largely hidden from the public accountability mechanisms that would render it visible. Through the civic-labor frame, it reveals a volunteer civic ecology whose work fills the infrastructural void at political, affective, and economic cost, and from which platforms extract operational value without acknowledgment.

First, platforms could better support the volunteer ecology, with appeals mechanisms independent of the classifiers that produced the flag, per-country and per-language enforcement disclosure, and access mechanisms for fact-checkers and civic-page administrators, rather than pursuing marginal automation improvements whose coverage gap is structural.

Second, regulatory frameworks mandating minimum disclosure standards, irrespective of ownership changes, are a prerequisite for sustained independent audit. The post-2022 X case illustrates what happens when such standards are absent: The visibility on which external accountability depends can be withdrawn unilaterally.

Civic resilience in Ethiopian online environments rests not on the technical infrastructure platforms claim to provide but on the uneven, politically contested, and structurally under-supported labor of a specific civic ecology. The conditions under which AI can support rather than undermine it remain largely unmet; the work is carried by humans whose labor the moderation infrastructure simultaneously depends on and refuses to see.

Acknowledgments

The author thanks the interview participants who generously shared their time and expertise.

Conflict of Interests

The author declares no conflict of interests.

Data Availability

The materials supporting the computational audit are available from the author on reasonable request. Interview data are not publicly available in order to protect the confidentiality of participants.

LLMs Disclosure

The author used Gemini (Google) solely to assist with language editing and formatting. The tool was not used to generate, analyze, or interpret data, or to produce the study's arguments, findings, or conclusions. The author reviewed and verified all edited text and takes full responsibility for the content of the article.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited). It documents the audit corpus and annotation scheme, the classifier audit protocol and model settings, and the inventory of platform documents analyzed. A separate file provides the underlying data. The codebook, datasheet, and ethics documentation for the computational audit are also available at https://github.com/Endalk-Chala/ethiopia-ai-moderation-audit

References

Ababu, T. M., & Woldeyohannis, M. M. (2022). Afaan Oromo hate speech detection and classification on social media. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth Language Resources and Evaluation Conference* (pp. 6612–6619). European Language Resources Association. https://aclanthology.org/2022.lrec-1.712

Amnesty International. (2023). *"A death sentence for my father": Meta's contribution to human rights abuses in northern Ethiopia* (Index No. AFR 25/7292/2023).https://www.amnesty.org/en/documents/afr25/7292/2023/en/

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Ayalew, Y. E. (2025). Linguistic blind spots in platform content moderation in Ethiopia: The case of Facebook. *Modern Languages Open*, (1), Article 11. https://doi.org/10.3828/mlo.v0i0.468

Ayele, A. A., Yimam, S. M., Belay, T. D., Asfaw, T., & Biemann, C. (2023). Exploring Amharic hate speech data collection and classification approaches. In G. Angelova, M. Kunilovskaya, & R. Mitkov (Eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 49–59). INCOMA. https://doi.org/10.26615/978-954-452-092-2_006

Azoulay, L. (2024). Truth in the crossfire: The case of Ethiopia and fact-checking in authoritarian contexts. *Media and Communication*, *12*, Article 8785. https://doi.org/10.17645/mac.8785

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 74. https://doi.org/10.1145/3449148

Bélair-Gagnon, V., Larsen, R., Graves, L., & Westlund, O. (2023). Knowledge work in platform fact-checking partnerships. *International Journal of Communication*, *17*, 1169–1189. https://ijoc.org/index.php/ijoc/article/view/19851

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention* (Working Paper). United States Holocaust Memorial Museum. https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky (Ed.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Borenstein, N., Warren, G., Elliott, D., & Augenstein, I. (2025). Can community notes replace professional fact-checkers? In R. Navigli (Ed.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 535–552). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-short.42

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Buerger, C. (2021). #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, *7*(4). https://doi.org/10.1177/20563051211063843

Cazzamatta, R. (2026). From moderation to chaos: Meta's fact-checking and the battle over truth and free speech. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251413687

Chala, E. (2026, May 5). Ethiopian TikTok grieves a popular creator. Are TikTok and Meta complicit in her death? *Global Voices*. https://globalvoices.org/2026/05/05/ethiopian-tiktok-grieves-a-popular-creator-some-wonder-whether-tiktok-and-meta-are-complicit-in-her-death

Corsi, G., Seger, E., & hÉigeartaigh, S. Ó. (2024). Crowdsourcing the mitigation of disinformation and misinformation: The case of spontaneous community-based moderation on Reddit. *Online Social Networks and Media*, *43–44*, Article 100291. https://doi.org/10.1016/j.osnem.2024.100291

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Demilie, W. B., & Salau, A. O. (2022). Detection of fake news and hate speech for Ethiopian languages: A systematic review of the approaches. *Journal of Big Data*, *9*, Article 66. https://doi.org/10.1186/s40537-022-00619-x

Dierickx, L., Lindén, C.-G., & Opdahl, A. L. (2023). Automated fact-checking to support professional practices: Systematic literature review and meta-analysis. *International Journal of Communication*, *17*, 5170–5190. https://ijoc.org/index.php/ijoc/article/view/21071

Endsley, M., & Kaber, D. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. https://doi.org/10.1080/001401399185595

Esberg, J. (2021). *What the Facebook whistleblower reveals about social media and conflict*. International Crisis Group. https://www.crisisgroup.org/united-states/united-states-internal/what-facebook-whistleblower-reveals-about-social-media-and-conflict

Flick, U. (2018). *Doing triangulation and mixed methods* (2nd ed.). Sage.

Flyverbom, M. (2019). *The digital prism: Transparency and managed visibilities in a datafied world*. Cambridge University Press.

Gibson, A. D. (2023). What teams do: Exploring volunteer content moderation team labor on Facebook. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186109

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gonçalves, A., Torre, L., Oliveira, F., & Jerónimo, P. (2024). AI and automation's role in Iberian fact-checking agencies. *Profesional de la Información*, *33*(2), Article e330212. https://doi.org/10.3145/epi.2024.0212

Gorwa, R., Binns, R., & Katzenbach, C. (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

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.

Graves, L. (2018). Boundaries not drawn. *Journalism Studies*, *19*(5), 613–631. https://doi.org/10.1080/1461670X.2016.1196602

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and Black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), Article 466. https://doi.org/10.1145/3479610

Hao, K. (2021). *The Facebook whistleblower says its algorithms are dangerous. Here's why*. MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms

Hitomi, K. (1994). Automation—Its concept and a short history. *Technovation*, *14*(2), 121–128. https://doi.org/10.1016/0166-4972(94)90101-5

Jacobsen, B. N. (2025). Deepfakes and the promise of algorithmic detectability. *European Journal of Cultural Studies*, *28*(2), 419–435. https://doi.org/10.1177/13675494241240028

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky (Ed.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Katzenbach, C. (2021). "AI will fix this"—The technical, discursive, and political turn to AI in governing communication. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211046182

Kemp, S. (2025). *Digital 2025: Ethiopia*. we are social; Meltwater. https://datareportal.com/reports/digital-2025-ethiopia

Koliska, M., & Roberts, J. (2025). Epistemology of fact checking: An examination of practices and beliefs of fact checkers around the world. *Digital Journalism*, *13*(3), 500–520. https://doi.org/10.1080/21670811.2024.2361264

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A new generation of Perspective API: Efficient multilingual character-level transformers. In A. Zhang & H. Rangwala (Eds.), *KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3197–3207). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539147

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Masud, S., Singh, S., Hangya, V., Fraser, A., & Chakraborty, T. (2024). Hate personified: Investigating the role of LLMs in content moderation. In T. Solorio (Ed.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 15847–15863). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.886

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2). https://doi.org/10.1177/2056305119836778

Meisner, C. (2023). The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy. *Policy & Internet*, *15*(4), 466–477. https://doi.org/10.1002/poi3.359

Meta. (n.d.-a). *Community standards enforcement report*. https://transparency.meta.com/reports/community-standards-enforcement

Meta. (n.d.-b). *Hateful conduct: 29 Feb 2024*. https://transparency.meta.com/policies/community-standards/hateful-conduct

Meta. (2021). *How Facebook is preparing for Ethiopia's 2021 general election*. https://about.fb.com/news/2021/06/how-facebook-is-preparing-for-ethiopias-2021-general-election

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Adelani, D. I., Ahmad, I. S., Aliyu, S. M., Röttger, P., Oppong, A., Bukula, A., Chukwuneke, C. I., Jibril, E. C., Ismail, E. A., Alemneh, E., Gebremichael, H. T., Aliyu, L. J., Beloucif, M., Hourrane, O., Mabuya, R., Osei, S., . . . Ousidhoum, N. (2025). AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In C. Cherry (Ed.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1854–1871). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.92

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In Y. Matsumoto (Ed.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9759–9777). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.653

Mun, J., Buerger, C., Liang, J. T., Garland, J., & Sap, M. (2024). Counterspeakers' perspectives: Unveiling barriers and AI needs in the fight against online hate. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 1085). ACM. https://doi.org/10.1145/3613904.3642025

Noy, C. (2008). Sampling knowledge: The hermeneutics of snowball sampling in qualitative research. *International Journal of Social Research Methodology*, *11*(4), 327–344. https://doi.org/10.1080/13645570701401305

Oliva, T. D., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Oversight Board. (2022). *Tigray communication affairs bureau*. https://www.oversightboard.com/decision/FB-E1154YLY

Oversight Board. (2024). *Content moderation in a new era for AI and automation*. https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage.

Pinchevski, A. (2023). Social media's canaries: Content moderators between digital labor and mediated trauma. *Media, Culture & Society*, *45*(1), 212–221. https://doi.org/10.1177/01634437221122226

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Pohjonen, M., & Gagliardone, I. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 1173–1191. https://ijoc.org/index.php/ijoc/article/view/5843

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal

algorithmic auditing. In M. Hildebrandt & C. Castillo (Eds.), *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014, May 22). *Auditing algorithms: Research methods for detecting discrimination on internet platforms* [Paper presentation]. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, Seattle, WA, United States. https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In L. Màrquez (Ed.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistic. https://doi.org/10.18653/v1/P19-1163

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, *21*(7), 1417–1443. https://doi.org/10.1177/1461444818821316

Sponholz, L. (2021). Hate speech and deliberation: Overcoming the "words-that-wound" trap. In M. Pérez-Escolar & J. M. Noguera-Vivo (Eds.), *Hate speech and polarization in participatory society* (pp. 49–64). Routledge. https://doi.org/10.4324/9781003109891-5

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In Y. Kitamura & A. Quigley (Eds.), *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 341). ACM. https://doi.org/10.1145/3411764.3445092

TikTok. (2025). *Community guidelines*. https://www.tiktok.com/community-guidelines/en

TikTok. (2026). *Community guidelines enforcement report*. https://www.tiktok.com/safety/en/transparency/cg-report

Vinhas, O., & Bastos, M. (2025). The WEIRD governance of fact-checking and the politics of content moderation. New Media & Society, *27*(5), 2768–2787. https://doi.org/10.1177/14614448231213942

X Corp. (2024). *Global transparency report: H1 2024*. https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf

X Help Center. (n.d.). *Hateful conduct*. https://help.x.com/en/rules-and-policies/hateful-conduct-policy

About the Author

Endalkachew H. Chala is an independent researcher and community fellow at the Center for an Informed Public, University of Washington. He holds a PhD from the University of Oregon and has taught communication studies in Ethiopia and at Hamline University. His research examines content moderation and multilingual online discourse.