



COMMENTS TO META OVERSIGHT BOARD

**How Meta Should Respect
Political Expression in Kenya**
JUNE, 2025

Comments to Meta Oversight Board

How Meta Should Respect Political Expression in Kenya

Authors: Samrudh, Prakhar Singh, Syed Kirdar Husain, Aadit Seth, Amishi Jain, Rajyavardhan

Research Consultant: Dr. Ivneet Walia, Associate Professor of Law and
Officiating Registrar, RGNUL



CENTRE FOR ADVANCED STUDIES IN CYBER LAW AND ARTIFICIAL INTELLIGENCE [CASCA] is a research-driven centre at RGNUL dedicated to advancing scholarly research and discourse in the field of Technology Law and Regulation. As a research centre of a leading institution in India, we are committed to promoting interdisciplinary research, fostering collaboration, and driving innovation in the fields of cyber law, artificial intelligence, and other allied areas.

For more information

Visit cascargnul.com

Disclaimer

The facts and information in this report may be reproduced only after giving due attribution to CASCA.

Comments to Meta's Oversight Board's Consideration of how Meta Should Respect Political Expression in Kenya

This comment addresses the issue on - how reliance on automated enforcement of Meta's Hateful Conduct policy impacts human rights, including freedom of expression, especially in countries with a recent history of intercommunal violence.

Introduction

In theory, the internet is imagined as the great "marketplace of ideas",¹ which in simpler terms, is to say a democratic space where truth outcompetes falsehood in a free exchange of perspectives. But that metaphor, foundational as it may be to free speech theory, is far too idealistic² when one exports it to the digital age. Today, virality trumps veracity. The architecture of online platforms privileges not thoughtful dialogue but performative outrage and shareability. In politically sensitive contexts, much like Kenya³ or any other country in the world, where intercommunal tensions run particularly high, this reconfiguration of the marketplace becomes deeply consequential.

Words, then, don't just carry meanings - but deep socio-political undertones that carry histories, identities, and so on. The vocabulary of ethnic politics is not neutral. In such spaces, even a single term (depending on its context and valence) can inflame or polarise. This is why the regulation of language online becomes both urgent and fraught.

Meta's reliance on automated moderation for its hate speech policies has repeatedly failed in high-risk environments, contributing to serious human rights violations. These failures are most acute where automated systems do not interpret the linguistic differences in under resourced languages and localised contexts.⁴ As a result, there is a difference in the scalability and speed of the enforcement, which becomes disconnected from the actual realities, especially in regions vulnerable to inter communal tensions, as transpired in Kenya.

Instances of Meta's inaction to curb hate speech

Meta's Facebook in the Rohingya genocide in Myanmar had its algorithms amplify anti-Rohingya sentiments via hate speech and misinformation. A hateful viral digital ecosystem flourished which led to unchecked violence as a result. Facebook's automated systems failed to detect inflammatory content in Burmese⁵, despite the existence of Community Standards.¹ This failure was due to linguistic and regional issues, as Meta's generic policy tools did not account for local speech contexts.¹ Independent assessments, including by the UN Fact-Finding Mission, found that Facebook's shortcomings significantly contributed to radicalization and mass atrocities by failing to curb hate speech⁶.

¹ David Schultz, 'Marketplace of Ideas' (Free Speech Center at Middle Tennessee State University, 1 January 2009, updated 9 July 2024) <https://firstamendment.mtsu.edu/article/marketplace-of-ideas/> accessed 15 June 2025.

² Tom Toles, 'Marketplace of Ideals' (Washington Post, 16 April 2024) <https://www.washingtonpost.com/news/opinions/wp/2014/04/16/marketplace-of-ideals/> accessed 15 June 2025.

³ Emma Elfversson, 'Patterns and Drivers of Communal Conflict in Kenya' in Steven Ratuva (ed), *The Palgrave Handbook of Ethnicity* (Palgrave Macmillan 2019) https://doi.org/10.1007/978-981-13-2898-5_50 accessed 15 June 2025.

⁴ Pat de Brún, 'Meta's New Content Policies Risk Fueling More Mass Violence and Genocide' (Amnesty International, 17 February 2025) <https://www.amnesty.org/en/latest/news/2025/02/meta-new-policy-changes/> accessed 15 June 2025.

⁵ Amnesty International, 'Myanmar: Facebook's systems promoted violence against Rohingya – Meta owes reparations, new report' (2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> accessed 15 June 2025.

⁶ Global Witness, 'Facebook approves adverts containing hate speech inciting violence and genocide against the Rohingya' (1 October 2021) <https://globalwitness.org/en/campaigns/digital-threats/facebook-approves-adverts-containing-hate-speech-inciting-violence-and-genocide-against-the-rohingya/> accessed 15 June 2025.

Firstly, in Ethiopia, Meta through Facebook approved paid advertisements explicitly calling for violence against ethnic groups, which again, was a failure of a policy.⁷ Automated moderation failed to recognize dehumanizing rhetoric in Amharic and other local languages, despite Meta's hate speech policy prohibiting Tier 1 attacks which define violent and dehumanizing acts towards people based on ethnicity or nationality. Secondly, in India, where communal tensions are politically sensitive, Facebook and Instagram algorithms promoted divisive content and failed to act upon flagged hate speech against minorities. Automated systems were ill-equipped to process linguistic nuance in regional languages.⁸ In 2020, Delhi riots, hate speech and incitement to violence were widely circulated on Facebook. Posts called for violence, glorified rioters, and used dehumanizing language. These posts, with specific violence calls, stayed up.

Internal whistleblower documents later revealed that automated moderation tools failed to flag inflammatory content in Hindi, and content reviewers lacked sufficient language and cultural knowledge.⁹

These failures are not merely algorithmic but institutional. Meta's content moderation framework applies a one-size-fits-all approach to curb global hate speech, which in many ways, disregards the linguistic, cultural, and political specificities of high-risk regions. This ignorance in automated detection policies by Meta, then, becomes an enabler of human rights violations - revealing the disconnect between ground reality and platform policy.

Shortcomings of Meta's Automated Moderation

Meta has a responsibility to respect international human rights, as the UN Guiding Principles on Business and Human Rights¹⁰ calls for businesses to uphold human rights. Under international human rights law, restrictions to rights such as freedom of expression¹¹ and free of assembly and association¹² can only be justified if there is a legal basis, legitimate aim and proportional action.

While automated systems can be helpful in moderating content at a scale, they have significant limitations. European Centre for Non-profit Law has cautioned that these systems usually accelerate existing challenges related to content moderation, i.e. lack of transparency and understanding local contexts.¹³ Algorithmic systems based on keyword detection and language models are not completely capable of understanding nuances of statements made on platforms.

Tech companies' recent mass layoffs have contributed to its increasing incapacities to moderate content in non-English languages. For instance, a third-party contractor in Nairobi that was responsible for content moderation in Ethiopia and Kenya was let go by

⁷ Global Witness, 'Now is the Time to Kill: Facebook Continues to Approve Hate Speech Inciting Violence and Genocide During Civil War in Ethiopia' (2023) <https://globalwitness.org/en/campaigns/digital-threats/now-is-the-time-to-kill-facebook-continues-to-approve-hate-speech-inciting-violence-and-genocide-during-civil-war-in-ethiopia/> accessed 15 June 2025.

⁸ The News Minute, 'Inflammatory content on FB was up 300% before Delhi riots, says internal report' (2020) <https://www.thenewsminute.com/atom/inflammatory-content-fb-was-300-delhi-riots-says-internal-report-156878> accessed 15 June 2025

⁹ TechScience, 'Linguistic Inequity in Facebook Content Moderation' (25 February 2025) <https://techscience.org/a/2025022501/> accessed 15 June 2025

¹⁰ Office of the United Nations High Commissioner for Human Rights, 'Guiding Principles on Business and Human Rights: Implementing the United Nations¹ "Protect, Respect and Remedy" Framework²' (2011) https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinessshr_en.pdf accessed 15 June 2025.

¹¹ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) art 19.

¹² International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) art 21.

¹³ European Center for Not-for-Profit Law, 'TECHNOLOGY AND COUNTER-TERRORISM' (2022) https://ecnl.org/sites/default/files/2023-03/TECHNOLOGY%20AND%20COUNTER-TERRORISM_NOV%202022.pdf accessed 15 June 2025.

Meta amid reduced revenues.¹⁴ In the United States, Facebook claimed that because of a lack of content moderators, it is relying more on automated systems.¹⁵

Evidence has shown that tech companies have implemented policies¹⁶ in moderating non-English language content that has impeded individuals' right of free expression and access to information (in their native languages). According to a FB whistleblower¹⁷ Facebook allocates 87% of its misinformation countermeasure budget on English content, despite only 9% of its users being primary English speakers. In 2023, the Oversight Board¹⁸ reported that 38% of content moderation cases originated from the United States and Canada, while 26% came from Europe, which accounted for a total of 64% of all cases. In contrast, only 5% of cases were reported from Central and South Asia, despite India having the largest number of users on Facebook and Instagram.

There is a lack of intent in addressing content moderation in the Global South, as Meta continues to only prioritise on English content moderation. Facebook entered markets in the Global South in the early 2000s and it is still not adopted to the cultural contexts of those specific regions. Meta had employed content moderators across 70 languages (including 20 Indian languages) on a global level, however, this is inadequate as India itself has over 100 languages.¹⁹ Furthermore, Meta's employees and contractors lack the cultural competency needed to address diverse issues related to caste, religious, gender and queer minorities.²⁰ Meta's moderation moved further away from understanding these nuanced linguist and cultural contexts by adopting ineffective and ill-trained automated systems.

Automated systems use techniques such as keyword filters, spam detection tools and hash-matching algorithms. In addition to this, large language models (LLMs) trained on data fed by human moderators have also been deployed to detect violating content. However, the use of LLMs is restricted due to lower volume and poorer quality of text data available for many non-English languages. This has created a "resourcedness gap" that makes it difficult to train LLMs.²¹ The Centre for Democracy and Technology²² identified that tech companies attempted to build multilingual LLMs to support their scaling of automated moderation, however the training data used for low-resource languages was translated or even mistranslated or scraped from low-

¹⁴ TechCrunch, 'Meta's main content moderation partner in Africa shuts down operations' (TechCrunch, 10 January 2023) https://techcrunch.com/2023/01/10/metas-main-content-moderation-partner-in-africa-shuts-down-operations/?gucounter=1&guc_referrer=aHR0cHM6Ly93d3cuZ29vZ2xILmNvbS8&guc_referrer_sig=AQAAAKJmwc3Z-xusQutQ1P1fu8_ecEkBO11QqDzxvn87reKqMJSE4mmofk_RpG57DPGlwdWbF0t5p-4O83h9ur9-f6gaaPlnMZ89fviSSc-nw9A_x4bAft_HtFgcas4zhbZotZKit5uhbNbTSUXwseQTowIDVeu7A0O97Ihk_oTyoPtG accessed 15 June 2025.

¹⁵ Axios, 'Facebook content moderation is a 'black hole' for human review' (Axios, 22 June 2023) <https://www.axios.com/2023/06/22/facebook-content-moderation-black-hole-human-review> accessed 15 June 2025.

¹⁶ Mahsa Alimardani and Mona Elswah, 'Digital Orientalism: #SaveSheikhJarrah and Arabic Content Moderation' (Project on Middle East Political Science, 2022) <https://pomeps.org/digital-orientalism-savesheikhjarrah-and-arabic-content-moderation> accessed 15 June 2025. See also, Mona Elswah and Philip N Howard, 'The Challenges of Monitoring Social Media in the Arab World: The Case of the 2019 Tunisian Elections' (COMPROP Data Memo 2020.1, 23 March 2020, Project on Computational Propaganda, Oxford Internet Institute) <https://demtech.oii.ox.ac.uk/research/posts/the-challenges-of-monitoring-social-media-in-the-arab-world-the-case-of-the-2019-tunisian-elections/#continue> accessed 15 June 2025.

¹⁷The Guardian, 'Facebook revelations: from misinformation to mental health – the key takeaways' (25 October 2021) <https://www.theguardian.com/technology/2021/oct/25/facebook-revelations-from-misinformation-to-mental-health> accessed 15 June 2025.

¹⁸Oversight Board, *Oversight Board 2023 Annual Report* (2024) <https://www.oversightboard.com/wp-content/uploads/2024/06/Oversight-Board-2023-Annual-Report.pdf> accessed 15 June 2025

¹⁹Deccan Herald, 'We are doing a lot to protect vulnerable users, says Meta' (4 November 2023) <https://www.deccanherald.com/technology/we-are-doing-a-lot-to-protect-vulnerable-users-says-meta-2764765> accessed 15 June 2025.

²⁰ Soorya Balendra, 'Meta's AI Moderation and Free Speech: Ongoing Challenges in the Global South' (2025) 1 Cambridge Forum on AI: Law and Governance e21 <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/2DB952F896DB5744A43CD3E6C1A6DCB4/S3033373325000055a.pdf/metas-ai-moderation-and-free-speech-ongoing-challenges-in-the-global-south.pdf> accessed 15 June 2025

²¹ Center for Democracy & Technology, *Investigating Content Moderation Systems in the Global South* (30 July 2024) <https://cdt.org/insights/investigating-content-moderation-systems-in-the-global-south/> accessed 15 June 2025

²² Gabriel Nicholas and Aliya Bhatia, *Lost in Translation: Large Language Models in Non-English Content Analysis* (Center for Democracy & Technology 2023) <https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/> accessed 15 June 2025

quality sources. This further signifies the inability of tech companies like Meta to moderate content in the Global South, even though its users are predominantly from these regions. The Draft Effective Guidelines on Hate Speech²³ published by the UN Special Rapporteur on Minority Issues urges social media platforms to offer protection to minorities against hate speech at least to the extent required under international human rights standards. The Guidelines gives an authoritative working definition of online hate speech as content that is “discriminatory, hostile, or violent towards those community members with protected characteristics on the basis of any identity factor.”²⁴

The Guidelines encourage social media platforms to clearly define broad, subjective or ambiguous terms in their policies. Furthermore, it suggests that the content moderation process should follow a “necessity and proportionality” analysis. Any moderation should be the least restricted means available for protecting the rights of the targeted person or group and there should be different enforcement options based on severity. While it is not possible to apply this test to every single post on these platforms, Meta should aim to design its policies and enforcement mechanism that only limit free speech when another person’s rights are at stake and should be done in the least restrictive way possible. The move to automation and the lack of algorithms well trained in non-English languages has inadvertently resulted in the suspension of user accounts/content,²⁵ the supercharging of hate speech²⁶ and the proliferation of misleading content²⁷ in the Global South. This has not only contributed to the suppression of the right to free speech and access to information but also has actively contributed to violence and hate crimes during times of intercommunal conflicts²⁸.

Possible Interventions for Meta

Automated moderation and translation technologies are falling short due to the scarcity of high-quality training data and resources like benchmarks and evaluation metrics to test these systems. While Meta may not always have the resources to evaluate its systems, there exists a number of researchers who have built open source and culturally specific benchmarks. For example, Tattle’s Uli²⁹ is an open-source crowdsourced lexicon of slurs against gender, caste and other marginalized identities across Tamil, Hindi and Indian English. Meta should effectively engage with these Natural Language Processing experts who are building such resources across the world to help identify, vet and make use of available resources to improve moderation in languages across the globe. Furthermore, algorithms deployed must be routinely updated in light of changes in language, as languages are dynamic and change with societal trends. Changes should only be made after consultation with civil societies, following which such changes must be made publicly available for the users.

²³ United Nations Special Rapporteur on Minority Issues (Fernand de Varennes), *Draft ‘Effective Guidelines on Hate Speech, Social Media and Minorities’* (June 2022) <https://www.ohchr.org/sites/default/files/2022-06/Draft-Effective-Guidelines-Hate-Speech-SR-Minorities.pdf> accessed 15 June 2025.

²⁴ United Nations Special Rapporteur on Minority Issues, *Draft ‘Effective Guidelines on Hate Speech, Social Media and Minorities’* (June 2022) <https://www.ohchr.org/sites/default/files/2022-06/Draft-Effective-Guidelines-Hate-Speech-SR-Minorities.pdf> accessed 15 June 2025.

²⁵ Human Rights Watch, *Meta’s Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook* (21 December 2023) <https://www.hrw.org/report/2023/12/21/metass-broken-promises/systemic-censorship-palestine-content-instagram-and> accessed 15 June 2025.

²⁶ The Conversation, ‘How TikTok became a breeding ground for hate speech in the latest Malaysia general election’ (2023) <https://theconversation.com/how-tiktok-became-a-breeding-ground-for-hate-speech-in-the-latest-malaysia-general-election-200542> accessed 15 June 2025

²⁷ France 24, ‘Zimbabwe election disinformation spreads on WhatsApp’ (3 August 2023) <https://www.france24.com/en/live-news/20230803-zimbabwe-election-disinformation-spreads-on-whatsapp-1> accessed 15 June 2025

²⁸ Amnesty International, ‘Meta’s new content policies risk fueling violence and genocide’ (19 February 2025) <https://www.amnesty.org/en/latest/news/2025/02/meta-new-policy-changes/> accessed 15 June 2025. See generally, Carnegie Endowment for International Peace, ‘Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech’ (2023) <https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech> accessed 15 June 2025; Human Rights Watch, *Meta’s Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook* (21 December 2023) <https://www.hrw.org/report/2023/12/21/metass-broken-promises/systemic-censorship-palestine-content-instagram-and> accessed 15 June 2025.

²⁹ Tattle, *Uli: Reclaim your Online Space* (GitHub repository, latest commit 30 May 2025) <https://github.com/tattle-made/Uli> accessed 15 June 2025

Even well-trained algorithms may not always identify hate speech and it may also misidentify certain content as hate speech. Given that currently, Meta's algorithms are letting through vast majority of hate speech (93% of all hate speech reported to remain on Facebook. This includes content advocating violence, bullying, and use of slurs and other forms of Tier 1 hate speech)³⁰. Thus, it becomes imperative to use automated moderation with human intervention, especially in high-risk contexts, like in cases of intercommunal conflict. Meta must invest in local content moderators and create a system of moderators within each of its markets, who are capable of understanding language, culture and political contexts. The Centre for Internet and Society has recommended that there must be meaningful representation of minorities across staff and contractors of tech companies like Facebook. Content moderation mechanism must be transparent and should include the hiring practices, contractor demographics, and slur lists. Thus, any takedown can be appropriately understood by the user and free speech can be upheld. Sole reliance on automated enforcement will lead to disproportionate application of Meta's Policies which will undermine the voices of the people and exacerbate hate speech and violence on its platforms.

³⁰ Equality Labs, *Facebook India: Towards The Tipping Point of Violence: Caste and Religious Hate Speech* (2019) https://equalitylabs.wpengine.com/wp-content/uploads/2023/10/Facebook_India_Report_Equality_Labs.pdf accessed 15 June 2025