The Legal Journal on Technology

Public Comment: Case 2025-023-FB-UA (Kenyan Politics Using a Designated Slur)

Summary: In this public comment to the Oversight Board, The Legal Journal on Technology (TLJT) addresses issues regarding Meta's content moderation of the Kikuyu term "tugeges" in Kenya, contending that Meta's definition of the word as "retarded Kikuyu" is an extreme misinterpretation that could incite violence, while local authorities like the Kenya National Cohesion and Integration Commission view it more as a political descriptor for blind followers. TLJT contends that Kenya's history of ethnic violence, particularly the 2007-08 post-election crisis, creates a volatile context where hate speech can escalate to physical violence but that Meta's automated content moderation systems lack the cultural and linguistic nuance needed to distinguish between harmful content and legitimate political discourse properly. We contend that over-reliance on AI moderation without adequate local language expertise and cultural understanding violates freedom of expression rights under both Kenyan and international law, as evidenced by instances of both under-enforcement (allowing actual hate speech to spread) and over-enforcement (removing legitimate content). We conclude with some policy recommendations for Meta to collaborate with State institutions and local linguistic institutions/bodies, publish specific annual risk assessments and incorporate systemic-risk audits and rights-respecting data APIs.

1. **Question 1:** The meaning of the term "tugeges" and its contemporary use in Kenyan politics and society, including evidence of its connection to harms or lack thereof.

TLJT Response: Tugeges means "*to describe people who support a certain political side is wrong."* The term Tugeges has been widely used by users of the X (formerly Twitter) app to refer to the supporters and members of the Kenya Kwanza government from the Mount Kenya region. Tugeges is a Kikuyu word (plural of Kagege) meaning someone who cannot make independent decisions despite having the ability to think. Their decisions are often wrong and based on emotions or hatred for someone. It can also refer to someone who follows another person blindly without knowing the destination. The Kenyan National Cohesion and Integration Commission ('KNCIC') has also observed that *"this word is hurtful, demeans and dehumanizes people. It is important that we stop using this and any other abusive words to describe fellow Kenyans."* However, at this moment, it should be noted that the Kenyan government or KNCIC has not banned the word.

The definition used by Meta for the term Tugeges means "retarded Kikuyu," noting it is the plural form of "kagege," which is a Kikuyu term to describe "a person who is extremely confused to the point of gaping vacantly at the world." This definition used by Meta is not only an extreme form of interpretation. Still, it may also have the tendency to incite violence when people know about this definition, considering Meta to be a popular social media platform. The KNCIC defined the word as someone supporting a particular political ideology or side, i.e., the Kikuyu people. However, the definition used by Meta exemplifies this and tends to mean that supporting a particular political ideology means 'retarded.' This is a very dangerous definition or meaning to take. Furthermore, while there is no denial that the word has often caught on a negative connotation, the extreme misinterpretation by Meta to avoid accountability and effort of understanding the context in each specific case makes it a fault on Meta's part.

While we understand the initiative Meta took to remove the usage of 'slurs' or 'harmful words/phrases' on the social media platform, which is an important and popular way of reaching out to mass. However, we wish to maintain that in doing so, Meta should not use an interpretation of its own, which may, instead of helping the case, cause harm or lead to over-enforcement—the harm not only in terms of protest but also affecting freedom of speech and expression. Hence, *firstly,* the Meta should not only rely on its definition but should also understand the local meaning of the word or phrase. *Secondly,* it should not take down posts just based on the

word(s)/phrase(s), which may be a slur as per its definition, but should see the usage of the word(s)/phrase(s) in the context of it being used.

2. **Question 2:** The political context in Kenya, including the situation for freedom of expression online and risks related to incitement of intercommunal violence.

TLJT Response: The Kenyan political landscape is deeply affected by ethnic divisions, which profoundly shape its online discourse and increase the risk of inter-communal violence. This analysis of Kenya's political situation shall attempt to simplify the political context in Kenya for the Board's decision on content moderation practices, particularly concerning the term "tugeges." The 2007-08 Kenyan Crisis (in which nearly 1200 people were killed and thousands displaced), which was post-election violence, is a lesson from history about how ethnically charged rhetoric can cause widespread violence. While constitutional reforms were enacted, ethnic strife is still exploited by political actors.

Political players in Kenya frequently impact online freedom of expression through various actions. They utilize restrictive laws like the Cybercrimes Act, leading to arrests and prosecutions of critics, such as bloggers Robert Alai and Cyprian Nyakundi, for "false publications." While less frequent, past internet slowdowns during elections hint at control over information flow. Furthermore, state-aligned "digital brigades" actively spread propaganda and disinformation, dominating online narratives. These combined efforts create a chilling effect, undermining open online discourse and self-expression.

Digital Platforms are central to public discourse in Kenya, with high penetration of the internet and mobiles, making them a concrete part of political engagement among the masses. However, the democratization of the digital platform may also result in the amplification of hate speech, misinformation, and incitement, which reaches its zenith during election periods. For example, the 2022 general elections saw widespread disinformation of a political nature, fake news, and manipulated videos on platforms like X and TikTok. The term "tugeges," a highly charged derogatory term specifically targeting the Kikuyu community, directly reinforces divisive "us vs. them" narratives and heightens the potential for inter-communal friction.

The translation of online rhetoric into offline repercussions in Kenya is not a new process and is pretty prevalent in Kenyan History. Hate Speech is often known as a step before violence, and political figures have often exploited the public platforms to meet their electoral ends. Examples from electoral cycles include widespread hate speech, incitement into violence during the 2007-2008 PEV on SMS and nascent social media, large numbers of threats and violence in 2013, and "protracted dispute, incitement to hatred and violence" in 2017. The 2022 elections are of particular relevance to the case at hand, where coded language like "madoadoa" and "tugeges" gained entry into popular discourse and caused violence among groups. Crossing into the period post-elections, the anti-government protests in 2023 have also featured online hate speech, leading to a lowered Freedom on the Net rating.

Despite the presence of robust protections for Free speech in Kenya, legal ambiguities in acts like the Computer Misuse and Cybercrimes Act of 2018 and the National Cohesion and Integration Act (NCIA) of 2008 are plagued with significant challenges such as slow action and reduced freedom of speech due to arbitrariness. Despite being mandated to combat hate speech, national enforcement bodies like the National Cohesion and Integration Commission (NCIC) face considerable challenges, mostly from political interference and corruption. This causes reduced freedom of speech in Kenya due to interference from politicians and the government. The inadequacy of enforcement in Kenya places a higher threshold of responsibility on Meta to protect freedom of speech on its platforms, prevent censorship, and reduce restrictions on free speech.

Meta's content moderation efforts in Kenya have been inadequate. Despite Meta's claim of maintaining "dedicated teams of Swahili speakers," independent research reveals that Meta's AI suffers from both over-removal" and "slow removal" of harmful content, restricting free speech in the Global South. A Global Witness investigation found that Facebook "appallingly failed to detect hate speech ads" in Swahili and English just weeks before the 2022 Kenyan election, including ads that dehumanized groups and called for violence. This

implementation gap, without adequate investment, increases the risk of ethnic violence. Meta's recent decision to scale back traditional fact-checking services in favor of "Community Notes" also raises substantial concerns, particularly in conflict-prone regions like Africa, as such systems are deemed "ill-equipped" for hate-driven narratives and vulnerable to manipulation by state-paid influencers.

3. **Question 3:** How ethnicity is discussed in relation to current affairs in Kenya, and the prevalence of contested or potentially insulting language or generalizations in everyday discourse?

TLJT Response: Kenyan society is divided into five major groups: Kikuyu, Luo, Luhya, Kamba, and Kalenjin, which account for 70% of its population. The Kikuyu have played a significant role in Kenya's independence and the Mau Mau movement. It has resulted in their dominance in business and politics, leading to interethnic rivalries and resentment, which has hindered the efforts towards national integration. There is a pattern by which interpersonal trust diminishes with social distance, which is a common feature of how all ethnic groups in Kenya view the wider world, thereby resulting in ethnicity as a major source of political and economic divisions. The 2007-2008 post-election violence revealed deep-seated inter-ethnic suspicion and hatred. According to the Commission of Inquiry into Post-Election Violence (Waki Commission), one of the root causes of the post-election violence was the underlying climate of tension and hate amongst the various ethnic groups in Kenya. To address these divisions, the Parliament enacted the National Cohesion and Integration Committee (NCIC) Act in December 2008, which mandated the elimination of discrimination on ethnic, racial, and religious basis and the promotion of national reconciliation and healing. It prohibits ethnic discrimination' at all times and places. This provision is complemented by Articles 27(4) and 27(5) of the Constitution, both of which prohibit discrimination on the basis of ethnic identity.

Hate speech and potentially insulting language are highly prevalent in Kenya's everyday political discourse, especially during elections. A key example of this is the referendum on a new Constitution, which was held in November 2025; it led to a confrontation between Kibaki, that represented the Kikuyu ethnic group, and Raila Odinga, who was seen as representing other ethnic groups. This period saw an intense use of hate speech at rallies, meetings, and on the airwaves, which had recently been liberalized and which had a fair component of ethnic language stations either owned by the political class on both sides or aligned with them. The Kenya National Commission on Human Rights found many statements that dehumanized people of different ethnic groups, as well as "elevated" those communities that the speakers hailed from. Yet, authorities refused to prosecute or even investigate the authors of the statements despite the existence of explicit legal provisions under Sections 96 and 391 of the Kenyan Penal Code. Most hate speech cases in Kenya have been dropped due to a lack of evidence or offenders, especially politicians who have received lenient punishments like public apologies. The political class are notorious producers of hate speech; however, they know how to get away with their acts. The NCIC has issued a list of hate speech terms that can be monitored on social media platforms and in public gatherings. Through this, the possible sources and potential trends of an ethnic conflict can be identified. During elections, underlying ethnic rivalries are rekindled through songs, words, euphemisms, epithets, and slurs. Ordinary words and catchphrases can assume a different meaning altogether when used during campaigns, depending on the context.

4. **Question 4:** 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?

TLJT Response: Kenya is a nation that has always battled with communal violence, and Meta's reliance on automated systems for its Hateful Conduct Policy does not help much to ease up the tensions and, in some cases, has had a negative impact on human rights, freedom of expression in a country like Kenya. The major drawback of relying on automated systems for content moderation is that they *lack the historical and cultural nuances* that

are very much needed to comprehend a language and its use. In Afghanistan, when Meta employed a few local language speakers to moderate the content being shared, less than 1 percent of content was taken down compared to 77 percent of non-violent Arabic content across the Middle East, where local language speakers were not employed.

This lacuna in some situations has led to violence-inciting content slipping through the cracks, as it was in the case of the 2022 Kenya elections, after which Meta has itself admitted how non-English languages often slip through the cracks in detecting hate speech. However, there have also been instances where these automated systems have *led to the over-enforcement* of rules, restricting people's freedom of speech. So was the case of all abortion-related posts being taken down the internet because they had the word 'kill' in them in the USA following the *Dobbs vs. Jackson Women's Health Organisation* judgment. The process also lacks transparency, leaving users with little to no recourse to appeal when their posts are wrongfully taken down. The words used in satire or have a cultural connotation are taken down due to the lack of ethnic inclusivity in automated systems while processing data. Meta has been accused of not sharing key information with its users, like its content moderation practices. Meta's response has also been vague in some cases, as words like hate speech have not been defined under its policy. The transparency reports provided only give statistics and do not provide any factors taken into consideration or the rationale behind the conclusion.

Also, before this controversy, there were instances when the Kenyan courts called Meta out and exerted their jurisdiction over Meta's operations in the country when its policy did not align with Kenya's local legal standards, particularly the Kenyan constitution. The cases were also filed against Meta for its involvement in the Ethiopian Conflict through its algorithmic bias. Similarly, an organization in Yugoslavia was falsely labeled as a criminal organization by the automated systems. These are the exact cases that portray why solely relying on automated systems is not only a bad idea but also leads to dangerous consequences and puts the whole community at risk and not just the platform users and the automated systems need to be backed by skilled human agents who are familiar with the country's cultural setting.

The wrongful takedowns of legitimate speech and the failure to take down the wrongful content by the automated systems used for content moderation also come in direct violation of the rights of freedom of expression granted under Article 33 of the Kenyan Constitution and Article 19 of the International Covenant of Civil and Political Rights which Kenya is a signatory of. Researchers have found that content moderation in Kenya is often done without understanding the local realities, which is made further worse by the lack of transparency in the operations of these systems. For regulating hate speech across the world within the human rights law framework, provisions have been given in both international human rights and local human rights treaties. Still, Meta has not provided an extensive and precise definition of hate speech, complicating the issue further. This makes it difficult for civil society and the media to hold platforms accountable for their human rights impact.

In short, as much as automation is needed to resolve the issues on a larger scale, Meta's present approach risks silencing valid voices and not preventing actual harm. A more balanced, context-sensitive system is overdue for Kenya and places like Kenya.

5. Policy Recommendations

Kenya's 2022 general election provided a textbook illustration of how digital platforms convert ethnic antagonism into offline violence. The Armed Conflict Location & Event Data Project recorded 206 discrete incidents of politically linked violence and demonstrations in the twelve months that followed the vote, many accompanied by Kikuyu-targeted epithets such as *tugeges*, a term the National Cohesion and Integration Commission (NCIC) later condemned for "demeaning and dehumanising" Mount-Kenya voters.' *Freedom on the Net 2024* likewise documents a steady erosion of digital liberty, noting blogger intimidation and the first

government-ordered messaging blackout. Such findings confirm that *tugeges* functions today much as *madoadoa* did on the eve of the 2007–08 killings: an ethnic dog-whistle whose online circulation tracks tangible spikes in street-level unrest.

Meta's present response remains largely lexical and belated. A July 2022 Global Witness "red-team" test showed Facebook approving paid advertisements that explicitly advocated ethnic cleansing in both English and Swahili. The company did not add *tugeges* to its Hateful Conduct lexicon until January 2024, and its May 2025 Community Standards Enforcement Report still shows significantly higher error rates in non-English hate-speech detection. Compounding the problem, Meta in January 2025 scrapped third-party fact-checking across Africa and narrowed its hateful-content definitions, moves that prompted the African Commission on Human and Peoples' Rights to adopt Resolution 630/2025, warning that such retrenchment could "impinge multiple human rights." These developments breach the due-diligence duty articulated in the UN Guiding Principles on Business and Human Rights, which oblige firms to prevent foreseeable rights harm where they operate. To overcome these problems and better train its automated system, Meta can collaborate with institutions like Kenya National Commission on Human Rights which works as a watchdog over the government and collaborates with the institutions like Meta to redress the issue of racism. Furthermore, Meta can collaborate with the universities and local linguistic institutions/bodies of the country through its "Language Technology Partner Program" to train its software on local languages while at the same time building a database for the local language with cultural context, improving real-time moderation accuracy.

A rigorous remedy begins with systemic-risk due-diligence that matches the European Digital Services Act (DSA) standard. Meta should publish a Kenya-specific annual risk assessment detailing how ranking, virality features and ad-targeting amplify tribe-coded speech. Summaries benchmarked against the Santa Clara Principles 2.0 must disclose precision-and-recall metrics for Swahili, Kikuyu, Luo and Sheng classifiers, and an independent auditor; appointed under the DSA's Article 37 methodology, should verify the findings. Such transparency would satisfy the UNGP duty of "knowing and showing," while giving civil society a baseline from which to measure progress.

Audit alone is insufficient without local linguistic capacity. Meta should establish a permanent Trust & Safety Hub in Nairobi, staffed around the clock by certified reviewers fluent in Kenya's four dominant languages. NCIC-flagged posts that meet the Rabat Plan of Action six-factor threshold for incitement would be removed within minutes; borderline content would enter a forty-eight-hour "friction" state that throttles resharing pending expert review. Nigeria's 2023 election provides a precedent: a fact-checker coalition processed 1.25 million items in two days and supplied real-time debunks to platforms and regulators. Finally, Meta must open a privacy-preserving researcher API; a key ask of ACHPR 630 - so accredited scholars and prosecutors can query aggregate enforcement logs. The API should incorporate annotated samples from AfriHate, a new 15-language corpus that sharply improves classifier recall in African contexts. Together, systemic-risk audits, a Nairobi hub and a rights-respecting data API would transform Meta's Kenyan presence from reactive slur policing to proactive human-rights stewardship, aligning commercial practice with the minimal expectations of the DSA and the UNGP while preserving the civic space Article 19 protects.