Meta's content moderation regulations and decisions have come under fire for a variety of reasons through the years. In the early 2020s, Meta made the decision to automate most of its content detection and moderation. More recently, Meta's decision to forego third-party fact checking organizations in favour of "community notes", a feature that is still in its testing phase, garnered significant backlash. Concerns were raised around

In the context of the current case around the usage of a 'designated slur' in Kenyan politics, based on the presented facts of the case, it is important to reconsider deferring to automation when it comes to moderating certain kinds of content. This is primarily because AI, no matter how robust and good, can sometimes lack the ability to provide nuance, especially in politically charged situations. As such, the main intention behind this particular public comment is to ensure that free speech is balanced with protecting other prevailing human rights. Over moderating can often have a chilling effect on political discourse online, especially if done so without understanding the specific social, cultural, political, and even language contexts.

1. The meaning and contemporary use of the word 'tugege' in Kenyan politics and society

'Tugege' is a Kikuyu word used to describe a 'dumb or ignorant person'. It derives from the Kikuyu word *'gega'*, which means 'to stare at something non-stop, as if open-minded'. The Kikuyu are Kenya's largest ethnic group, and are native to lands near Mt. Kenya. Consequently, Meta translates the term *'tugege'* to mean a 'retarded Kikuyu', even though the prevailing definition of the word from Kenyan sources does not make this distinction with respect to any specific ethnicity.

The word gained popularity during Kenya's 2022 elections, and has remained in the political discourse on and off, since. In September of 2023, Kenya's National Cohesion and Integration Commission (NICC) refuted claims on social media that it was going to ban the usage of the word *'tugege'*. In April of 2022, prior to Kenya's presidential elections, the NCIC published a list of 'banned' words—words that per the Commission, amounted to hate speech. Notably, the word *'tugege'* was not a part of this list.

However, in a more recent speech from 18th June 2024, on the International Day for Countering Hate Speech, Rev. Dr. Samuel Kobia, the NCIC Chairman stated that "[t]he use of the word words such as *"tugeges"* to describe people who support a certain political side is wrong. 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." In spite of this, the word has not been formally designated as a slur by the NCIC. Further, even though Meta's reasoning for removing the comment containing this word follows the January 2024 designation of the word as a slur, presumably from one of its 'market-specific slur lists', the fact also remains that these lists aren't made public. This means that people are more likely than not unaware of what word, even outside of the specific social, cultural, and other contexts, is considered a slur under Meta's community standards.

That is not to qualify whether or not the word *"tugege"* is offensive, but rather that it is important to understand the specific socio-political context in which the word was used, especially given Kenya's fraught electoral history.

In spite of how politically charged it is; thus, it remains to be determined whether *'tugege'* is just a crude insult, or a slur that should be categorised as hate speech. Considering the context, it's important to

leave that decision up to the Kenyans and/or the NCIC first, before we make the decision of removing any content that so uses the word.

2. The political context in Kenya

Kenya was marred by a violent political crisis in late 2007 and early 2008 following the 2007 presidential elections. The Commission to Investigate Post-Election Violence (CIPEV), more commonly known as the Waki Commission estimated 1,133 people dead and 663,921 internally displaced in the aftermath of the electoral violence. The general consensus among the international community was that all parties engaged in corrupt and violent activities leading up to the elections.

Former United Nations Secretary General Kofi Annan then mediated a deal between the two contenders Mwai Kibaki and Raila Odinga, which led to a coalition government being formed, effectively putting an end to the conflict. Later, the International Criminal Court issued warrants charging crimes against humanity to arrest many considered responsible for these deaths—deaths that were largely linked to tribalism and ethnic conflict.

This is an especially important history lesson to keep in mind in the context of free speech, because following the announcement of the election results, the then Information Ministry in Kenya banned all coverage of the same, citing fear of increased violence.

Following this, in 2008, the National Cohesion and Integration (NCI) Act established the National Cohesion and Integration Commission to regulate hate speech and promote national cohesion and integration. Among other things, the Act (and the Commission) work to ensure that speech does not lead to any kind of ethnic hatred, nor incite violence. In 2010, the Kenyan constitutional referendum also expressly prohibited hate speech, stating that "the right to freedom of expression does not extend to incitement to violence, hate speech and advocacy of hatred that constitutes ethnic incitement, vilification of others or incitement to cause harm."

Prior to the 2022 Kenyan elections, social media platforms, including Meta introduced a series of measures to reduce the spread of misinformation, hate speech, and other harmful content. Among its many measures, Meta also partnered with local election experts, and independent fact-checkers proficient in both English and Swahili. To combat mis- and disinformation, Meta also worked with local radio stations (in a variety of local languages, including Luo, Kikuyu, and Swahili) in Kenya to help users better spot and protect themselves from hate speech and misinformation.

Working with local experts, journalists, and fact-checkers helps bolster automated efforts to moderate and regulate content. This is because the locals can provide much needed context and other information that can be crucial in determining the intent, and sometimes even the meaning of some forms of speech. In places like Kenya, where political violence and censorship have both existed in tandem in the past, this context can become especially important. The lack thereof could lead to a slippery descent in either direction.

Meta has had problems with this in the past, most famously in the instance of Myanmar. The Buddhist majority military officials in Myanmar used the platform to spread disinformation about the minority Rohingya Muslims in its Rakhine State. This led to what humanitarian agencies called a textbook example of ethnic cleansing—over 700,000 Rohingya Muslims had fled the state into neighbouring Bangladesh by the August of 2018.

Facebook had had almost no system of content moderation here. In a country where 'Facebook' was synonymous with 'internet', and millions used the platform, almost all the content moderators hired spoke no Burmese. In 2015, there were two moderators who spoke Burmese, reviewing content from 18 million users. In 2018, when the situation came to light, Meta had no content moderators in Myanmar itself. Instead, it used to outsource its content moderation to countries like Malaysia. Even here, there were only 60 people to review posts from the entire country.

3. The impact of automated enforcement of Meta's Hateful Conduct (and other) policies on human rights

Mark Zuckerberg claimed rather accurately, in 2018, that it is easier to make an AI detect nipples than it is to detect hate speech. This was indeed Meta's content moderation history at the time. Nudity and advertisements about safe sex had been censored more easily than actual hate speech.

At the time, that was because of the inherent imagery of the posts in the first two categories, making automated detection much easier. Detecting, identifying, and removing content from the other categories, on the other hand, is inherently more difficult. For one, it is very difficult to determine intent, and for another, no matter how robust, such determination will always be arbitrary. Thus, people reposting racist rants that they were on the receiving end of, or an ad for an exam that featured the word 'sexy', tended to get flagged, creating tremendous uproar.

Since 2018, sadly, hate speech remains one of the more difficult 'categories' to moderate. Recent studies have shown the unique problem that automated content moderation poses to countries in the global south, mainly due to a lack of understanding of the prevailing cultural contexts. Such content is often plagued by both censorship of content that does not violate any laws, guidelines, or regulations—'over removal' of lawful content, as well as a failure to tackle actually harmful content in an accurate and timely manner—'slow removal' of harmful content. This problem is only exacerbated by the lack of resources and data to train models to moderate content in languages other than English. Simply, there isn't enough data in 'local', Global South languages to study harmful content in these languages, nor enough data to use to train models to more accurately moderate content automatically.

There is also a growing concern among speakers of the language around the increased proliferation of hate speech and misinformation in Swahili online. Meta's decision to apply uniform content moderation policies to all Swahili speakers and their content does not consider specific national, cultural, economic, and any other contexts whilst regulating speech in that language. The same word, phrase, or statement may hold different meaning to Swahili speakers in say, Tanzania, as opposed to those in Kenya. Instead, working with both scientists who design the AI, as well as human moderators, from all Swahili-speaking nations and regions (as some other platforms do), who are aware of specific regional and cultural contexts along with the language itself, could go a long way in reducing both over removal and slow removal of content.

The High Court in Kenya also recently ruled that charges can be brought against Meta for how it handled hate speech on its digital platforms during the recent Ethiopian civil war. The suit alleges not just that Meta failed to properly regulate hate speech and misinformation, but rather that its algorithms pushed such content.

The point remains here on whether automated moderation is really the solution to regulating content, in a situation (and country) where it doesn't really seem to be working. Further, Meta's recent decision

to move away from independent, third-party fact checkers would only exacerbate this problem. Misinformation and hate speech are looming problems, especially during elections, and robust content moderation plays an important role in safeguarding democracy and human rights. In light of Meta's recent announcements to also roll back its automated flagging of hate speech and other harmful content, albeit only in the United States currently, there's also a lack of surety around where automated moderation itself stands. Simply, automated moderation, at least as it stands now, is not sophisticated enough to understand the nuances of human language and speech, and this is especially true for languages like Swahili.