Considering Kenya’s fraught history with election-related violence, any discussion around election preparedness ought to address hate speech and incitement to violence. On 25 October 2021, the Government of Kenya announced the convening of a multiagency team tasked with charting a course to free, fair, and credible general elections. The team is chaired by Chief Justice Martha Koome, who announced that the Judiciary of Kenya would be setting up five specialized courts in Nairobi, Mombasa, Nakuru, Kisumu and Eldoret to deal with hate speech cases in the run up to, and during, the general elections.

Concerns around the likelihood of political rhetoric stirring violence are understandably higher when such rhetoric is disseminated through social media platforms that are, by nature, peer-to-peer, instantaneous, and in some case, encrypted. In fact, the National Cohesion and Integration Commission (NCIC), while embarking on a nationwide civic education drive, noted that hate speech disseminated through social media is currently the biggest challenge it faces. With recent revelations that social media platforms such as Facebook are often ill-equipped or unwilling to handle the spread of harmful content in the first place, let alone in contexts within the Global South, it is worth exploring how best stakeholders can work together to mitigate the potential impact of
hate speech or inciteful rhetoric disseminated through such platforms.

**What is hate speech?**

During the 2007/8 election cycle, several local radio stations are believed to have facilitated the spread of inciteful political rhetoric, often in vernacular. At the time, Kenya did not have a law specifically defining or criminalising hate speech. The dissemination of such inciteful political rhetoric contributed to election-related violence that resulted in numerous deaths and massive internal displacement. Following this deeply tragic episode, Kenya enacted the National Cohesion and Integration Act (the Act) which, for the first time, defined hate speech in Kenya and established the NCIC. The definition of hate speech adopted in the Act broadly entails two major components: (i) the use or spread of content that is threatening, abusive or insulting, and (ii) the intent to stir up ethnic hatred (or, if such hatred would be the likely outcome, whether intended or not). Given the context in which this definition was developed, it is unsurprising that the core conceptual focus is ethnicity (though this has been defined in the Act to include race and nationality). A person convicted of hate speech is liable to a fine not exceeding KES 1,000,000 or to imprisonment of up to 3 years, or to both.

**Precarious balance**

Much like several hate speech laws around the world, the breadth of the definition of hate speech in the Act has been criticised as potentially stifling free expression, particularly due to the criminal sanction. While the Constitution of Kenya—which was promulgated less than three years after the Act—provides hate speech as an exception to the freedom of expression, restrictions on speech under the Constitution ought to be proportionate. In other words, the restriction of speech which is of a hateful or inciteful nature should not be done in a way that would put other forms of speech (such as healthy political debate) at risk. Indeed, some attempts by the government to prosecute hate speech have been met with the criticism that they are politically motivated. While these actions (which are grounded in law) have faced backlash, social media platforms operating in Kenya have continued to regulate hateful and inciteful content on their own terms.

Despite the NCIC being tasked with the investigation of matters relating to ethnic hatred under the Act, only a small portion of hateful content disseminated through social media seems to have come under its radar based on available evidence. Recently, a legislator—Oscar Sudi—was charged with hate speech over remarks made on a video which was uploaded to Facebook. This prosecution likely occurred due to the notoriety of the individual in question. Outside such instances, hateful or inciteful rhetoric which may qualify as hate speech under the Act have remained under the purview of social media platforms which have continued to regulate such content rather opaquely, and primarily based on instructions provided by foreign regulators. Leaving such a consequential task to private platforms raises several concerns and also calls into question the effectiveness of current law enforcement efforts.

**Content moderation by platforms**

A large amount of content is shared online through social media. Consider Twitter, for example, where on average, 6,000 tweets are sent out each second. The likelihood of some of this content being problematic in one way or another is quite high, even more so where there is no agreement on what constitutes “problematic” content. Some form of control is advisable to mitigate arising harm. These platforms are, to varying degrees, self-regulating; due to their private nature and their free speech rights in the jurisdiction in which they are established (the US), social media platforms can develop and enforce guidelines that their users ought to adhere to. They often use artificial intelligence (AI) technologies to scan user content for any infringing material and to take a
predetermined action such as taking the content down, downranking it, or flagging it for human review. This practice is suspect, as it has been well established that algorithms suffer from various forms of bias, and they are rarely optimized for “foreign” speech nuances and customs.

Social media platforms operating in Kenya have continued to regulate hateful and inciteful content on their own terms.

The platforms also permit other users to report content for human review. The prescriptions in these guidelines which determine the treatment of content (both by human reviewers and by AI) sometimes overlap with speech restrictions in the jurisdictions in which they operate. For example, Facebook’s community standards define and prohibit hate speech. However, there are sometimes glaring conceptual differences between the private sector definitions and those imposed by law or a government regulator. In such cases, a government may directly request the platform to take down the content based on a local law violation. According to these platforms, when faced with such requests, they assess the content in question primarily against their own guidelines. In other words, where the content does not run afoul of their guidelines, they would only make it unavailable in the jurisdiction where the government has made a takedown request. For example, if the government of Kenya, through the NCIC and the Communications Authority, were to request Facebook to take down a particular post on the basis that it falls under the Kenyan definition of hate speech, Facebook would consider its own definition of hate speech in its community standards. If there is an overlap and Facebook believes that the content truly amounts to hate speech, then it would comply. Otherwise, it would simply restrict access to the content within Kenya and leave it up for the rest of the world, which would also not prevent Kenyan users from accessing it using virtual private networks (VPNs).

The enforcement of these guidelines, and the design of these platforms (particularly how content is promoted to certain users), have long been the subject of criticism for, among other things, opacity and a lack of accountability. These concerns, as has been made clear in recent weeks, are more pronounced in the Global South. While these platforms place themselves in a position where they make consequential decisions on how to handle content such as hate speech, they are often ill-equipped to do so. From AI technologies which have been trained on biased datasets, to human reviewers with insufficient contextual background and local knowledge, these platforms are presently not up to the task of moderating content at scale in countries such as Kenya, more so in electoral contexts.

Recent reporting following the leak of the “Facebook Papers” chronicles a story of neglect of Third World countries when it comes to the deployment of adequate moderation tools often resulting in real-world violence or harm. For example, despite flagging Ethiopia as an at-risk country with insufficient resources to detect hate speech in local dialects, Facebook failed to improve its AI detection technologies and to hire additional moderators familiar with the local context. During the conflict in Tigray (which has since exacerbated), Facebook’s internal teams were aware of the insufficiency of their efforts. Facebook recently indicated that it has since improved its moderation efforts in Ethiopia.

However, the problem facing Facebook, and other platforms, is systemic. Reactionary, one-time solutions to a problem with grave real-world outcomes such as violence and death are unsustainable. Bearing in mind the similar accounts of failures in content moderation resulting in real-world harm that have been documented in Myanmar, India, Nigeria, and Palestine (to name a few), it is crucial to reconsider the extent of oversight these platforms are subjected to, and the level of collaboration between stakeholders to ensure harms are mitigated during politically charged situations.
Mitigation of harms

In prior elections, both sides of the political contest used inciteful speech to fuel the emotions of voters, with disastrous consequences. There is no reason to expect that the coming election will not include this tactic, supporting the argument that there is need for some action to mitigate the potency of this tactic. The government recently launched the National Computer and Cybercrime Coordination Committee (dubbed “NC4”), which is a committee provided for under the Computer Misuse and Cybercrimes Act. The NC4 is responsible for consolidating action on the detection, investigation, and prosecution of cybercrimes. The Cabinet Secretary for Interior and Coordination of National Government recently indicated that the NC4 would prioritize the misuse of social media in the run-up to the general elections, raising the likelihood of arrests and prosecutions under the Computer Misuse and Cybercrimes Act over the next year. It would therefore seem that through the NCIC and the NC4, the government has doubled down on policing the spread of inciteful rhetoric online – a herculean task that may well jeopardise the space for political speech.

Reactionary, one-time solutions to a problem with grave real-world outcomes such as violence and death are unsustainable.

Considering the incompatibility of the online communication ecosystem with traditional detection and prosecution methods, attempts at regulation that do not factor in the role of social media platforms and other stakeholders are bound to encounter challenges. Without adopting a collaborative policy attitude toward the issue, the government may easily find itself turning to internet shutdowns to mitigate the perceived harm of inciteful rhetoric or to silence criticism.

As opposed to priming law enforcement agencies for crackdowns on content disseminated through social media, entities such as the NCIC, NC4, as well as the Independent Electoral and Boundaries Commission (IEBC) should consider working more closely with social media platforms and other stakeholders in media and civil society. Such collaborations can be aimed at building the capacity of social media platforms’ content moderation tools used in Kenya and fostering transparency in the conduct of these platforms to enable oversight. The inclusion of civil society would also serve to hold both government and platforms accountable for their conduct.

The risk posed by inciteful political rhetoric demands a comprehensive and inclusive approach. Kenya cannot afford to entrench mistrust by relying solely on prosecutorial action that may, in some instances, be politically motivated, and is typically wholly oblivious to the harms posed by the conduct of social media platforms. Any efforts at mitigating the impact of hate speech on social media should not ignore the fact that numerous stakeholders have a role to play, though with varying degrees of importance. Crucially, these efforts should not detract from the space for healthy civic engagement.

Political actors must recognise their centrality to the nature of discourse around the forthcoming election that takes place online. It is imperative for them to publicly commit to avoiding engaging in the spread of hateful, inciteful or false content. Examples such as the Election Pledge developed by the Transatlantic Commission on Electoral Integrity are instructive in this regard. Through public pledges acting as rules of engagement, political actors can signal their commitment to healthy democratic debate. These political actors should also recognise the sway they have over their supporters and proxies and should do their best to encourage positive conduct. In political party meetings and rallies, political actors should ensure that they communicate a zero-tolerance policy towards hateful or divisive rhetoric. To entrench a culture of healthy discourse, political actors should collaborate with civil society to engage the citizenry in civic education.
This is the fifth of a five-part op-ed series that seeks to explore the use of personal data in campaigns, the spread of misinformation and disinformation, social media censorship, and incitement to violence and hate speech, and the practical measures various stakeholders can adopt to safeguard Kenya’s electoral integrity in the digital age ahead of the 2022 elections. This op-ed series is in partnership with Kofi Annan Foundation and is made possible through the support of the United Nations Democracy Fund.

Published by the good folks at The Elephant.

The Elephant is a platform for engaging citizens to reflect, re-member and re-envision their society by interrogating the past, the present, to fashion a future.

Follow us on Twitter.