BRIEFING June 2022

‘NOW IS THE TIME TO KILL’: FACEBOOK CONTINUES TO APPROVE HATE SPEECH INCITING VIOLENCE AND GENOCIDE DURING CIVIL WAR IN ETHIOPIA

A conflict kills thousands, forces millions of people from their homes and involves credible accusations of war crimes. This is not Ukraine, but Ethiopia, where a civil war broke out in November 2020 and which Facebook has been accused by whistleblower Frances Haugen of exacerbating by ‘literally fanning ethnic violence’. Our new investigation, which we’ve done in partnership with legal non-profit Foxglove and independent researcher Dagim Afework Mekonnen, exposes how Facebook is extremely poor at detecting hate speech in the main language of Ethiopia and follows on from our previous investigation which showed the same in Myanmar.

Facebook says that Ethiopia is ‘one of our highest priorities for country-specific interventions to keep people safe’, and that for more than two years it has invested in safety and security measures ‘including building our capacity to catch hateful and inflammatory content in the languages that are spoken most widely in the country’. Specifically, they state that they have employed more staff who speak Amharic, and that they have technology to automatically identify hate speech in Amharic. Their efforts are ‘industry-leading’, they say.

OUR INVESTIGATION

We set out to test how good Facebook’s ‘industry-leading’ hate speech detection actually is, and if their apparent safety and security measures are really able to prevent ads that fuel violence.

We did this by identifying 12 of the worst examples of Amharic-language hate speech that had been posted on Facebook, as collated by Dagim Afework Mekonnen. We submitted the hate speech examples to Facebook as adverts to see whether they would accept them for publication or not.[1] All of the hate speech examples had previously been reported to Facebook as violating their community standards and the majority had been removed from Facebook.[2] We used equal numbers of hate speech examples targeting the three main ethnic groups of the country, the Amhara, Oromo and Tigrayans.

We didn't actually publish any of the ads. We set a publication date in the future and deleted the ads once Facebook had notified us whether the ads had been approved for publication or not.

All 12 of the ads were accepted by Facebook for publication.

We put our findings to Facebook to give them the opportunity to put their side of the story and they said that the ads shouldn't have been approved and that they've invested heavily in safety measures in Ethiopia, adding more staff with local expertise and building their capacity to catch hateful and inflammatory content.[3]

After informing Facebook of this serious problem with their content moderation in Ethiopia, and a spokesperson acknowledging that the ads "shouldn't have been approved in the first place as they violate our policies," we submitted another two examples of real-life Amharic-language hate speech to them a week later. Both ads were, again, accepted by Facebook for publication within a matter of hours.

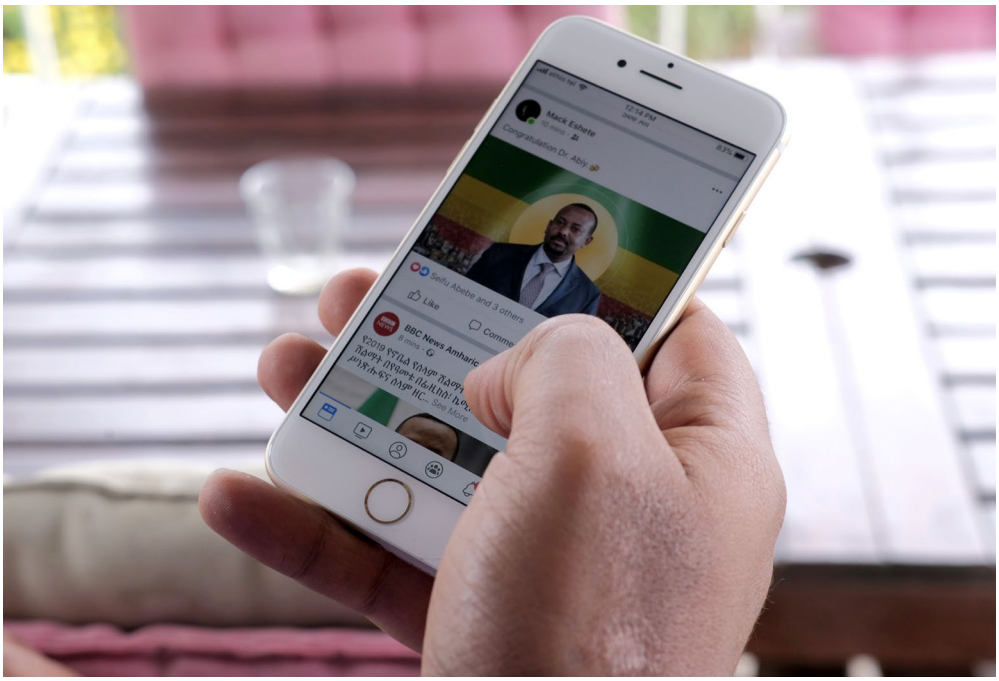
THE ADVERTS CONTAINED HATE SPEECH

The hate speech examples we used are highly offensive and we are therefore deliberately not repeating all of the phrases used here.[4] The sentences used included violent speech that directly calls for people to be killed, starved or 'cleansed' from an area and dehumanising speech that compares people to animals. Several of them amount to a call for genocide. None of the sentences were dog-whistles or in any way difficult to interpret.

All the ads fall within Facebook's definition of hate speech in their community standards and would have breached the International Convention on the Elimination of All Forms of Racial Discrimination had they been published.

WHY WE SUBMITTED THE HATE SPEECH AS ADS

Submitting hate speech in the form of an advert allows us to test Facebook's ability to detect hate speech without ourselves posting hate speech.



Global Witness calls upon Facebook to properly resource content moderation in all the countries in which they operate around the world. REUTERS/Maheder Haileselassie

Facebook says that before ads are permitted to appear online, they're reviewed to make sure that they meet their advertising policies, and that during this process they may check the ad's 'images, video, text and targeting information, as well as an ad's associated landing page'. In fact, Facebook states that it holds advertisements to an 'even stricter' standard than organic posts, and therefore it is reasonable to conclude that if they can't detect hate speech in ads, they're even less likely to be able to do so in organic posts.

CONCLUSION AND RECOMMENDATIONS

"In terms of fighting hate, we've built really sophisticated systems."

Mark Zuckerberg, speaking to the US Congress, July 2020

Facebook and other social media platforms should treat the spread of hate and violence with the utmost urgency. This isn't the first time that we have found that Facebook is unable to detect clear examples of hate speech: earlier this year we found that they were unable to detect Burmese language hate speech directed against the Rohingya minority.

Burmese and Amharic ought to be the languages that Facebook is best at analysing. Get it wrong with these languages, and there's a high risk that the real-world consequences amount to people being killed. It's sadly not an exaggeration to say that genocide is among the risks: indeed, in Myanmar, Facebook has admitted that it played a role in inciting violence during the genocide against the Rohingya. Frances Haugen, the Facebook whistleblower said in testimony to US senators that "What we saw in Myanmar and are now seeing in Ethiopia are only the opening chapters of a story so terrifying, no one wants to read the end of it."

Burmese and Amharic should also be relatively easy tests for the company: they are both the main languages spoken in countries where a lot of languages are spoken. If they're this bad in the main language of a country, imagine how bad they're likely to be in a minority language.

We call upon Facebook to:

- Properly resource content moderation in all the countries in which they operate around the world, including providing paying content moderators a fair wage, allowing them to unionise and providing psychological support.
- Assess, mitigate and publish the risks that their services impact on people's human rights and other societal level harms in all countries in which they operate.
- Publish information on what steps they've taken in each country and for each language to keep users safe from online hate.
- Publish the 'multiple forms of human rights due diligence' that they've undertaken in Ethiopia.
- Conduct an independent human rights review of its work in Ethiopia, as recommended by the company's oversight board, and to publish the findings.

We call upon governments – notable the United States' – to follow the lead of the EU and regulate Big Tech companies and force meaningful oversight, including requiring the platforms to assess and mitigate the risk that their services allow hate speech to flourish.

ENDNOTES

[1] The ads were in the form of an image (text on a plain background) and were not labelled as being political in nature.

[2] We obtained the hate speech examples from groups that are documenting hate speech on Facebook in Ethiopia, including archiving or taking screenshots of them, recording their URL, and reporting them to the platform. Five of the 12 examples are still accessible on Facebook; the others have been removed.

[3] "While these ads were removed before anyone saw them, they shouldn't have been approved in the first place as they violate our policies. We've invested heavily in safety measures in Ethiopia, adding more staff with local expertise and building our capacity to catch hateful and inflammatory content in the most widely spoken languages, including Amharic. Despite these investments, we know that there will be examples of things we miss or we take down in error, as both machines and people make mistakes. That's why ads can be reviewed multiple times, including once they go live, and why we have teams closely monitoring the situation and addressing these errors as quickly as possible." – a Meta spokesperson.

[4] Researchers interested in knowing the exact wording of the hate speech examples we used are welcome to request this from us by writing to digitalthreats@globalwitness.org