Facebook fails again to detect hate speech in ads

June 9 2022, by Barbara Ortutay

Facebook's Meta logo sign is seen at the company headquarters in Menlo Park, Calif. on Oct. 28, 2021. According to a report released Thursday, June 9, 2022, Facebook and parent company Meta once again failed to detect blatant, violent hate speech in advertisements submitted to the platform by the nonprofit groups Global Witness and Foxglove. Credit: AP Photo/Tony Avelar, File

The test couldn't have been much easier—and Facebook still failed.

Facebook and its parent company Meta flopped once again in a test of how well they could detect obviously violent hate speech in advertisements submitted to the platform by the nonprofit groups Global Witness and Foxglove.

The hateful messages focused on Ethiopia, where internal documents obtained by whistleblower Frances Haugen showed that Facebook's ineffective moderation is "literally fanning ethnic violence," as she said in her 2021 congressional testimony. In March, Global Witness ran a similar test with hate speech in Myanmar, which Facebook also failed to detect.

The group created 12 text-based ads that used dehumanizing hate speech to call for the murder of people belonging to each of Ethiopia's three main ethnic groups—the Amhara, the Oromo and the Tigrayans. Facebook's systems approved the ads for publication, just as they did with the Myanmar ads. The ads were not actually published on Facebook.

This time around, though, the group informed Meta about the undetected violations. The company said the ads shouldn't have been approved and pointed to the work it has done to catch hateful content on its platforms.

A week after hearing from Meta, Global Witness submitted two more ads for approval, again with blatant hate speech. The two ads, written in Amharic, the most widely used language in Ethiopia, were approved.

Meta said the ads shouldn't have been approved.

"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," the company said in an emailed statement, adding that machines and people can still make mistakes. The statement was identical to the one Global Witness received.

"We picked out the worst cases we could think of," said Rosie Sharpe, a campaigner at Global Witness. "The ones that ought to be the easiest for Facebook to detect. They weren't coded language. They weren't dog whistles. They were explicit statements saying that this type of person is not a human or these type of people should be starved to death."

Meta has consistently refused to say how many content moderators it has in countries where English is not the primary language. This includes moderators in Ethiopia, Myanmar and other regions where material posted on the company's platforms has been linked to real-world violence.

In November, Meta said it removed a post by Ethiopia's prime minister that urged citizens to rise up and "bury" rival Tigray forces who threatened the country's capital.

In the since-deleted post, Abiy said the "obligation to die for Ethiopia belongs to all of us." He called on citizens to mobilize "by holding any weapon or capacity."

Abiy has continued to post on the platform, though, where he has 4.1 million followers. The U.S. and others have warned Ethiopia about "dehumanizing rhetoric" after the prime minister described the Tigray forces as "cancer" and "weeds" in comments made in July 2021.

"When ads calling for genocide in Ethiopia repeatedly get through Facebook's net—even after the issue is flagged with Facebook—there's only one possible conclusion: there's nobody home," said Rosa Curling,

director of Foxglove, a London-based legal nonprofit that partnered with Global Witness in its investigation. "Years after the Myanmar genocide, it is clear Facebook hasn't learned its lesson."