NYU TANDON CYBERSECURITY FOR DEMOCRACY global witness

BRIEFING December 2022

"WE'RE GOING TO KILL YOU ALL": FACEBOOK FAILS TO DETECT DEATH THREATS AGAINST ELECTION WORKERS IN THE US, WHILE YOUTUBE AND TIKTOK SUCCEED

An investigation by Global Witness and the NYU Cybersecurity for Democracy (C4D) team looked at Facebook, TikTok, and YouTube's ability to detect and remove death threats against election workers in the run up to the US midterm elections.

The investigation revealed starkly contrasting results for the social media giants: YouTube and TikTok suspended our accounts for violating their policies, whereas Facebook accepted 15 of the 20 advertisements containing death threats that we submitted to them for publication.

OUR INVESTIGATION

We tested the ability of Meta's Facebook, Google's YouTube, and TikTok to enforce their own policies on harmful content. We did this by identifying 10 of the worst examples of death threats issued against election workers in the US and then submitting them to the three platforms in the form of advertisements and recording whether the platforms accepted them for publication or not. All the death threats were real examples of previous threats against election workers that had been reported in the media. They included statements that people would be killed, hanged, or executed, and that children would be molested.[1] We removed profanity from the death threats and corrected grammatical errors, as in a previous investigation Facebook initially rejected ads containing hate speech for these reasons and then accepted them for publication once we'd edited them.

We submitted the death threats in the form of ads as this enables us to schedule them in the future and, importantly, to remove them before they go live, while still being reviewed by the platforms and undergoing their content moderation processes.

The death threats we used as sources were all originally in English; we submitted them to the platforms in both English and Spanish. The adverts consisted of an image of a US election worker with the death threat written in clearly legible text over the top. All of the death threats were chillingly clear in their language; none were coded or difficult to interpret. All of the ads violate Meta, TikTok and Google's ad policies.

We submitted the ads on the day of or the day before the 2022 US midterm elections.

Our experimental protocols were reviewed and deemed to not be human subjects research by New York University's Institutional Review Board, which reviews the ethics of experiments involving human research subjects.

OUR FINDINGS

After we submitted the ads containing death threats, TikTok and YouTube suspended our accounts for violating their policies.

Facebook, however, behaved very differently. The platform approved nine of the ten English-language death threats for publication and six of the ten Spanish-language death threats. Our account was not closed down despite a handful of ads having been identified as violating their policies.

Global Witness approached Facebook's owner, Meta, for comment on these findings and a spokesperson responded: "This is a small sample of ads that are not representative of what people see on our platforms. Content that incites violence against election workers or anyone else has no place on our apps and recent reporting has made clear that Meta's ability to deal with these issues effectively exceeds that of other platforms. We remain committed to continuing to improve our systems."

We asked Meta for the evidence that supports the claim that the platform is better at dealing with incitement to violence than other platforms. Meta provided quotes from technology experts published in the media that note that Meta has more resources devoted than other platforms and that it does better at moderation than some alt right platforms. While these assertions may be factual, they don't constitute evidence that Meta is better at detecting incitement to violence than other mainstream platforms. In addition, there should be no tolerance for failure before a major election, when tensions and potential for harm are high.

We have carried out similar investigations into platforms' ability to detect harmful content in Brazil and the US. In both cases we looked at blatant election disinformation.

In Brazil, we found that both Facebook and YouTube accepted all of the election disinformation we submitted to them (and that Facebook continued to accept some of the ads when we submitted them a second time). We didn't test TikTok.

In the US, we found that TikTok accepted all of the election disinformation we submitted to the platform, Facebook accepted 20-50% (depending on the day and the language) and, as we found here, YouTube suspended our account.

WHAT NEEDS TO CHANGE

Platforms need to treat all users, no matter where they are in the world, equally. We are encouraged to see that YouTube has successfully detected election disinformation and death threats to election workers in the US, but discouraged to see that they failed to detect election disinformation in Brazil.

Platforms need to demonstrate that they can enforce their own policies. In particular, the track record of Facebook in being able to detect and remove the worst kinds of dangerous content is appallingly bad: their policies may look reasonable on paper but they are meaningless unless they are enforced.

The fact that YouTube and TikTok managed to detect the death threats and suspend our account whereas Facebook permitted the majority of the ads to be published shows that what we are asking is technically possible.

While the EU is taking a lead globally to regulate Big Tech companies and force meaningful oversight, platforms should also be acting of their volition to protect their users fully and equally.

We call on Meta to:

- Urgently increase the content moderation capabilities and integrity systems deployed to mitigate risk around elections.
- 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 unionize and providing psychological support.
- Routinely 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 ensure election safety.
- Include full details of all ads (including intended target audience, actual audience, ad spend, and ad buyer) in its ad library.
- Allow verified independent third party auditing so that they can be held accountable for what they say they are doing.
- Publish their pre-election risk assessment for the United States.

Global Witness is a not-for-profit organisation working to hold companies and governments to account for their destruction of the environment, their disregard for the planet and their failure to protect human rights. We have offices in London, Brussels and Washington DC.

NYU Cybersecurity for Democracy is a research-based, nonpartisan, and independent effort to expose online threats to our social fabric – and recommend how to counter them. We are part of the Center for Cybersecurity at the NYU Tandon School of Engineering.

ENDNOTES

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