GlobalResearch
Center for Research on Globalization

Facebook Approved an Israeli Ad Calling for Assassination of Pro-Palestine Activist

After the ad was discovered, digital rights advocates ran an experiment testing the limits of Facebook's machine-learning moderation.

By Sam Biddle
Global Research, January 03, 2024
The Intercept 21 November 2023

Theme: Intelligence, Law and Justice
In-depth Report: PALESTINE

*

A series of advertisements dehumanizing and calling for violence against Palestinians, intended to test Facebook's content moderation standards, were all approved by the social network, according to materials shared with The Intercept.

The submitted ads, in both Hebrew and Arabic, included flagrant violations of policies for Facebook and its parent company Meta.

Some contained violent content directly calling for the murder of Palestinian civilians, like ads demanding a "holocaust for the Palestinians" and to wipe out "Gazan women and children and the elderly." Other posts, like those describing kids from Gaza as "future terrorists" and a reference to "Arab pigs," contained dehumanizing language.

> "The approval of these ads is just the latest in a series of Meta's failures towards the Palestinian people," Nadim Nashif, founder of the Palestinian social media research and advocacy group 7amleh, which submitted the test ads, told The Intercept. "Throughout this crisis, we have seen a continued pattern of Meta's clear bias and discrimination against Palestinians."

7amleh's idea to test Facebook's machine-learning censorship apparatus arose last month, when Nashif discovered an ad on his Facebook feed explicitly calling for the assassination of

American activist Paul Larudee, a co-founder of the Free Gaza Movement.

Facebook's automatic translation of the text ad read: "It's time to assassinate Paul Larudi [sic], the anti-Semitic and 'human rights' terrorist from the United States." Nashif reported the ad to Facebook, and it was taken down.

The ad had been placed by Ad Kan, a right-wing Israeli group founded by former Israel Defense Force and intelligence officers to combat "anti-Israeli organizations" whose funding comes from purportedly antisemitic sources, according to its website. (Neither Larudee nor Ad Kan immediately responded to requests for comment.)

Calling for the assassination of a political activist is a violation of Facebook's advertising rules. That the post sponsored by Ad Kan appeared on the platform indicates Facebook approved it despite those rules. The ad likely passed through filtering by Facebook's automated process, based on machine-learning, that allows its global advertising business to operate at a rapid clip.

"Our ad review system is designed to review all ads before they go live," accordingOpens in a new tab to a Facebook ad policy overview. As Meta's human-based moderation, which historically relied almost entirely on outsourced contractor labor, has drawn greater scrutiny and criticism, the company has come to lean more heavily on automated text-scanning software to enforce its speech rules and censorship policies.

While these technologies allow the company to skirt the labor issues associated with human moderators, they also obscure how moderation decisions are made behind secret algorithms.

Last year, an external audit commissioned by Meta found that while the company was routinely using algorithmic censorship to delete Arabic posts, the company had no equivalent algorithm in place to detect "Hebrew hostile speech" like racist rhetoric and violent incitement. Following the audit, Meta claimed it had "launched a Hebrew 'hostile speech' classifier to help us proactively detect more violating Hebrew content." Content, that is, like an ad espousing murder.

Incitement to Violence on Facebook

Amid the Israeli war on Palestinians in Gaza, Nashif was troubled enough by the explicit call in the ad to murder Larudee that he worried similar paid posts might contribute to violence against Palestinians.

Large-scale incitement to violence jumping from social media into the real world is not a mere hypothetical: In 2018, United Nations investigators foundOpens in a new tab violently inflammatory Facebook posts played a "determining role" in Myanmar's Rohingya genocide. (Last year, another group ran test ads inciting against Rohingya, a project along the same lines as 7amleh's experiment; in that case, all the ads were also approvedOpens in a new tab.)

The quick removal of the Larudee post didn't explain how the ad was approved in the first place. In light of assurances from Facebook that safeguards were in place, Nashif and 7amleh, which formally partners with Meta on censorship and free expression issues, were puzzled.

Curious if the approval was a fluke, 7amleh created and submitted 19 ads, in both Hebrew and Arabic, with text deliberately, flagrantly violating company rules — a test for Meta and Facebook. 7amleh's ads were designed to test the approval process and see whether Meta's ability to automatically screen violent and racist incitement had gotten better, even with unambiguous examples of violent incitement.

"We knew from the example of what happened to the Rohingya in Myanmar that Meta has a track record of not doing enough to protect marginalized communities," Nashif said, "and that their ads manager system was particularly vulnerable."

Meta's appears to have failed 7amleh's test.

The company's Community Standards rulebook — which ads are supposed to comply with to be approved — prohibit not just text advocating for violence, but also any dehumanizing statements against people based on their race, ethnicity, religion, or nationality. Despite this, confirmation emails shared with The Intercept show Facebook approved every single ad.

Though 7amleh told The Intercept the organization had no intention to actually run these ads and was going to pull them before they were scheduled to appear, it believes their approval demonstrates the social platform remains fundamentally myopic around non-English speech — languages used by a great majority of its over 4 billion users. (Meta retroactively rejected 7amleh's Hebrew ads after The Intercept brought them to the company's attention, but the Arabic versions remain approved within Facebook's ad system.)

Facebook spokesperson Erin McPike confirmed the ads had been approved accidentally.

> "Despite our ongoing 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," she said. "That's why ads can be reviewed multiple times, including once they go live."

Just days after its own experimental ads were approved, 7amleh discovered an Arabic ad run by a group calling itself "Migrate Now" calling on "Arabs in Judea and Sumaria" — the name Israelis, particularly settlers, use to refer to the occupied Palestinian West Bank — to relocate to Jordan.

According to Facebook documentationOpens in a new tab, automated, software-based screening is the "primary method" used to approve or deny ads. But it's unclear if the "hostile speech" algorithms used to detect violent or racist posts are also used in the ad approval process. In its official response to last year's audit, Facebook said its new Hebrew-language classifier would "significantly improve" its ability to handle "major spikes in violating content," such as around flare-ups of conflict between Israel and Palestine. Based on 7amleh's experiment, however, this classifier either doesn't work very well or is for some reason not being used to screen advertisements. (McPike did not answer when asked if the approval of 7amleh's ads reflected an underlying issue with the hostile speech classifier.)

Either way, according to Nashif, the fact that these ads were approved points to an overall problem: Meta claims it can effectively use machine learning to deter explicit incitement to violence, while it clearly cannot.

"We know that Meta's Hebrew classifiers are not operating effectively, and we have not

seen the company respond to almost any of our concerns," Nashif said in his statement. "Due to this lack of action, we feel that Meta may hold at least partial responsibility for some of the harm and violence Palestinians are suffering on the ground."

The approval of the Arabic versions of the ads come as a particular surprise following a recent report by the Wall Street JournalOpens in a new tab that Meta had lowered the level of certainty its algorithmic censorship system needed to remove Arabic posts — from 80 percent confidence that the post broke the rules, to just 25 percent. In other words, Meta was less sure that the Arabic posts it was suppressing or deleting actually contained policy violations.

Nashif said, "There have been sustained actions resulting in the silencing of Palestinian voices."

*

Note to readers: Please click the share button above. Follow us on Instagram and Twitter and subscribe to our Telegram Channel. Feel free to repost and share widely Global Research articles.

Featured image is from Legal Loop
