The Washington Times

Facebook's boast of flagging 99% of terrorist content obscures as much as it reveals, experts say

James Varney

August 14, 2018

It wasn't a lie, experts believe, but it was a damned statistic nonetheless.

Facebook inventor and CEO Mark Zuckerberg boasted that his company was able to flag "99 percent of the ISIS and al Qaeda content … before any human sees it" when he testified before Congress in April.

Facebook, like other major internet platforms such as YouTube and Twitter, is under increasing pressure to scrub its platforms of violence, extremism and the more nebulous concept of "hate speech."

But as the debate about social media's responsibility and the First Amendment rages, Mr. Zuckerberg's comment obscures as much as it reveals, experts said.

"He's saying that of the content they ultimately take down, 99 percent is caught by their algorithms rather than being flagged by human beings," said Julian Sanchez, a senior fellow at the Cato Institute. "Now, I assume that this is technically true. But that's not the same as saying their algorithms are so good they accurately identify ISIS and al Qaeda content, and I think the focus on those two groups — as opposed to jihadist or 'violent extremist' content — generally should make us question how meaningful the figure is."

There's no doubt Facebook wants it to look meaningful.

"That's a carefully phrased statement, and that's the figure they've been saying all over the world since the beginning of the year," said Faiza Patel, a co-director of the Brennan Center for Justice's Liberty and National Security Program. "But given the universe of content, and what we don't know about it, that doesn't tell you much."

Social media companies say they have made progress in tackling the threats from jihadis who use online platforms to brag about their doings, to recruit members and to encourage lone-wolf attacks.

Facebook has gone the furthest of any platform, touting the 99 percent success rate. That stems from the 1.9 million pieces of content removed in one three-month period this year — 99 percent of which was flagged and removed by artificial intelligence algorithms, Facebook says.

The company did not respond to questions posed by The Washington Times, pointing instead to various policy statements it has issued in recent months.

Experts said that since Facebook's information is privately held, there is no way to know for sure.

In the past, more specific figures on exactly how much content was being scrubbed from platforms could be gleaned by reading white papers that the companies put online or in trade publications, but that hasn't been true since around 2013, said Sarah Roberts, an assistant professor of information studies at UCLA.

"I think that's the thing that is key and it's unknown," she said. "Part of this is the companies like to keep that hidden under the aegis of a trade secret, but the fact is everybody is not clear, not just with Facebook but with all the companies."

"I agree it doesn't tell you as much as it appears," said Ryan Radia, a research fellow with the Competitive Enterprise Institute. "I mean, what is the sample size? For all we know, half of it could still be appearing and we're only talking about that percentage of what's removed that artificial intelligence intercepts. But I suspect it's accurate and he didn't just make it up. That's not very likely."

The building storm over social media has made Facebook more transparent, Ms. Patel said, noting that the company released a "transparency report" this year. In it, Facebook said it had taken down 1.9 million terrorist-related pieces of content in the first quarter of the year.

Of that total, 600,000 pieces were "old content they just discovered in that quarter," said Christopher Meserole, a fellow at the Center for Middle East Policy at the Brookings Institution.

Mr. Meserole also cited another concern about the widely touted statistic: how one defines objectionable content.

"Imagine if a medical company boasted that their tests flagged cancer in 100 percent of patients that ended up having cancer," he said. "That's not a very helpful statistic, because if you flag 100 percent of all patients as potentially having cancer, then you'll get a 100 percent 'success' rate, too. To gauge how effective the test is, you would also want to know how often it predicted cancer in patients that didn't have it. The same thing is true of Facebook's 99 percent figure — it doesn't tell you anything about all the times Facebook's AI flagged content that wasn't actually problematic, so it doesn't tell you much about how effective the AI actually is."

Facebook, in a report in April, defined a terrorist organization as "any non-governmental organization that engages in premeditated acts of violence against persons or property to intimidate a civilian population, government, or international organizations in order to achieve a political, religious or ideological aim."

That definition diminishes the picture Mr. Zuckerberg's testimony provided, since Islamic State and al Qaeda are the most infamous terrorist groups but aren't alone.

“The human skill that can’t be taught to the algorithm is processing the semantic content of a post without those markers to know if it’s substantively terrorist advocacy,” Mr. Sanchez said. “By framing this answer in terms of affiliation, Zuckerberg is basically just saying that machines are better at the sort of thing we already knew machines were better at.”

Ms. Roberts also said many “terrorist” posts are just recirculating items already in the public domain, such as the infamous video of Wall Street Journal correspondent Daniel Pearl’s beheading.

What’s more, Mr. Zuckerberg carefully referred only to the terrorism genre of material.

When it comes to weeding out a broader category of “hate speech,” Facebook is much worse, with just a 38 percent success rate for the automatic algorithms. That means most of the 2.5 million pieces of information flagged from January to March came from user reports.

By contrast, the algorithm was responsible for 96 percent of the 21 million pieces of nudity or sexually explicit posts removed during the same time.

Even there, however, problems can arise. Breastfeeding photos, for instance, have created headaches for social media, and in one much-publicized instance Facebook initially removed posts that featured Nick Ut’s famous photo from the Vietnam War of a village girl running naked after a napalm attack by the South Vietnamese air force.

Posts such as grisly videos of Islamic State beheadings and other forms of execution are easy calls for either artificial intelligence or people. It’s when nuance gets involved that artificial intelligence has more problems.

Mr. Zuckerberg acknowledged that limitation in his testimony.

“Some problems lend themselves more easily to AI solutions than others,” he said. “So hate speech is one of the hardest, because determining if something is hate speech is very linguistically nuanced, right? It’s — you need to understand, you know, what is a slur and what — whether something is hateful not just in English, but the majority of people on Facebook use it in languages that are different across the world.”

Even among English speakers, words that might constitute a humorous insult in, say, England and Australia, could be considered quite offensive in the U.S., Mr. Radia said.

Ms. Patel said the debate is moving quickly amid a rush of public pressure, and some concerns aren’t being thought through enough.

“Things like ‘hate speech’ are slippery concepts and incredibly hard to define,” she said. “The evidence is anecdotal so far, but we’ve seen it happen, and I’m worried they will sweep too broadly and too much political discourse will get swept up.”