Facebook says AI getting better at spying unwanted content

November 14 2019, by Glenn Chapman

Facebook has been investing heavily in artificial intelligence to automatically spot banned content

Facebook on Wednesday said that its software is getting more skilled at spying banned content at the social network, then working with humans

to quickly remove terrorist videos and more.

"While we err on the side of free expression, we generally draw the line at anything that could result in real harm," Facebook chief executive Mark Zuckerberg said during a briefing on the company's latest report on ferreting out posts that violate its policies.

"This is a tiny fraction of the content on Facebook and Instagram, and we remove much of it before anyone sees it."

Facebook has been investing heavily in artificial intelligence (AI) to automatically spot banned content, often before it is seen by users, and human teams of reviewers who check whether the software was on target.

Facebook has more than 35,000 people working on safety and security, and spends billions of dollars annually on that mission, according to Zuckerberg.

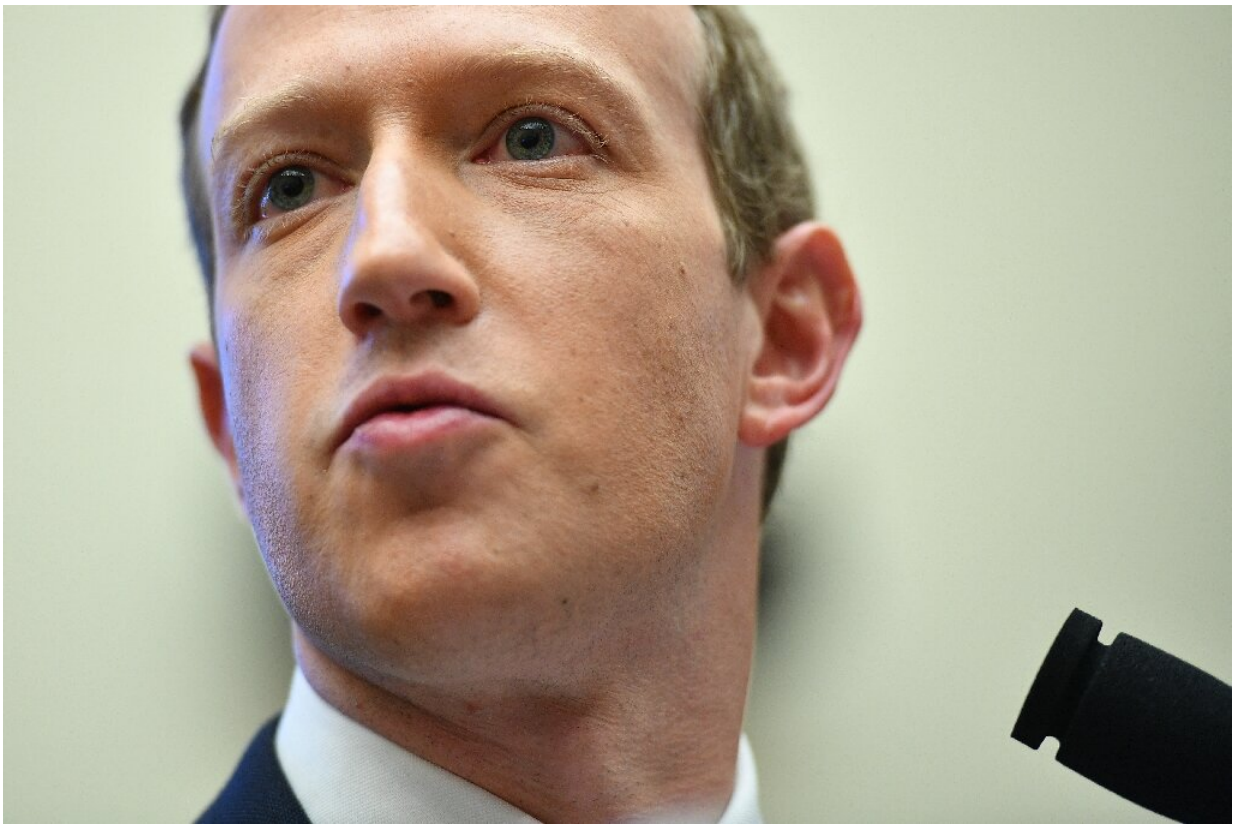
"Our efforts are paying off," Zuckerberg said. "Systems we built for addressing these issues are more advanced."

When it comes to detecting hate speech, Facebook software now automatically finds 80 percent of the content removed in a massive improvement from two years ago, when nearly all such material was not dealt with until being reported by users, according to the California-based firm.

Nettling nuance

Zuckerberg noted that hate speech is tougher for AI to detect than nudity in images or video because of "linguistic nuances" that require context that could make even common words menacing.

Add to that videos of attacks driven by bias against a race, gender or religion could be shared to condemn such violence rather than glorify it.



Facebook Chairman and CEO Mark Zuckerberg testifies before the House Financial Services Committee on October 23, 2019

People at Facebook continue to try to share video of a horrific mosque attacks in Christchurch, New Zealand, with social network systems blocking 95 percent of those attempts, according to executives.

A lone gunman opened fire on two mosques in the city of Christchurch killing and wounding scores of Muslims in March, broadcasting the assaults live on Facebook.

Facebook has terrorism experts as part of a team of more than 350 people devoted to preventing terrorist groups from using the social network, according to head of global policy management Monika Bickert.

Systems honed to find and delete banned content on Facebook are being used also at Instagram, which has ramped up efforts to filter out content likely to encourage suicide or self harm while allowing people to share such traumatic experiences in the spirit of recovery or healing,

Rules put in place recently at Instagram added drawings and other fictional content to the list to a ban on images that might encourage suicide or self harm.

The Facebook-owned image and video sharing service early this year clamped down on images of self-injury after a British teen who went online to read about suicide took her own life in 2017.

The 14-year-old's social media history revealed that she followed accounts about depression and suicide.

The case sparked a vigorous debate in Britain about parental control and state regulation of children's social media use.

Instagram has never allowed posts that promote or encourage suicide or self-harm.

With a rule change early this year, Instagram began removing references to non-graphic content related to people hurting themselves from its searches and recommendation features.

The measures were meant to make such images more difficult to find for depressed teens who might have suicidal tendencies.