Meta says generative AI deception held in check—for now

May 30 2024

Meta says Russian operation called 'Doppelganger' has persisted in efforts to undermine support for Ukraine by using fake accounts on social media.

Social media giant Meta says its bid to thwart coordinated disinformation campaigns created through ever-improving generative AI

is working, despite widespread concerns.

Meta's latest study on "coordinated inauthentic behavior" on its platforms comes as fears mount that generative AI will be used to trick or confuse people in upcoming elections worldwide, notably in the United States.

"What we've seen so far is that our industry's existing defenses, including our focus on behavior rather than content in countering adversarial threats, already apply and appear to be effective," said David Agranovich, Meta's threat disruption policy director, in a press briefing Wednesday.

"We're not seeing generative AI being used in terribly sophisticated ways, but we know that these networks are going to keep evolving their tactics as this technology changes."

Facebook has been accused for years of being used as a powerful platform for election disinformation.

Russian operatives used Facebook and other US-based social media to stir political tensions in the 2016 election won by Donald Trump. The European Union is currently investigating Meta's Facebook and Instagram over alleged failure to counter disinformation ahead of June EU elections.

But experts now also fear an unprecedented deluge of disinformation from bad actors on Meta apps because of the ease of using generative AI tools such as ChatGPT or the Dall-E image generator to make content on demand and in seconds.

Meta said it had seen "threat actors" put AI to work to create bogus photos, videos, and text, but no realistic imagery of politicians,

according to the report.

Generative AI has been used to make profile pictures for false accounts in Meta's family of apps, and a deception network from China apparently used the technology to create posters for a fictitious pro-Sikh activist movement called Operation K, the report indicated.

Meanwhile, an Israel-based network posted what appeared to be AI-generated comments about Middle Eastern politics on Facebook pages of media organizations and public figures, Meta reported.

Comparing them to spam, Meta said those comments, some of which were on pages of US lawmakers, were criticized in responses posted by real users, who called them propaganda.

Meta attributed the campaign to a Tel Aviv-based political marketing firm.

"This is an exciting space to watch," said Mike Dvilyanski, Meta's head of threat investigations. "So far, we haven't seen a disruptive use of generative AI tooling by adversaries."

The report also showed that efforts by a Russia-linked group called "Doppelganger" to use Meta apps to undermine support for Ukraine persisted but are being thwarted on the platform.

"Doppelganger has taken it to a new level over the last 20 months while remaining crude and largely ineffective in building authentic audiences on social media," according to Meta.

Meta also removed small clusters of inauthentic Facebook and Instagram accounts that originated in China and aimed at the Sikh community in Australia, Canada, India, Pakistan, and other countries, the report

showed.

Posts on those fake accounts called for pro-Sikh protests.