Meta to start labeling AI-generated content in May

April 5 2024, by Alex PIGMAN

Meta's new "Made with AI" labels will identify content created or altered with AI, including video, audio, and images.

Facebook and Instagram giant Meta on Friday said it will begin labeling AI-generated media beginning in May, as it tries to reassure users and

governments over the risks of deepfakes.

The social media juggernaut added that it will no longer remove manipulated images and audio that don't otherwise break its rules, relying instead on labeling and contextualization, so as to not infringe on freedom of speech.

The changes come as a response to criticism from the tech giant's oversight board, which independently reviews Meta's content moderation decisions.

The board in February requested that Meta urgently overhaul its approach to manipulated media given the huge advances in AI and the ease of manipulating media into highly convincing deepfakes.

The board's warning came amid fears of rampant misuse of artificial intelligence-powered applications for disinformation on platforms in a pivotal election year not only in the United States but worldwide.

Meta's new "Made with AI" labels will identify content created or altered with AI, including video, audio, and images. Additionally, a more prominent label will be used for content deemed at high risk of misleading the public.

"We agree that providing transparency and additional context is now the better way to address this content," Monika Bickert, Meta's Vice President of Content Policy, said in a blog post.

"The labels will cover a broader range of content in addition to the manipulated content that the Oversight Board recommended labeling," she added.

These new labeling techniques are linked to an agreement made in

February among major tech giants and AI players to cooperate on ways to crack down on manipulated content intended to deceive voters.

Meta, Google and OpenAI had already agreed to use a common watermarking standard that would invisibly tag images generated by their AI applications.

Identifying AI content "is better than nothing, but there are bound to be holes," Nicolas Gaudemet, AI Director at Onepoint, told AFP.

He took the example of some open source software, which doesn't always use this type of watermarking adopted by AI's big players.

Biden deepfakes

Meta said its rollout will occur in two phases with AI-generated content labeling beginning in May 2024, while the removal of manipulated media solely based on the old policy will cease in July.

According to the new standard, content, even if manipulated with AI, will remain on the platform unless it violates other rules, such as those prohibiting hate speech or voter interference.

Recent examples of convincing AI deepfakes have only heightened worries about the easily accessible technology.

The board's list of requests was part of its review of Meta's decision to leave a manipulated video of US President Joe Biden online last year.

The video showed Biden voting with his adult granddaughter, but was manipulated to falsely appear that he inappropriately touched her chest.

In a separate incident not linked to Meta, a robocall impersonation of

Biden pushed out to tens of thousands of voters urged people to not cast ballots in the New Hampshire primary.

In Pakistan, the party of former prime minister Imran Khan has used AI to generate speeches from their jailed leader.