Facebook bans deepfakes in fight against online manipulation

Nation Jan 7, 2020 3:35 PM EST

LONDON (AP) — Facebook says it is banning “deepfake” videos, the false but realistic clips created with artificial intelligence and sophisticated tools, as it steps up efforts to fight online manipulation. But the policy leaves plenty of loopholes.

The social network said late Monday that it’s beefing up its policies to remove videos edited or synthesized in ways that aren’t apparent to the average person, and which could dupe someone into thinking the video’s subject said something he or she didn’t actually say.

Created by artificial intelligence or machine learning, deepfakes combine or replace content to create images that can be almost impossible to tell are not authentic.

“While these videos are still rare on the internet, they present a significant challenge for our industry and society as their use increases,” Facebook’s vice president of global policy management, Monika Bickert, said in a blog post.

However, she said the new rules won’t include parody or satire, or clips edited just to change the order of words. The exceptions underscore the balancing act Facebook and other social media services face in their struggle to stop the spread of online misinformation and “fake news” while also respecting free speech and fending off allegations of censorship.

The U.S. tech company has been grappling with how to handle the rise of deepfakes after facing criticism last year for refusing to remove a doctored video of House Speaker Nancy Pelosi slurring her words, which was viewed more than 3 million times. Experts said the crudely edited clip was more of a “cheap fake” than a deepfake.

Then, a pair of artists posted fake footage of Facebook CEO Mark Zuckerberg showing him gloating over his one-man domination of the world. Facebook also left that clip online. The company said at the time that neither video violated its policies.

The problem of altered videos is taking on increasing urgency as experts and lawmakers try to figure out how to prevent deepfakes from being used to interfere with U.S. presidential elections in November.

The new policy is a “strong starting point,” but doesn’t address broader problems, said Sam Gregory, program director at Witness, a nonprofit working on using video technology for human rights.

“The reality is there aren’t that many political deepfakes at the moment. They’re mainly nonconsensual sexual images,” and the bigger problem is videos that are either shown without context or lightly edited, which some have dubbed “shallow fakes,” he said. These include the Pelosi clip or one that made the rounds last week of Democratic presidential candidate Joe Biden that was selectively edited to falsely suggest he made racist remarks.

Gregory, whose group was among those that gave feedback to Facebook for the policy, said that while the new rules looked strong on paper, there are questions around how effective the company will be at uncovering synthetic videos.

Facebook has built deepfake-detecting algorithms and can also look at an account’s behavior to get an idea of whether it’s intention is to spread disinformation, so the company will have an edge over users or journalists at sniffing them out, Gregory said.

But those algorithms haven’t been used widely for deepfakes “in the wild. So it is an open question how effective detection will be,” he said. “This is an algorithmic kind of game of cat and mouse, where the forgeries will get better alongside the detection.”

Facebook said any videos, deepfake or not, will also be removed if they violate existing standards for nudity, graphic violence or hate speech. Those that aren't removed can still be reviewed by independent third-party fact-checkers and any deemed false will be flagged as such to people trying to share or view them, which Bickert said was a better approach than just taking them down.

"If we simply removed all manipulated videos flagged by fact-checkers as false, the videos would still be available elsewhere on the internet or social media ecosystem," Bickert said. "By leaving them up and labelling them as false, we're providing people with important information and context."

Twitter, which has been another hotbed for misinformation and altered videos, said it's "in the process" of creating a policy for "synthetic and manipulated media," which would include deepfakes and other doctored videos. The company has asked for public feedback on the issue. The responses it's considering include putting a notice next to tweets that include manipulated material. The tweets might also be removed if they're misleading and could cause serious harm to someone.

YouTube, meanwhile, has a policy against "deceptive practices" but does not appear to specifically ban deepfakes. Google did not immediately respond to a message for comment on Tuesday.

By — **Kelvin Chan, Associated Press**