Meta pushes to label all AI images on Instagram and Facebook in crackdown on deceptive content

Publication Date: 2024-02-06

Author: Josh Taylor

Section: Technology

Tags: Meta, Artificial intelligence (AI), Social media, Instagram, Facebook, news

Article URL: https://www.theguardian.com/technology/2024/feb/06/meta-ai-images-instagram-facebook-label-crackdown

Meta is working to detect and label AI-generated images on Facebook, Instagram and Threads as the company pushes to call out “people and organisations that actively want to deceive people”. Photorealistic images created using Meta’s AI imaging tool are already labelled as AI, but the company’s president of global affairs, Nick Clegg, announced in a blog post on Tuesday that the company would work to begin labelling AI-generated images developed on rival services. Meta’s AI images already contain metadata and invisible watermarks that can tell other organisations that the image was developed by AI, and the company is developing tools to identify these types of markers when used by other companies, such as Google, OpenAI, Microsoft, Adobe, Midjourney and Shutterstock in their AI image generators, Clegg said. “As the difference between human and synthetic content gets blurred, people want to know where the boundary lies,” Clegg said. “People are often coming across AI-generated content for the first time and our users have told us they appreciate transparency around this new technology. So it’s important that we help people know when photorealistic content they’re seeing has been created using AI.” Clegg said the capability was being built and the labels would be applied in all languages in the coming months. “We’re taking this approach through the next year, during which a number of important elections are taking place around the world,” Clegg said. Clegg noted it was limited to images and AI tools that generate audio and video do not currently include these markers, but the company would allow people to disclose and add labels to this content when posted online. He said the company would also place a more prominent label on “digitally created or altered” images, video or audio that “creates a particularly high risk of materially deceiving the public on a matter of importance”. The company was also looking at developing technology to automatically detect AI-generated content, even if the content does not have the invisible markers, or where those markers have been removed. “This work is especially important as this is likely to become an increasingly adversarial space in the years ahead,” Clegg said. “People and organisations that actively want to deceive people with AI-generated content will look for ways around safeguards that are put in place to detect it. Across our industry and society more generally, we’ll need to keep looking for ways to stay one step ahead.” AI deepfakes have already entered the US presidential election cycle, with robocalls of what is believed to have been an AI-generated deepfake of US president Joe Biden’s voice discouraging voters from attending the Democratic primary in New Hampshire. Nine News in Australia last week also faced criticism for altering an image of the Victorian Animal Justice party MP Georgie Purcell to expose her midriff and alter her chest in an image broadcast in the evening news. The network blamed “automation” in Adobe’s Photoshop product, which features AI image tools.