TikTok to auto-flag AI videos – even if created on other platforms

Publication Date: 2024-05-09

Author: Alex Hern

Section: Technology

Tags: TikTok, Artificial intelligence (AI), Digital media, Deepfake, Computing, Technology sector, Social media, news

Article URL: https://www.theguardian.com/technology/article/2024/may/09/tiktok-auto-flag-ai-videos-digital-watermarking

TikTok will flag users who upload artificial intelligence-generated content (AIGC) to the video-sharing site from other platforms, the company says, becoming the first big video site to automatically label such content for users to see. Content created using TikTok's own AI tools is already automatically marked as such to viewers, and the company has required creators to manually add the same labels to their own content, but until now they have been able to evade the rules and pass off generated material as authentic by uploading it from other platforms. Now, the company will begin using digital watermarks created by the cross-industry group Coalition for Content Provenance and Authenticity (C2PA) to identify and label as much AIGC as it can. "AI enables incredible creative opportunities but can confuse or mislead viewers if they don't know content was AI-generated," said Adam Presser, the head of operations and trust and safety at TikTok. "Labelling helps make that context clear – which is why we label AIGC made with TikTok AI effects, and have required creators to label realistic AIGC for over a year." The labelling goes both ways: TikTok will also begin to apply the same digital watermarking technology, called Content Credentials, to content downloaded from its own platform, which will let other platforms identify "when, where and how the content was made or edited", Presser said. But the ability to label generated content is limited to that created by other platforms that are also members of C2PA. That includes most major players in AI, such as Microsoft, Google and Adobe. Until this week, it did not count OpenAI among its membership, but the research lab joined the steering committeeon Tuesday. It had already started to use the Content Credentials technology earlier this year, and plans to include it in its video-creation AI, Sora, once it is released to the public. But smaller, less scrupulous and less commercially focused AI groups will still continue to produce unlabelled content for some time. Open source tools such as Stable Diffusion, which have the underlying code free to download, can always be tweaked to remove any attempts to label images (although Stability.AI, one of the creators of Stable Diffusion and the current developer of the tool, is a member of the group). The AI startup Midjourney, one of the most popular image generating tools, is entirely absent from the list of members. TikTok's labelling plans follow Meta, which made a similar announcement in February. But the labelling attempts have done little to slow the huge proliferation of AI-generated imagery on Facebook in particular, leading some to call the site the "zombie internet". Other social media apps have been slower off the mark: Snapchat labels AI-generated content created using its own tools, but warns users that "images created with non-Snap products may not be labelled as AI-generated", while Elon Musk's X, previously known as Twitter, has no automatic labelling system at all, instead relying on its user-submitted "community notes" to flag fake imagery.