Taylor Swift's deepfake 'nudes': a wake-up call to the dangers of AI

The social media platform X has stopped people from searching for Taylor Swift due to explicit AI-generated pictures of the singer circulating on the site.

Image by Alexandra_Koch from Pixabay

According to X's head of business operations, Joe Benarroch, this is a temporary measure to prioritise the singer's safety. Since this weekend, when users search for Swift on the site, they receive a message saying, "Something went wrong. Try reloading."

The fake images of the singer gained widespread attention earlier recently, going viral and being viewed millions of times. This caused concern among US officials and the singer's fans.

AI's new low: Taylor Swift targeted with explicit deepfake images

Karabo Ledwaba 26 Jan 2024

Because of this, X, previously known as Twitter, released a statement on Friday, 26 January 2024, stating that they strictly prohibit posting non-consensual nudity on the platform. The statement also mentioned that they have a zero-tolerance policy towards such content, and their teams are actively removing all identified images and taking appropriate actions against the accounts responsible for posting them.

According to Anna Collard, SVP of content strategy and evangelist at KnowBe4 Africa, advancements in AI and technologies make it difficult to tell if a human or a machine did something.

While these advancements have great potential for good, there are also significant risks, especially in an election year.

“Apart from abusing these platforms such as in the case of creating deepfake pornographic images of Taylor Swift, these tools can also increase the effectiveness of phishing and business email compromise (BEC) attacks, when used to impersonate people we know. These deepfake platforms can create civil and societal unrest when used to spread mis- or disinformation in political and election campaigns, and is a dangerous element in modern digital society.”

“This is cause for concern and asks for more awareness and understanding among the public and policymakers, especially now with important elections coming up in South Africa and the USA. As an example of this being in use already, between 8 December 2023 and 8 January 2024, 100+ deepfake video advertisements were identified impersonating the British prime minister, Rishi Sunak, on Meta, many of which elicited emotional responses. The potential of deepfakes driving disinformation to disrupt democratic processes, tarnish reputations, and incite public unrest cannot be underestimated.”

SABC investigates deepfake videos impersonating journalists

15 Nov 2023

“In a recent survey undertaken by KnowBe4, across 800 employees aged 18-54 in Mauritius, Egypt, Botswana, South Africa and Kenya, 74% of respondents said that they had believed a communication via email or direct message, or a photo or video, was true when, in fact, it was a deepfake. Considering how deepfake technology uses both machine learning and AI to manipulate data and imagery using real-world images and information, it is easy to see how they were tricked. The problem is, awareness of deepfakes and how they work is very low in Africa and this puts users at risk.”

“It is crucial that we have more education and awareness training. These are the only tools that will help users understand the risks and recognise the red flags when it comes to fake photo and video content. They should also be trained to understand that they should not believe everything they see and should not act on any unusual instructions without first confirming they are legitimate.”

The full results from the recent KnowBe4 survey on cybersecurity awareness in Africa can be read here.