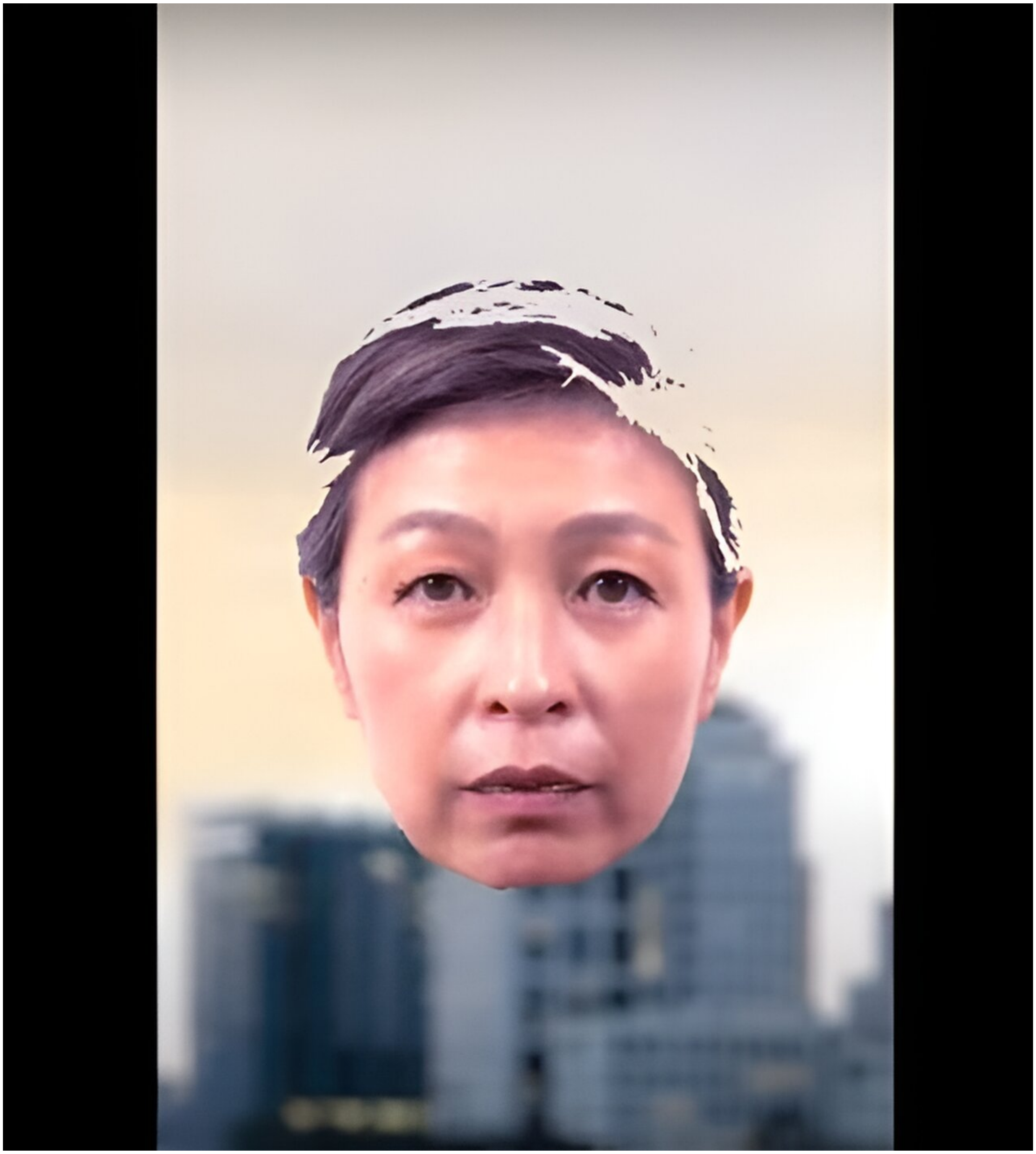
Detecting AI-manipulated content is a challenging arms race

March 11 2024, by Magnus Stenaa Jensen



A group of DTU students developed a deepfake model that allowed South Korean artist Haegue Yang to speak Danish at an exhibition at SMK (Statens Museum for Kunst). Credit: DTU

Nude photos of pop superstar Taylor Swift recently flooded social media X, where they were viewed and shared by millions of users. But the images weren't realâ&#128;"they were deepfakes created using artificial intelligence (AI). The incident fanned the debate about regulating deepfakes, and drew so much attention that the White House got involved.

Solutions are now being discussed among politicians and big tech companies. But is it even possible to guard against this kind of attack? According to Morten MÃ¸rup, who does research in artificial intelligence at DTU Compute, it could be a very difficult task.

Deepfake uses deep learning to create text, images, speech, or video that are presented as being real but are far from it. The development of deepfakes took off in 2014 when the Generative Adversarial Network (GAN) AI learning principle was developed.

The principle allows AI models to train against other AI models designed to detect deepfakes. Since then, AI models have been developed that make it even harder to tell the difference between fake and reality, and there are many deepfake tools available on the Internet.

"The GAN learning principle is based on an arms race between two AI models: one that generates deepfakes, and another that tries to distinguish between what is real and AI-created. The deepfake models test themselves against models designed to detect them. If the AI model designed to produce deepfakes is trained against another AI model designed to determine whether, say, an image is real or fake, it will learn from this AI model how it needs to improve.

"This race between two AI modelsâ&#128;"one trying to generate and the other to detect fake materialâ&#128;"will continue until the detection model can no longer distinguish between reality and fake. This is what makes it so difficult for both people and AI models to tell the difference between what is real and fake," says MÃ¸rup.

What are deepfakes?

Deepfakes are computer-generated text, images, sounds, speech, or video that present as being real, but are not. It can be very difficult for a human to discern whether something is real or AI-generatedâ&#128;"in some cases almost impossible. Deepfakes are often generated based on the GAN learning principle, but many other methods for generating deepfakes also exist today.

Tech companies and researchers are working to develop AI models that can detect deepfakes, but even these models can have difficulty knowing the difference.

See if you can tell the difference. (This website from 2019 is based on GAN, and even better deepfake tools have since been developed.)

Declarations are no guarantee

A study by DR (in Danish) found that 1 in 3 children aged 9â&#128;"14 never considers the possibility that photos and videos on social media could be manipulated. The many deepfakes have now led Meta, which owns Facebook and Instagram, to make an effort to detect images and videos that are computer-generated.

At the same time, the EU's AI Actâ&#128;"the world's first AI legislationâ&#128;"will make it mandatory to declare computer-generated content. However, even though regulation is being drafted in this area and tech companies will seek to detect deepfakes, MÃ¸rup believes there is no guarantee that we can avoid seeing much more deepfake content in the future.

"Starting to declare deepfakes is definitely an important step, but there will still be people who can generate content without it being declared. Research is being done on developing AI-based deepfake detectors, but then we are back to the arms race. And it's an arms race that I think will be very hard to win. We must therefore not turn a blind eye and just assume that anything that is not declared as deepfake is real," says MÃ¸rup.

There is currently another method for detecting deepfakes that completely bypasses AI and is based on thorough research.

"You can try to check a deepfake against other information. If a video clip shows something that happened in Ukraine, for example, you can compare it with satellite photos and weather information at the time to see if everything matches the video clip. For example, was it raining that day, yet the video clip shows a cloudless sky?

"The only problem is that the AI models can potentially also have access to the information we are checking the video against. So a good deepfake will ensure that it is raining in the video," says MÃ¸rup.

A world of misinformation

In 2019, the CEO of a British energy company received a call from what he believed to be his superior in the parent company in Germany, telling him to transfer EUR 220,000 to a bank account. In reality, the CEO had been tricked by a deepfake. A con man had used AI to generate his superior's voice so convincingly that the director transferred the money immediately. In February 2024, a large company in Hong Kong experienced a similar incident and was defrauded USD 25.6 million.

In Denmark, the Ministry of Foreign Affairs will more closely monitor diplomatic video conversations after the Foreign Minister, Lars LÃ¸kke Rasmussen (M), experienced a deepfake call last year from a group of Russian comedians who had faked the face and voice of Moussa Faki, commission chairman of the African Union.

While it can be very difficult to prevent similar scams and the spread of deepfake-generated misinformation, MÃ¸rup believes that greater awareness of the issues is key to limiting the problem.

"Declaration requirements will make it harder for regular users to make deepfakes without being detected, but there will continue to be major players out there who will defraud others or influence democratic processes. We therefore need to recognize that these technologies exist and act accordingly.

"We need to practice source criticism and understand that we live in a world of misinformation, where manipulation exists that can be very difficult to detect. As a society, our common understanding of what is real can be threatened. It will be a big problem if we start to reject truths as misinformation and fake because they don't mesh with our worldview," says MÃ¸rup.

The images of Taylor Swift were subsequently deleted, and searches using the singer's name were disabled for a period on X to prevent new images being shared. Since then, several U.S. politicians, including Congresswoman Yvette Clarke (Democrat), have called for legislation to ban the creation and sharing of deepfakes as pornographic content on the internet.

Provided by Technical University of Denmark

Citation: Detecting AI-manipulated content is a challenging arms race (2024, March 11) retrieved 28 April 2024 from https://techxplore.com/news/2024-03-ai-content-arms.html