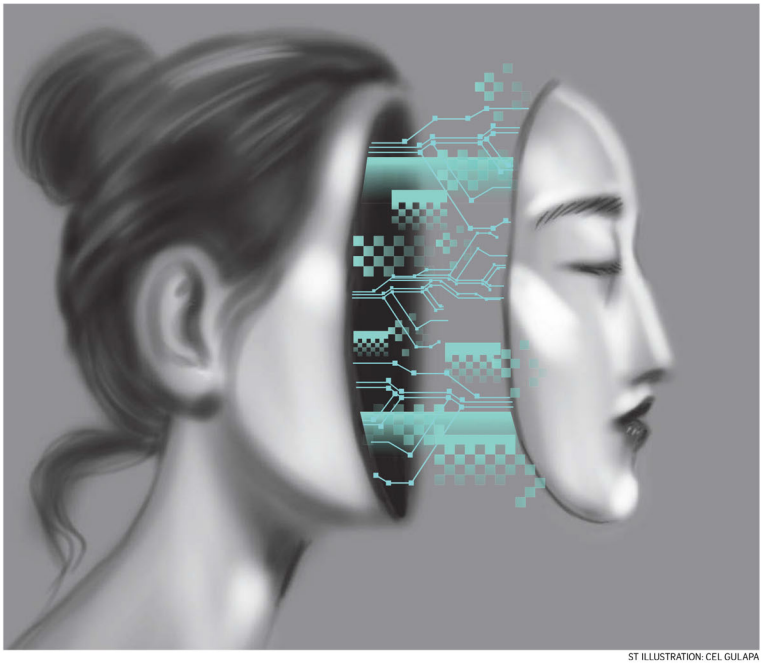
NUS National University of Singapore

Source: *The Straits Times*, pB22
Date: 9 August 2022

Growing ease of access to deepfake tools a worry, say experts



ST ILLUSTRATION: CEL GULAPA

Technology at stage where images created can likely fool big groups and soon, the trained eye

Rei Kurohi
Tech Correspondent

Last year, Spanish TV viewers were treated to never-before-seen footage of their beloved late folk singer Lola Flores encouraging them to be proud of their accents and cultural roots.

What looked like archival clips of Flores, who died in 1995, were used in a new advertising campaign called Con Mucho Acento (Heavily Accented), produced by public relations firm Ogilvy for beer brewery Cruzcampo.

Though the uplifting and empowering message resonated with many viewers, the ad generated much controversy because Ms Flores had never actually been recorded saying the words in the video.

Instead, the singer had been digitally resurrected using deepfake technology.

The tools used to create deepfakes – videos of a person whose face or body has been altered using artificial intelligence (AI) to resemble someone else's – have become more powerful in recent years.

They have also been made more easily available online since 2020, causing concern among experts who study the role of technology in misinformation.

For instance, deepfake creators and enthusiasts are sharing their pre-trained AI models on online forums, making it easier than ever for users without much technical know-how to download one and apply it to their own videos in mere hours.

Without a pre-trained model, it could take days or even weeks of processing time to train one from scratch and generate a decently convincing fake face.

The technology has already reached the stage where a skilled deepfake maker can create images that are highly convincing and likely to fool large groups of people, said AI expert Terence Sim of the National University of Singapore's (NUS) School of Computing.

Associate Professor Sim, who studies deepfakes and other kinds of digitally altered images at the NUS Centre for Trusted Internet and Community (CTIC), said that people let their guard down and could easily be deceived in the heat of, say, political campaigning.

"You could be in the midst of an election, where all the candidates are campaigning and certain words are being twisted deliberately, maliciously," he said.

In the near future, the technology may even be applied to manipulate other kinds of images, such as a person's full body, inanimate objects or parts of the environment, Prof Sim said.

This is because AI and machine learning algorithms used to generate deepfakes are "agnostic" about the type of media they are fed.

So future deepfake creators may be able to turn a video of a person tearing up a piece of paper or burning a pile of rubbish into one of the person destroying a sensitive artefact, which could incite strong feelings and have serious consequences.

"The barriers to entry are definitely much lower and the threat is real, so we do have to be watchful," he said.

Dr Maria Teresa Soto-Sanfiel, principal researcher at the CTIC, believes the technology might advance to the point where even trained experts could be fooled in as short as 10 years' time.

For now, one can spot potential fakes by looking out for mismatched resolution between the face and the rest of the video, or visible seams between the face and the body of the actor.

These technological advances also raise ethical questions.

Dr Soto-Sanfield cited the case of the late Spanish singer Flores. "The image was so perfect. It was so amazing how they recreated the character, the personality and the features of Lola Flores," she said.

Besides using AI to generate a deepfake face from more than 5,000 images of the real Flores, Ogilvy also created a 3D model of her face onto which the deepfake could be projected for greater realism. This was pasted over a video of an actress who performed the ad's script.

Ogilvy also used video compositing techniques to alter the face shape and hairline of the actress to better resemble Flores.

"The thing is, we don't know if Lola Flores (would have) accepted being part of this commercial selling beer. We don't know if she wanted to be recreated artificially by a machine."

Another consequence of the spread of deepfakes is that they could undermine public trust just by existing.

Dr Soto-Sanfiel pointed out that people may begin to doubt video messages from politicians, for instance, even if they are real.

This mindset has already fed into online conspiracy theories such as those surrounding Chinese athlete Peng Shuai and actor Zhang Zhehan. Both have been the target of online speculation about whether their social media photos and videos really feature them or whether they have actually "disappeared" and been replaced by deepfakes.

And altering images of historical figures like Flores may even have an impact on how people recall past events, Dr Soto-Sanfiel said.

"One thing we're investigating here at the CTIC is how deepfakes affect memory," she added.

"We don't know exactly if our brains can really distinguish something that is real from something that is not, even after being told that it's fake.

"We don't know if it is enough to inform people that something is fake to avoid them integrating that message into their perception of reality."

rei@sph.com.sg

SCAN TO WATCH Deepfakes: Behind the digital deception
https://str.sg/deep09

Rei Kurohi (left), with a photo of ST executive editor Sumiko Tan's deepfake (centre), next to one of the real journalist. The writer says that while features such as Ms Tan's eyes, nose and mouth were recreated accurately, the body and face shape were significantly different from hers, making it obvious to anyone familiar with Ms Tan that the figure in the video was not her. ST PHOTOS: PHILIP CHEONG, DESMOND FOO, KUA CHEE SIONG

How I created Sumiko Tan's deepfake overnight

Home-made deepfakes are more common than ever as tools to create them have become more powerful and more readily available.

The Straits Times attempted to create one to see just how easy it would be.

I began by picking a target, or a "person of interest", whose appearance I will attempt to recreate on my home computer. ST executive editor Sumiko Tan gamely agreed to be the person of interest for this experiment.

I started by collecting footage of Ms Tan to train the artificial intelligence (AI) model on, including episodes of her Lunch With Sumiko interview series and her 2020 General Election analysis videos.

In total, I downloaded 60 videos featuring Ms Tan, which added up to about 5½ hours' worth of footage. Using video editing software, I trimmed this down to about 15 minutes' worth of clips featuring shots of her face.

Next, I downloaded a free, open-source deepfake tool called DeepFaceLab. It is one of the most widely used tools in the creation of both professional and amateur deepfake videos. I looked up tutorials on forums and on YouTube to learn to use it.

I also downloaded a pre-trained model from a forum so I would not have to train my own AI model from scratch.

I was ready to work on the footage I had collected. Extracting 10 frames a second from the source video, I collected over 9,000 images of Ms Tan's face. DeepFaceLab then generated "masks" or facial maps from the images, based on markers such as her eyes, nose, mouth and jawline.

Next, I needed some destination videos onto which the deepfake of Ms Tan could be applied. I selected a short clip of a Parliamentary sitting in which her face was superimposed over a politician's. I also had my colleagues film me as I delivered some short statements while pretending to be Ms Tan.

DeepFaceLab then performed the same processing routine on the destination videos – extraction of frames, followed by facial mapping.

With the various face sets ready, I could then train the model to recognise both Ms Tan and the destination actors, namely the politician and myself. This was the most time-consuming step.

The program made use of the processing power of my computer's graphics card to "learn" how to map Ms Tan's face onto that of the actors.

My computer features an Nvidia RTX 3070 graphics card, as well as a Ryzen 5 5600X processor and 16GB of system memory or RAM. This is a typical modern gaming set-up which can be purchased for about $2,000.

I left the program running overnight for 12 hours, and it was able to complete about 100,000 iterations, or learning cycles, and had become capable of generating fairly realistic replicas of Ms Tan's features.

At that point, I stopped the training and proceeded to the merging stage, where I used DeepFaceLab's built-in tools to blend the deepfake "masks" onto the destination videos.

The resolution of the deepfake face was also noticeably poorer and lower quality than the rest of the image, which was shot at 4K resolution.

At full size, the mask is a square image that includes a blur background and other visual artefacts. DeepFaceLab allows the user to shrink the edges of the mask and blur out the seams to create a better blend. It also offers automatic colour matching to account for different lighting conditions and shadows.

The result was not bad, if a little uncanny. While features such as Ms Tan's eyes, nose and mouth were recreated accurately, my body and face shape were significantly different from hers, making it obvious to anyone familiar with Ms Tan that the figure in the video was not her.

The resolution of the deepfake face was also noticeably poorer and lower quality than the rest of the image, which was shot at 4K resolution.

The second video featuring the politician looked more convincing, perhaps because the politician's body and face shape more closely resemble Ms Tan's.

The original clip of the Parliament sitting was also of a lower resolution, resulting in a closer match between the deepfake face and the rest of the video.

I showed the deepfake clips to Associate Professor Terence Sim, who studies deepfakes and other kinds of digitally altered images at the National University of Singapore's Centre for Trusted Internet and Community.

While he could tell they were fake, Prof Sim said it was not a bad effort and could be convincing to viewers who are less familiar with Ms Tan and me.

"I could tell straight away that the face is of a lower resolution than the rest of the body because of the visual quality, but overall, I think if you are not expecting to see a fake, you may be fooled," he said. Prof Sim said the telltale signs generally fall into three categories: physical artefacts, semantic features and content.

Physical artefacts could include visual imperfections such as poorly blended seams between the fake and real images and flickering colours.

Semantic properties could include poorly rendered components that do not make sense such as mismatched eyes, malformed features, misalignment of the face relative to the head pose, or an expression that does not line up with the emotional content of the video.

Finally, at the content level, the viewer should ask if the person purportedly being featured is likely to say or do what they appear to be saying or doing.

"For example, if you have Steve Jobs selling Samsung phones, it throws you off, because Steve Jobs sells iPhones, not Samsungs," said Prof Sim.

"This requires a lot of more high-level, general knowledge of who this person is and what he is likely to say or not say."

But while making a deepfake may not be too difficult, creating a convincing one goes beyond just processing power. At the current state of the technology, at least, the actor still needs to do a good job of impersonation and should ideally bear some resemblance to the target to begin with.

When I showed the deepfake videos to Ms Tan, she appeared bemused.

"Is that supposed to be me? I don't really think so," she said.

She was not too worried about deepfakes for now, given the limitations of the technology.

"I guess if the technology is very advanced, then people will have to be very careful, but based on the two examples I saw, I don't think we need to worry that much for now," she added.

Rei Kurohi