Publication: The Straits Times, pg B3
Date: 11 April 2023
Headline: Are we ready for a post-trust world?

Are we ready for a post-trust world?

With generative AI becoming freely accessible, brace yourself for deepfake overload. Time to marshal truth tools, and your own critical thinking.

Lim Sun Sun

Fake images of Pope Francis in a white designer puffer coat and former US president Donald Trump's arrest are two examples of the ability of generative artificial intelligence to produce highly realistic photos. PHOTOS: MIDJOURNEY/REDDIT, WEIRD AI CREATIONS/TWITTER

A video of Ukrainian President Volodymyr Zelensky telling his soldiers to surrender to Russia? Plausible. A photo of French President Emmanuel Macron running away from pension reform protesters? Intriguing. The image of Pope Francis in a white designer puffer coat? Convincing.

Yet all three were fake, all three went viral and there are heaps more where they came from. Generative artificial intelligence (AI) is capable of fabricating endless hyper-realistic images and videos, enhanced with compelling audio produced through AI voice-cloning. As image generators such as Dall-E and Stable Diffusion become freely accessible to the masses, we must also brace ourselves for misinformation and disinformation to proliferate more than ever.

In the face of such shifting circumstances, forget about living in a post-truth world. Now that the genie of generative AI is out of the bottle, we are decidedly living in a post-trust world.

Today, it seems that even tried and tested markers of content authenticity and credibility are stretched to near capacity. With highly realistic photos produced by generative AI, we can no longer believe our eyes. With remarkably accurate AI voice-cloning, we can no longer believe our ears. With so many age-old methods of verifying truth increasingly outmoded, can we believe our minds?

HOW TO SPOT A FAKE

As countless technological innovations emerge to warp our perceptions of reality, navigating a post-trust world is decidedly more challenging. Let's look at how you could have worked out that the photos of former United States president Donald Trump's arrest – done using AI image generator Midjourney – were bogus.

Experts advise that images produced by generative AI tools have particular trademark flaws. So here, the centre of attention, in this case Trump, is usually clear and sharp, whereas people and scenery in the background are more likely to be blurry. Due to the sheer volume of their images in circulation, famous faces like Mr Macron's and the Pope's will be distinct and well defined, but not so the peripheral characters in these photos.

Finer details such as people's hands are often poorly rendered with too many fingers, or a lack of distinction between them. Their bodies also tend to be disproportionate, having perhaps oversized torsos, missing body parts or deformed limbs. The text on smaller objects such as policemen's badges is often gibberish.

Such tips, while helpful, assume that everyone who receives such faux images is actually able or inclined to scrutinise them thoroughly. However, we all know that as we scroll through the flood of notifications on our phones, and scan the posts on social media apps and messages our well-intentioned friends send us, we are not always on our guard. How could we possibly be? Our attention economy requires that we focus our energies on certain crucial content.

But a picture of Trump that resonates with current headlines, aligns with his tumultuous record and accords with our cognitive biases towards him will easily escape scrutiny. Instead of critically eyeballing the image, our propensity to take a glance, like and share is high. Such is the social engineering underpinning online falsehoods – they are convincing because they are highly plausible and designed to trigger a strong emotional response.

DETECTION INNOVATION PROVIDES HOPE

Couple such tactics with the sophistication of the latest generative AI tools and we have a potent cocktail of deception and virality. Layer on the lightning speed of AI advancements and even telltale signs that images have been falsified will soon evaporate. Content moderation by social media platforms is already struggling to keep up. Twitter, for example, did not label the fake Trump and Macron images as they went viral.

But we should not throw in the towel and resign ourselves to the daily hazards of a post-trust society where nothing is to be believed – not just yet. In the same way that the technology behind the production of convincing falsehoods is progressing swiftly, so, too, are the inventions sprouting to detect them.

There is ongoing innovation to capture the provenance of digital content so that such information is portable and tamper-proof. One method is digital watermarking that involves altering the pixels in digital images according to pre-determined patterns to embed various kinds of information about the source and authenticity.

But we should not throw in the towel and resign ourselves to the daily hazards of a post-trust society where nothing is to be believed – not just yet. In the same way that the technology behind the production of convincing falsehoods is progressing swiftly, so, too, are the inventions sprouting to detect them.

Since images comprise millions of pixels, there is room for a comprehensive range of indicators to be incorporated into the watermark, including metadata such as the date, time and location of content creation, software used, attribution and so on. The same method can also be used to brand audio and digital content, but further research is needed to ensure that these watermarks are completely resistant to manipulation despite rounds of doctoring.

Alternative methods are geared towards making the news ecosystem more robust, including through the use of blockchain technology. By design, blockchains are distributed ledgers of recorded transactions that are invulnerable to modification. Once news or information is recorded on a blockchain, it becomes tamper-resistant and allows for the verification and authentication of news content by anyone at any time.

The alteration and deletion of news content will be recorded, enhancing transparency and accountability. While blockchain technology is already well developed, the collective buy-in of diverse industry players and state actors is required to get such a repository off the ground. In other words, this solution, while visible, is far away on the horizon.

The destruction that deepfake videos and images can wreak is incalculable. Beyond the societal schisms they could trigger by fabricating controversies around polarising public figures like Trump, lawsuits for civil disputes that rest on pivotal photographic evidence will also become more contentious when the fidelity and authenticity of such evidence are called into question.

Egregiously, too, deepfakes have already inflicted significant damage on regular people, especially women whose images have been misused for pornography that features them in humiliating sexual acts without their consent.

Technology-facilitated sexual violence could escalate even more sharply with the entry of generative AI.

Research company Sensity AI has been tracking online deepfake videos since December 2018 and found that non-consensual porn constitutes 90 per cent of deepfake videos. MrDeepFakes is a leading purveyor of deepfake porn, hosting thousands of sexually explicit deepfake videos for its 17 million monthly visitors.

Clearly, the other critical components of a trustworthy media landscape are industry self-regulation and state regulation. Notably, Midjourney chief executive officer David Holz has announced plans to end free access to his firm's service due to "extraordinary demand and trial abuse" after those high-profile deepfakes.

However, in tech companies' race to reap handsome dividends from running costly generative AI services, we cannot expect them to impose restrictions that inhibit their growth. That is, unless lawmakers worldwide mandate that generative AI companies introduce methods to inviolably watermark AI-generated content, establish digital content repositories that track content provenance, develop effective guardrails against misuse and abuse, and ramp up efforts to improve the digital literacy of consumers.

Until the day arrives when we have in our palm apps that allow us to verify content in one click, we must steel ourselves for the trials of living in a post-trust world and marshal our best facilities for critical thinking and circumspection.

• Lim Sun Sun is vice-president (partnerships and engagement) and professor of communication and technology at the Singapore Management University. She is also a member of the Media Literacy Council.