Generative AI • Need for laws to protect against harms from deepfakes | B4

THE STRAITS TIMES, 18 OCT 2023, PAGE B4

Are we equipped to confront AI-generated deepfakes?

The ease of committing online harms using generative AI and the impact on victims need to be tackled urgently.

Hannah Yee-Fen Lim

As generative artificial intelligence (AI) continues to take the world by storm, from ChatGPT being able to seemingly generate human-like texts to its technology being touted as able to help defendants in court who can't afford lawyers, there's been an explosion in unsavoury uses of it, too.

Generative AI technology is already with us, and these tools are readily available on the Internet. It is essential to understand that generative AI algorithms are simply programs that can be easily created by those trained in computer science. They consist of computer code that uses mathematics and statistics, and are developed by feeding the algorithm vast amounts of scraped source materials freely available on the Internet, such as photos and videos.

The technology can be used to generate fake text, audio, images and videos that are nearly impossible to detect, and has become a harmful weapon as it is easily available to the average person and doesn't require much expertise or resources.

Now, even school-age children can create high-quality fakes that perpetrate damaging and permanent harms including misinformation, disinformation and distress to individuals.

In order for an average person to use the technology to generate fake pornographic images, for example, all that person needs to do is to provide the program an image or two of the target or victim, and these can be easily found on the Internet, especially on social media accounts.

The ease and availability with which such generative AI tools can be created, accessed and used would make it near impossible to outlaw their use. Indeed, as with all tools, technological or otherwise, generative AI can be used for both good and bad, and thus it is not feasible to restrict their availability.

DEEPFAKE PORN AND LIP-SYNCH MISINFORMATION

At an Online Harms Symposium in September, Minister for Law and Home Affairs K. Shanmugam said that further laws are needed to protect victims. The need to address online safety is made more pressing with the rise of harassment AI, Mr Shanmugam said. He referred to an example of AI-generated naked images of young girls in the Spanish town of Almendralejo, which was reported by the BBC after the images circulated on social media without the victims' knowledge.

The earliest deepfake porn appeared around 2017, when "face swops" were done on celebrities by attaching images of their faces to pornographic videos. Videos purporting to show actresses Gal Gadot, Kristin Bell and Scarlett Johansson were of such good quality that most viewers believed they did engage in those sex acts.

The victims are not just celebrities.

Noelle Martin was an ordinary 18-year-old Australian student when she discovered her face had been used in "cheapfake" pornography without her consent. Cheapfakes are fakes that utilise less sophisticated digital technology techniques, often without the use of generative AI. She reported the matter to the police, only to be told that there was not much they could do to help her.

Women are often the victims of such deepfakes. According to Sensity AI, a firm specialising in deepfake content detection, 90 per cent to 95 per cent of deepfake videos since 2018 were non-consensual pornography, and based on figures in October 2020, more than 100,000 computer-generated fake nude images of women were created without their consent or knowledge. More alarmingly, some of these nude images were apparently of underage individuals. In some cases, it is strongly believed that they were created by schoolboys.

Another emerging harm is lip-synching deepfakes generated by generative AI. These give the perpetrators the ability to make the victim appear to say things they never said. Lip-synching is a valuable tool in the film industry, enabling bloopers to be easily rectified and eliminating the need to re-shoot scenes.

However, it has magnified the ability of perpetrators to distort the truth and manipulate reality. This has significant ramifications in many contexts, at the level of individuals to the national level, such as the manipulation of political elections, where politicians can be the target of smear campaigns when their speeches, policies and actions can be manipulated and distorted.

THE NEED FOR SOLUTIONS

There are existing laws in Singapore that cover some of the acts implicated in nefarious uses of generative AI, but most don't result in redress for the victims, nor do they deter the bad actors. For starters, it may not always be possible to discover who the perpetrators are – or if they can be discovered, they may not be within the jurisdiction.

Even if the perpetrators are within the jurisdiction, the legal actions that can be taken may not be helpful to address the harm, distress and humiliation suffered by the victims. For example, victims could take legal action claiming copyright infringement for using images protected by copyright law without permission, but this would not correct the misinformation or address the harm in reputation suffered.

The Personal Data Protection Act (PDPA) unfortunately does not give victims clear protection against deepfakes and the framework is ill-suited, being more targeted towards the collection, use or disclosure of personal data without consent and, at the same time, contains sweeping exceptions.

False personal data such as deepfakes is within the scope of the PDPA and there are provisions on the correction of data, and the requirement to cease the collection, use and disclosure of personal data when consent is withdrawn. But these provisions are peppered with many exceptions, including if the material is already publicly available, which in almost all cases it would be, especially content such as deepfake pornography.

The recently introduced Online Safety (Miscellaneous Amendments) Act would not be of assistance to victims generally as it is not a framework intended for individuals to seek redress or any other kind of actions. Rather, it only regulates providers of online communication services, which are defined as services that enable end users to access or communicate content on the Internet using that service.

The new Part 10A introduced by the amendments currently only applies to social media services, and as such, it does not apply to content generally available on the Internet. Further, the regime is such that the social media service would need to adhere to Codes of Practice. Indeed, while the Infocomm Media Development Authority has the mandate to direct egregious content to be made inaccessible to Singapore end-users, such directions exclude communication between two or more end-users that is of a private or domestic nature. Thus, there are significant hurdles for a victim to rely on this new amendment.

Similarly, the Online Criminal Harms Act passed in July 2023 is not a framework intended for individuals to seek redress or any other kind of actions. Its intention is to tackle online content that is criminal in nature or used to facilitate or abet crimes and to counter harms that may be committed at great speed or scale, such as scams and malicious cyber activities.

In Singapore, the Protection from Harassment Act (Poha) goes some way towards helping victims, in giving them an avenue. The law aims to protect individuals from being harassed or stalked in the physical world and online, as well as being cyber-bullied and other acts causing alarm or distress. The acts caught by Poha, however, may not capture all instances of deepfakes, such as those that do not cause harassment, alarm or distress, but perhaps only cross the thresholds of causing discomfort, or insults.

The provisions in Poha also deal with false statements, which are defined to include the likes of images. The issue for deepfake images, however, is the burden of proof. It may not be easy for the victim to prove on the balance of probabilities that the content is fake.

In any event, it is not without costs and the legal process may be intimidating for victims. More often than not, in cases of online harms, the victim would not know the identity of the perpetrator, and a criminal case would be the only avenue. Apart from filing a police report, a victim can file a magistrate's complaint, which is then followed by a legal process, all of which can be daunting for the layperson.

Ultimately, as pointed out by Mr Shanmugam, going to court will take time and money, and is something the average person may not wish to undergo. Stronger laws are needed in Singapore to bridge the gaps.

LESSONS FROM ABROAD

Laws criminalising image-based sexual abuse were introduced in parts of Australia, such as Victoria, as early as 2013. These provide that it is an offence to record, distribute, or threaten to record or distribute, nude, sexual or otherwise intimate images without consent. A court may order rectification through orders to remove, retract, recover, delete or destroy any such intimate images.

Women are often the victims of such deepfakes. According to Sensity AI, a firm specialising in deepfake content detection, 90 per cent to 95 per cent of deepfake videos since 2018 were non-consensual pornography.

Some of these laws in some states were unclear as to whether digitally doctored images were covered by the legislation. After various public consultation processes in the middle of the last decade across Australia, legislation was strengthened to expressly include fakes.

At the national level in Australia, the federal Online Safety Act came into effect only in 2021, and doctored images are expressly within the ambit of the law. Today, corporations can be fined up to A$782,500 (S$680,200) for failure to comply with a removal notice under this piece of legislation. This goes some way towards prevention and holding perpetrators accountable. Such image-based abuse legislation would be helpful in Singapore, too.

In order for faster processes to remove offending images from the Internet, a "take-down" notice scheme targeted at those who host the content, similar to that for copyright-infringing materials, would be helpful, but to be truly effective, such a scheme would need to be formulated at the international level.

Laws could also mandate deepfakes to be labelled as such when they are uploaded to online platforms and for any alterations to be declared.

For instance, China's regulation on deepfakes requires any content that was created using AI to be clearly labelled with a watermark. China's laws go one step further to prohibit synthetic content that endangers national security and interests, harms the national image, harms societal public interest, disturbs economic or social order, or harms the legitimate rights and interests of others.

Deepfakes abound, not just pornography deepfakes. Equally offensive and harmful to individuals would be deepfakes pertaining to other aspects of their lives.

The time is ripe for further laws in Singapore to tackle generative AI harms.

• Hannah Yee-Fen Lim is an associate professor of Business Law at Nanyang Technological University who is qualified in law and computer science. She is an author of six books and an internationally recognised legal expert who has advised international bodies such as the World Health Organisation and the Law Commission of England and Wales.

Artificial intelligence technology can be used to generate fake text, audio, images and videos that are nearly impossible to detect, and has become a harmful weapon as it is easily available to the average person and doesn't require much expertise or resources, says the writer. PHOTO: PIXABAY