GENERATIVE AI AND DEEPFAKES

Posted on April 2, 2024

Category: CNPupdates

Authors: Wong Pei-Ling and Samuel Lee

1. Introduction

Deepfakes have become increasingly prevalent in recent years. In December 2023, several deepfakes promoting investment scams surfaced, with these deepfakes using the appearance of Singapore government officials like Prime Minister Lee Hsien Loong and Deputy Prime Minister Lawrence Wong.

The launch of generative Artificial Intelligence ("AI") tools such as ChatGPT have given rise to a new myriad of concerns in politics, business, education, arts and entertainment. Unlike predictive AI, generative AI has the ability to take its training data and create new content, which may be manipulated and misused in certain quarters.

Deepfakes created by adversaries may involve the use of generative AI to create fake content of persons doing or saying something that they did not do or say. When deployed by experts, deepfakes are able to effortlessly mimic a person's facial expression, mannerism, voice and inflections, making it very difficult to distinguish what is real or fake from the content. The increasing availability of free applications like Wombo, Reface and MyHeritage also makes it much easier for anyone, particularly malicious actors, to create deepfakes and produce disinformation at scale.

The issue of deepfakes is all the more pressing given that 2024 is shaping up to be the "super election year", with many countries and territories holding elections (according to a CNA article, at least 40 countries and territories are holding elections this year, including populous countries such as India, Pakistan, Indonesia and the USA). Deepfakes may serve as powerful tools to disrupt elections. This was recently observed in the USA, where, during New Hampshire's primary election in January 2024, AI-generated robocalls mimicking President Joe Biden's voice tried to discourage people from turning up to vote. More recently in February, the UK's home secretary warned that hostile countries could flood social media platforms with highly realistic AI-generated deepfake videos to sway voters in the upcoming UK election.

This article examines Singapore's current regulatory approach to addressing two key harms of deepfakes: (a) the facilitation of online scams and fraud; and (b) electoral disinformation (i.e. information that has been deliberately falsified).

2. Deepfakes and Online Scams

In a parliamentary sitting on 5 February 2024, the Minister for Communications and Information responded to a question on the "safeguards and regulations being put in place to tackle the issue of deepfake software being used in scam and fraud cases". This article notes from the Minister's response two features of the Singapore government's approach to addressing deepfake scams.

First, the Online Criminal Harms Act 2023 ("**OCHA**"), which came into effect on 1 February 2024, allows

the government to issue directions to online service providers, entities, and individuals to prevent potential scam related accounts or content from reaching Singapore users. Under the OCHA, designated online service providers may also be required to implement measures (if not already taken) to proactively disrupt online scams, including those facilitated by deepfakes.

Five types of directions may be issued under the OCHA:
• **Stop communication direction**: This requires the recipient of the direction to take steps to stop communicating specified online material to persons in Singapore. The recipients of the direction are persons and entities who communicated such online content.
• **Disabling direction**: This requires online service providers to disable access by persons in Singapore to specified online material (e.g. a post or page) on their service. This may include identical copies of the content.
• **Access blocking direction**: This requires internet access service providers to block access by persons in Singapore to an online location (e.g., a web domain) or specified online material.
• **Account restriction direction**: This requires online service providers to disallow or restrict interaction between an account on their service and persons in Singapore.
• **App removal direction**: This requires app distribution services (e.g., app stores) to stop distributing a specified app to persons in Singapore and to stop enabling persons in Singapore from downloading that specified app.

In order to counter the scale and speed of harm created through scams and malicious cyber activities, section 6(1)(b) OCHA specifies a lower threshold for issuing directions for scams and malicious cyber activity offences specified under Part 2 of the First Schedule. This is when there is suspicion or reason to believe that any online activity is being carried out in preparation to or in furtherance of the commission of a scam or malicious cyber activity offence.

The OCHA also provides a framework to strengthen partnership with online services to counter scams and malicious activities. Under this framework, the government is able to direct designated online service providers to proactively disrupt scams and malicious activities through the Codes of Practice issued by a Competent Authority responsible for administering OCHA.

Second, the Singapore government is collaborating with industry partners to improve local capabilities to deal with deepfake scams and fraud. For example, the Centre for Advanced Technologies in Online Safety ("**CATOS**") will be launched in the first half of 2024. The aim of CATOS is to bring together a community of research partners, companies and practitioners in Singapore to build capabilities for a safer Internet. Such capabilities may include tools and measures to: (a) detect harmful content, such as deepfakes and non-factual claims; (b) inject watermarks or trace the origin of digital content; and (c) empower vulnerable groups with resources to verify information they encounter online. These research efforts will also help inform new legislation or regulations needed to address issues, such as deepfakes.

3. Deepfakes, Disinformation, and Democracy

Deepfakes present two main threats to elections: (a) the use of deepfakes to alter voter preferences by spreading misinformation and "fake news" online; and (b) the undermining of trust in elections as the increasing sophistication of deepfakes makes it harder for the public to distinguish between what is real and what is not real. This article sets out some initiatives and regulations which address the issue of political deepfakes.

Singapore's Infocomm Media Development Authority ("**IMDA**") announced on 16 January 2024 that it had developed a draft Model AI Governance Framework for Generative AI ("**the Model Framework for Generative AI**"). This framework expands on the existing model governance framework (issued by Singapore's Personal Data Protection Commission) that covers traditional AI, which was last updated in 2020. The IMDA is currently holding a public consultation on the Model Framework for Generative AI and aims to finalise it in mid-2024.

The Model Framework for Generative AI notes that the rise of generative AI, which enables the rapid creation of realistic synthetic content at scale, has made it harder for consumers to distinguish between AI-generated content (e.g., deepfakes) and original content. The framework therefore recognises that there is a need for technical solutions, such as digital watermarking and cryptographic provenance, to catch up with the speed and scale of AI-generated content. Digital watermarking techniques embed information within content (with such information being used to identify AI-generated content), while cryptographic provenance solutions track and verify the digital content's origin and any edits made, with the records being cryptographically protected. In short, digital watermarking and cryptographic provenance both aim to label and provide additional information, and are used to flag content created with or modified by AI.

The proposed Model Framework for Generative AI also sets out nine dimensions to create a trusted environment – one that enables end-users to use generative AI confidently and safely, while allowing space for cutting-edge innovation. The framework also aims to facilitate international conversations among policymakers, industry, and the research community, to enable trusted development of generative AI globally.

The nine dimensions are as follows:
- **Accountability**: Putting in place the right incentive structure for different players in the AI system development life cycle to be responsible to end-users.
- **Data**: Ensuring data quality and addressing potentially contentious training data in a pragmatic way, as data is core to model development.
- **Trusted Development and Deployment**: Enhancing transparency around baseline safety and hygiene measures based on industry best practices in development, evaluation and disclosure.
- **Incident Reporting**: Implementing an incident management system for timely notification, remediation and continuous improvements, as no AI system is foolproof.
- **Testing and Assurance**: Providing external validation and added trust through third-party testing, and

developing common AI testing standards for consistency.
• **Security**: Addressing new threat vectors that arise through generative AI models.
• **Content Provenance**: Transparency about where content comes from as useful signals for end-users.
• **Safety and Alignment R&D**: Accelerating R&D through global cooperation among AI safety institutes to improve model alignment with human intention and values.
• **AI for Public Good**: Responsible AI includes harnessing AI to benefit the public by democratising access, improving public sector adoption, upskilling workers and developing AI systems sustainably.

Persons who use deepfakes to spread misinformation and disinformation online may be guilty of the offence of communicating a false message under section 14D of the Miscellaneous Offences (Public Order and Nuisance) Act 1906. A person who transmits or causes to be transmitted a message which he knows to be false or fabricated shall be guilty of an offence, and shall be liable on conviction to a fine not exceeding S$10,000 or to imprisonment for a term not exceeding 3 years or to both.

Under the Protection from Online Falsehoods and Manipulation Act 2019 ("**POFMA**"), a person who knowingly communicates online falsehoods or misleading content (e.g., deepfakes) is guilty of an offence if the communication of that falsehood is inter alia, likely to influence the outcome of an election in Singapore. It is also an offence for a person to make or alter a bot with the intention of communicating (or enabling any other person to communicate), by means of the bot, a false statement of fact in Singapore. The Minister may issue directions under Part 3 of POFMA to any person communicating a falsehood to: (a) put up a correction notice; or (b) to stop communicating the falsehood.

Additionally, under Part 4 of POFMA, internet intermediaries (e.g., social networking services and search engine services) may be subject to the following directions:
• **Targeted Correction Direction**: This requires the internet intermediary to communicate a correction notice to all end-users in Singapore who accessed the falsehood via its service.
• **Disabling Direction**: This requires the internet intermediary to disable access by end-users in Singapore to the falsehood.
• **General Correction Direction**: This requires prescribed internet intermediaries, newspapers, broadcasters, telecommunications service providers, and any other prescribed person to communicate a correction notice to its end-users.

However, the aforesaid directions can only be issued if: (a) a false statement of fact has been or is being communicated in Singapore through the internet; and (b) it is in the public interest to issue the direction.

"Public interest" means doing anything that is necessary or expedient in the interest of the security of Singapore, to protect public health or public finances or to secure public safety or public tranquillity, in the interest of friendly relations with other countries, to prevent any influence of the outcome of an election, to prevent incitement of feelings of enmity, hatred or ill-will between different groups of persons or to prevent diminution of public confidence in the performance of any duty or function of or in the exercise of any power by the government. Hence, individuals would not be able to rely on POFMA to correct every

instance of online falsehoods against themselves (for example, the spread of false news targeting celebrities), unless it satisfies the public interest element.

4. Conclusion

As noted in a January 2024 article by the Straits Times, "2024 has been described as the year of the deepfake, one in which a tsunami of disinformation will flood the world and people will be challenged as never before to tell truth from falsehood, what to believe and who to trust." This article observes that the Singapore government has adopted a fairly comprehensive approach to addressing the harms of deepfakes, namely: (a) regulations which enhance the government's powers to deal with deepfakes (e.g., the OCHA); (b) guidelines dealing with the development and use of AI (e.g., the Model Framework for Generative AI); and (c) research centres to improve local capabilities to deal with deepfakes (e.g., CATOS). Ultimately, the regulation of AI will require constant monitoring and fine-tuning to keep pace with a rapidly changing technological landscape.