Monetary Authority of Singapore

Information Paper

July 2024

Cyber Risks Associated with Generative Artificial Intelligence

Contents

1. Introduction

Since early 2023, Generative Artificial Intelligence ("GenAI") technology (e.g. ChatGPT) has gained significant interest and attention globally. While traditional AI focuses on analysing data and recognising patterns to make predictions or classifications, GenAI goes a step further to create new data or media content that can come in various formats, including text, image, video or audio. This allows GenAI to perform tasks, such as answering questions, summarising reports, or generating codes based on user-provided data or questions.

GenAI technology brings about potential opportunities for financial institutions ("FIs") to make their business processes more productive, efficient and convenient. Some of the possible use cases could be to automate data extraction, accelerate research, streamline operations, and enhance customer experiences. While there are many benefits to be reaped from the adoption of GenAI, FIs should also be mindful of the multifaceted risks and pitfalls that are associated with this development.

This paper aims to raise FIs' awareness by providing an overview of key cyber and technology related threats arising from GenAI, the associated risk implications, and some of the mitigation measures that FIs could take to address the risks. Specifically, section 2 details how GenAI can be used by threat actors, in the form of deepfakes and GenAI-enabled phishing, as well as malware generation and enhancement. Section 3 covers the threats and risks on GenAI deployments, namely, unauthorised information disclosure and data leakage, as well as GenAI model and output manipulation. A summary of the identified threats and their countermeasures can be found in the Appendix.

2. GenAI-Enabled Threats

2.1 Deepfakes and GenAI-Enabled Phishing

Deepfakes, which are AI-generated replicas of human video, image, or audio, can be used by threat actors to enhance the effectiveness of their social engineering when carrying out cyber attacks and financial scams. Security firms and agencies have identified numerous instances of deepfakes in the profile photos of fake personas. Threat actors can also create fake replicas of people within the victim's trusted circle, making the deception more believable by manipulating the inherent trust and familiarity.

Threat actors have also been observed to leverage GenAI models to facilitate more targeted phishing campaigns. Recently, ChatGPT introduced a feature, called GPTs[1], that allows users to add custom instructions or upload additional documents for a more tailored experience. This feature can potentially be misused[2] by threat actors to commit malicious acts, such as crafting professional looking phishing emails that may appeal to the victim's style of communication based on his/her public profile information.

Recent Incidents:

- In February 2024, fraudsters reportedly used a combination of malware and deepfake technology to target banking customers in Vietnam and Thailand. The malware extracted videos and images of the victims, along with their banking credentials and identity related documents, from their mobile phones. The fraudsters then used these extracted videos and images to create deepfakes of the victims' faces in their attempts to circumvent the facial biometric authentication of banks in Vietnam and Thailand.[3]

[1] https://openai.com/blog/introducing-gpts
[2] https://bbc.com/news/technology-67614065?at_medium=RSS&at_campaign=KARANGA
[3] https://www.group-ib.com/blog/goldfactory-ios-trojan/

- In January 2024, fraudsters used deepfake technology to target an employee, who worked in the finance department of a multi-national company in Hong Kong. Using publicly available video footages of the chief financial officer ("CFO") and his colleagues, the fraudsters were able to create deepfake videos of them for use in a video conference to social engineer the employee. The employee was successfully tricked into paying out US$25 million to the fraudsters.[4]

- In August 2023, GenAI tools were used to create doctored images of victims for loan scams targeting FIs and moneylenders in Hong Kong. These FIs required applicants to scan and upload identification documents, including real-time selfies, for loan applications. This resulted in a total of US$25,000 in fraudulent loans being approved.[5]

Possible Mitigating Measures:

- **Implement liveness detection techniques in facial recognition authentication solutions to counter deepfakes.** FIs using facial recognition for authentication can use a few techniques to make their solutions effective against deepfakes. For example, some FIs have used virtual liveness injection techniques[6], which detect subtle differences between live and synthetic faces through motion analysis, texture analysis, thermal imaging, 3D Depth analysis and behavioural analysis.

- **Conduct campaigns to raise user awareness on deepfakes and GenAI-enabled phishing.**[7] On top of phishing exercises, FIs could also consider running regular video and voice deepfake simulation exercises on employees. As part of the user awareness campaigns, staff can be taught to watch out for signs, such as unnatural face movements, inconsistent skin texture and

[4] https://www.straitstimes.com/asia/east-asia/hk-firm-scammed-of-34-million-after-employee-is-duped-by-video-call-with-deepfake-of-cfo
[5] https://www.scmp.com/news/hong-kong/law-and-crime/article/3232273/hong-kong-police-arrest-6-crackdown-fraud-syndicate-using-ai-deepfake-technology-apply-loans
[6] https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF
[7] https://academic.oup.com/cybersecurity/article/9/1/tyad011/7205694

edges around the face and distorted audio recordings. FIs can also advise their customers and staff to install applications like ScamShield[8], which could help protect users from deepfake-enabled scams by detecting and warning users of suspicious calls and phishing messages, and blocking blacklisted phone numbers.

- **Enable additional verification for high-risk transactions and for staff in high privileged roles.** FIs should implement multi-factor authentication for high privileged user accounts, such as those used by system administrators and financial officers, and especially for high-risk transactions, such as wire transfers, accessing sensitive customer information, or phone requests by customers or employees to reset account passwords.

- **Include deepfake attack scenarios in incident response.** FIs are encouraged to outline the steps to be taken in the event of a deepfake attack. This includes processes for reporting incidents, conducting investigations, communicating with stakeholders, and taking down deepfake content.

2.2 Malware Generation and Enhancement

GenAI tools, such as WormGPT[9] and DarkBard[10], are trained on malicious datasets, including malware-related data and phishing emails, and can be used to generate malicious scripts and develop malware codes. Such tools can enable threat actors, even those without much technical expertise, to create and circulate malware more cheaply and quickly.

Additionally, some malware were observed to use GenAI to help them implement polymorphism[11] to bypass traditional signature-based filtering and evade detection. These trends suggest that the

[8] Scamshield was developed by Open Government Products in collaboration with SPF with the aim to safeguard users against suspicious messages and calls.
[9] https://flowgpt.com/p/wormgpt-6
[10] https://outpost24.com/blog/dark-ai-tools/
[11] Polymorphism, which has been observed in various malware, is a process which enables a malware to make frequent changes to its binary code to avoid detection

availability of malicious GenAI tools could make malware capabilities more advanced and sophisticated to bypass existing defences.

Examples of Malware Leveraging GenAI/AI:

- **BlackMamba[12,13]:** BlackMamba is known to target FIs in European countries, such as Germany and Czech Republic. This is an example of malware that uses GenAI to implement polymorphism, making it hard for traditional security systems to detect. Once installed, the malware uses keylogging to steal sensitive information such as passwords and user credentials, leading to financial theft.

- **DeepLocker[14]:** This malware uses AI to conceal its malicious intent and avoid detection until a specific target or trigger conditions have been recognized. Once the target or trigger conditions are recognised, the AI model de-obfuscates the hidden malware and executes it. To demonstrate DeepLocker's potential, security researchers created a proof-of-concept ("PoC") in which WannaCry ransomware was hidden in a video conferencing application. The malware was not detected by anti-virus engines or sandboxing.[15]

Possible Mitigating Measures:

- **Adopt a multi-layered cyber defence strategy.** Given the increasing sophistication of cyber attacks, it has become all the more important for FIs to implement a multi-layered cyber defence strategy, so that even when one or more of the security measures get circumvented, the other measures can help to mitigate the risks. It is fundamental for FIs to maintain basic cyber hygiene which continues to be relevant to counter AI-enabled cyber attack modus operandi.

[12] https://hitachi-systems-security.com/black-mamba/
[13] https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware
[14] https://www.bitdefender.com/blog/hotforsecurity/deeplocker-new-breed-of-malware-that-uses-ai-to-fly-under-the-radar/
[15] https://www.zdnet.com/article/deeplocker-when-malware-turns-artificial-intelligence-into-a-weapon/

- **Monitor and consider incorporating AI tools to detect polymorphic malware on corporate endpoints.** FIs are encouraged to monitor developments in AI-enabled malware detection and endpoint security solutions, and adopt them where appropriate. Such solutions leverage machine learning and heuristics-based behavioural detection to stop both legacy and new malware threats, which can evolve quickly to circumvent conventional security measures. These solutions complement existing signature-based anti-virus to protect against both known and unknown threats, file-less and signature-less attacks.

- **Incorporate AI and integrate threat intelligence into log monitoring to better identify anomalies and suspicious activities.** FIs are encouraged to incorporate AI solutions with log management systems, which work by ingesting data points from devices throughout the network. These logs are then analysed with machine learning models in real-time to help with faster anomaly detection, event correlation, making predictions and auto-remediation. FIs are also encouraged to integrate such tools with threat intelligence platforms and services to keep up with the evolving tactics, techniques and procedures of attackers, and use the information to enhance their log analysis to better identify suspicious activities or a potential breach.

3. Threats Targeting GenAI Deployment

3.1 Data Leakage from GenAI Deployment

FIs that allow employees to use publicly accessible GenAI tools (e.g. ChatGPT and Gemini) could be subject to potential data leaks if their employees submit or upload sensitive data while using those tools. FIs could also be exposed to data leakage risks through unauthorised insider actions and improper data handling when using the GenAI solutions. Where such data leaks involve confidential information and intellectual property, the FI could end up facing legal, regulatory and reputational consequences.

Data leaks could also arise from prompt injection attacks against the GenAI solution. In a prompt injection attack, the threat actor provides a malicious input to lead the GenAI model into revealing sensitive information in the generated response. There have also been instances of jailbreak attacks, where the threat actor uses specially crafted prompts to bypass the security controls and safeguards implemented to make the GenAI solution disclose sensitive information.

While using third-party or open-source GenAI models, FIs will need to manage the attendant supply chain and third-party risks. For example, malicious open-source models or open-source models with poor security controls may introduce vulnerabilities or backdoors, which could also result in data leaks.

Recent Incidents:

- In March 2024, security vulnerabilities were discovered in GenAI browser plugins (e.g. AskTheCode, Charts) which could enable attackers to access information browsed by the victim or access victim's data stored in other services such as Google drive and Github.[16]

[16] https://www.securityweek.com/chatgpt-plugin-vulnerabilities-exposed-data-accounts/

- In March 2023, a bug in redis-py, a Redis[17] client open-source library used by ChatGPT to cache user information, exposed users to another user's personal information, partial payment card details and historical chat queries.[18]

- In March 2023, a major electronics company reportedly experienced three separate cases of employees leaking sensitive data while using ChatGPT. This occurred when employees entered corporate information, such as confidential source code and meeting minutes, in the ChatGPT prompt.[19]

Possible Mitigating Measures:

- **Establish user policies and conduct employee awareness campaigns on security best practices in relation to GenAI usage.** FIs should have clear data classification and GenAI usage policies to guide employees on how to use GenAI safely, and the type of data that can be used on public GenAI solutions. It is important for FIs to raise the awareness of their employees on how to use public GenAI models safely, for example to desensitise data inputs provided, and not input any confidential information.

- **Adopt security best practices when developing in-house GenAI models:**

    - **Implement security-by-design approach and secure coding.** FIs are encouraged to adopt a security-by-design approach and secure coding practices while developing in-house GenAI models to minimize the vulnerabilities introduced. This would include performing threat modelling and incorporating security considerations during the requirements gathering and design stages, as well as implementing secure coding and code reviews during the development stage.

[17] Redis, short for Remote Dictionary Server, is an open-source, in-memory data structure store which can serve multiple purposes such as database and cache.
[18] https://openai.com/blog/march-20-chatgpt-outage
[19] https://news.cgtn.com/news/2023-04-03/Samsung-finds-data-leak-due-to-use-of-ChatGPT-Korean-media-1iHSzPcMDEk/index.html

- **Perform vulnerability assessments and security testing.** Similar to the testing done for other applications, FIs are encouraged to perform vulnerability assessments, penetration testing and red teaming on their GenAI solutions. FIs are encouraged to test their GenAI models against common types of model attacks, such as those listed by the Open Web Application Security Project ("OWASP")[20]. These tests serve to evaluate and validate the security posture of the GenAI model, and to identify potential security vulnerabilities for remediation.

- **Perform proper due diligence when using third-party or open-source GenAI solutions.** FIs should conduct risk assessments, robust testing and model validation on third-party or open-source GenAI solutions before using them. Steps should be taken to check and ensure that the data used to train these models are not tainted. To facilitate this, FIs could use model cards[21] to document model technical details, such as model capabilities, security vulnerabilities, information on the training data, and training methodology used to train the model.

- **Implement data loss prevention ("DLP") and firewalls for GenAI models.** FIs are encouraged to implement appropriate DLP controls to check for any sensitive data provided in the prompts, as well as the responses generated by their GenAI solutions. GenAI firewalls that are purpose-built for GenAI models could also be implemented to analyse user inputs to detect any attempts to extract data or exploit the GenAI solutions.

[20] https://owasp.org/www-project-top-10-for-large-language-model-applications/

[21] Model cards provide details on a model's capabilities, limitations, biases, security vulnerabilities, and performance metrics, enhancing transparency, compliance, and risk management for FIs using these models.

3.2 GenAI Model and Output Manipulation

Threat actors can introduce malicious or inaccurate data, for example through data poisoning attacks, to manipulate the GenAI models and their outputs. This can take place during the training stage, or while using the models:

- During training of the GenAI models, a threat actor who has unauthorised access to the Foundation Model ("FM") can introduce incorrect or fabricated data into the training data. The GenAI model would then be trained to generate outputs and responses that benefit and favour the threat actor, at the expense of the users. For example, threat actors with access to the FM can potentially manipulate the model to make investment recommendations based on false information, or inject false information into models for algorithmic trading, leading to inaccurate predictions and financial losses to the users.

- While using GenAI solutions, risks can be introduced in the form of (i) data poisoning attacks in which the users of the GenAI solution feed malicious data to the GenAI model, or (ii) jailbreak attacks. Such attacks may lead to undesirable or unpredictable behaviour, including inaccurate, offensive, or harmful outputs. For example, if a user of a GenAI-enabled chatbot feeds a piece of dangerous or offensive information to the chatbot, the chatbot could in turn use that data in its response to other users.[22] Such methods can be used by malicious actors to spread disinformation through the chatbot.

Recent Incidents:

Although there are no publicly known incidents of data manipulation in the financial sector, there have been such incidents reported in other sectors:

[22] https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

- In June 2023, researchers in the US reported[23] about how an AI-powered medical diagnosis system could be misled by manipulated data to influence and affect the accuracy of medical diagnosis. This shows the potential implications that poisoning attacks can have on AI models used in crucial decision-making systems.

- In July 2023, a cybersecurity firm demonstrated[24] how GenAI model manipulation could be performed by training the GenAI models on fabricated data, or making unauthorised modifications to inject falsified information into the model, to influence their responses. Such model poisoning could become vectors for spreading misinformation or inserting harmful backdoors, especially in applications such as AI coding assistants.

Possible Mitigating Measures:

- **Put in place proper GenAI model and data governance.** FIs should establish proper model governance to ensure the end-to-end integrity, accountability, and auditability of GenAI models. FIs should also implement robust data governance processes and perform data quality checks to ensure the accuracy, consistency, and completeness of the data used with the GenAI model. This includes incorporating measures for data cleaning, data validation and anomaly detection.

- **Ensure robust access controls to the GenAI training data and foundation model.** FIs should implement strict access controls to limit who can access and modify the AI resources and training data. FIs could have a maker-checker process, such as human-in-the-loop approach, to ensure that changes to the training data or foundation models are vetted by two or more individuals. FIs should also implement the principle of least privilege to limit the personnel who have access to training data and implement logging to monitor changes to the training data set.

[23] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10235827/
[24] https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/

- **Implement continuous monitoring and validation of GenAI models.** FIs are encouraged to implement comprehensive logging and continuously monitor GenAI models for performance degradation, model drift, or unexpected behaviour that could indicate a possibility of data poisoning. FIs that implement retrieval-augmented generation ("RAG")[25] with their GenAI models should also update their knowledge bases with fresh and clean data that have been checked and validated.

- **Incorporate contingency measures for GenAI solutions into business continuity plans ("BCP").** FIs are encouraged to incorporate contingency measures for GenAI solutions in their business continuity plans. This could include keeping a backup of the training data which can be reused should there be signs of degradation or manipulation of the GenAI model. For critical processes using GenAI models, FIs are encouraged to ensure that contingency measures are in place in case of disruptions or security incidents related to the GenAI model.

- **Participate in information sharing to identify issues related to GenAI model deployment.** FIs are encouraged to share information, such as useful learning points and challenges encountered while deploying GenAI models, to the wider FI community via information sharing platforms, such as Information Sharing and Analysis Centers ("ISACs") and industry associations' Standing Committee on Cyber Security ("SCCS").

[25] Retrieval-augmented generation ("RAG") is the process of optimising the output of a GenAI model by referencing knowledge bases outside of its training data sources before generating a response.

4. Conclusion

In conclusion, the rapid advancement and widening application of GenAI technology, along with the potential of existing tools and controls being rendered ineffective by GenAI-enabled threats, have brought about new cyber risks to FIs. This paper has highlighted four key areas of risks that FIs should pay attention to, namely (1) deepfakes and GenAI-enabled phishing, (2) malware generation and enhancement, (3) data leakage from GenAI deployment, and (4) GenAI model and output manipulation. FIs should be cognizant of the evolving GenAI developments and their risk implications, as well as keep abreast of the relevant industry best practices and risk mitigation strategies to safely harness the benefits of GenAI.

Appendix: Summary of Threats, Impact and Countermeasures

Aspect	Threats	Impact	Countermeasures
2.1 Deepfakes and GenAI-Enabled Phishing			
People	• Deepfakes can make impersonation more convincing, posing a significant risk for organizations	• Business e-mail compromise (BEC) • Loss of personally identifiable information (PII) or FI confidential data • Financial Loss	• Conduct deepfake awareness campaigns • Reconfirm identity with another factor • Leverage new technologies to detect deepfakes and AI-generated content
Process	• Threat actors can leverage GenAI to create realistic images, videos, speech, and personas for fraud and deception	• Financial loss • Identity theft • Data exposure	• Implement additional user verification for high-risk transactions and high-risk roles by different means such as challenge question, email etc.
Technology	• New account creation fraud, and bypassing existing authentication mechanisms with fake identities • Spread of fake news with botnets (with possibility of market manipulation)	• Fraud risk • Security compromised • Money laundering schemes • Evading security protocol • Market Manipulation	• Implement deepfake detection tools • Implement additional verification for high-risk transactions
2.2. Malware Generation and Enhancement			
People	• Malware sent through phishing links	• Monetary loss • Loss of PII	• Conduct user awareness campaigns • Maintain basic cyber hygiene
Process	• Inability to detect new malware	• Delayed detection of threats • Delayed containment of malware	• Adopt a multi-layered cyber defence strategy • Incorporate solutions that leverage machine learning and heuristics-based behavioural detection to stop both legacy and new malware threats
Technology	• Polymorphic AI malware which evades detection • Outdated enterprise systems which are unable to detect malware	• Bypass of security measures leading to loss of sensitive information, financial loss and reputation damage	• Implement AI-powered tools to detect polymorphic malware • Incorporate AI and integrate threat intelligence into log monitoring to better identify anomalies and suspicious activities

Aspect	Threats	Impact	Countermeasures
3.1. Data Leakage from GenAI Deployment			
People	• Intentional or unintentional data leaks by employees to public GenAI models	• Loss of customer data/ PII and FI secrets • Regulatory consequences and reputational damage	• Conduct awareness campaigns • Implement data classification for data which can be entered into GenAI models
Process	• Vulnerabilities or security weaknesses in in-house developed GenAI models • Risks of supply chain attack arising from the use of third party or open-source GenAI models	• Data Leakage leading to loss of sensitive information • Backdoors and in-built vulnerabilities	• Adopt security best practices while developing GenAI models • Conduct third-party provided or open-source GenAI model risk assessment
Technology	• Inability to detect unusual user inputs • Bypass of GenAI model guardrails	• Loss of sensitive information • Data Leak of PII • Reputational damage	• Implement DLP tools and firewalls for GenAI models to mitigate loss of confidential data to GenAI models • Introduce controls while developing and using the GenAI models • Conduct vulnerability assessments and security testing on GenAI models
3.2. GenAI Model and Output Manipulation			
People	• Insider threats • Access to foundation model and training data is not limited	• Unauthorized data access and loss of data integrity	• Implement maker-checker function to edit data in foundation models • Implement human-in-the-loop to verify that the output is as expected.
Process	• Lack of proper access control to GenAI model data • Improper data governance for data used to train GenAI models if the data is not sanitised and verified • Lack of contingency measures for GenAI solutions	• Unauthorized data access • Poisoning of foundation model data • Impact to business operations due to disruptions to GenAI solutions	• Ensure robust access controls to the GenAI training data and foundation model • Establish proper GenAI model and data governance • Include contingency measures for GenAI solutions into BCP • Conduct information sharing on issues and challenges faced during GenAI model deployment
Technology	• Inability to monitor model performance, model drift, or unexpected behaviours • Inability to detect unusual model outputs	• Incorrect information provided to users • Reputational damage • Regulatory consequences	• Implement tools to log and monitor output of GenAI models