THREAT REPORT

Title	Cybersecurity Implications of Generative AI
Reference	II-TR-014
Date	Janurary 2024
Produced by	Insights & Intelligence, CERT NZ - Computer Emergency Response Team New Zealand

Overview

CERT NZ's Insights and Intelligence team produces periodic threat reports to illustrate and identify notable trends in information we collect. This includes highlighting threats and possible actions or interventions for consideration.

This report aims to inform IT management of the adversarial and defensive applications of recent advances in generative artificial intelligence (GenAI).

The rise of generative AI

GenAI is a category of AI that emerged in the 1960s in the form of simple chatbots. In early 2023, new models, capable of a wide variety of tasks, were adopted by a wider audience. One of the most popular of these is OpenAI's ChatGPT[1] which is proving useful as an assistant for work and creative endeavours.

ChatGPT is an advanced chatbot that applies machine learning techniques to a large language model (LLM) consisting of billions of parameters. When a user asks a question as an input, the LLM finds strongly related words in its dataset and strings them together into a human-like response.

At time of writing, OpenAI's most advanced LLM, GPT-4, has multimodal capabilities, accepting either text or images as inputs. For example, it can transform a hand-drawn sketch into a fully functioning web page.

A growing number of products, both commercial and open source, employ GenAI in novel ways.

- Summarising information.
- Generating music from a description.
- Reviewing code.
- Detecting fraud.
- Converting text-to-speech.
- Modelling and simulating biological systems.
- Designing fashion items.

The many practical applications of GenAI is not surprising, as some products have been shown to perform as well as humans in various academic and professional benchmarks.

[1] Introducing ChatGPT (openai.com)

Malicious cyber applications

GenAI's ability to mimic human skills and abilities can also be used for nefarious purposes. In response, some vendors have put so-called 'guardrails' in place. While this can make generating potentially harmful content more difficult, determined individuals continue to discover workarounds. Making guardrails too restrictive though may be a double-edged sword – individuals may turn to an alternative GenAI tool that is easier to work with or better suited to malicious purposes.

The following are some examples of malicious cyber applications of GenAI.

Malware development

GenAI could create malware from scratch although, depending on its complexity, knowledge of the computer language may be required to make the output functional. Also, rather than limiting the output to work on a single architecture, GenAI could generate multiple builds to infect a broader target set.

Bypassing threat detection

GenAI could examine a malware sample and suggest changes to the code or design. Drawing on its training data, it could generate a new strain that is more intelligent, adaptable and less likely to be detected by antivirus / intrusion detection systems.

Building on capabilities

New capabilities can take a lot of time and effort for a threat group to develop. In place of a team, a GenAI could help assess whether a threat group's objectives can be met with their existing capabilities and provide step-by-step instructions for making improvements.

Target profiling

GenAI can be tasked with scraping personally identifiable information (PII) from websites and social media. This information could be used to guess somebody's password or to relate to them better in a social engineering attack.

Impostor scams

Recent advances in voice impersonation have led to a rise in so-called impostor scams. A neural codec language model known as Vall-E[2] can learn to match a person's voice and manner of speaking within seconds. A malicious actor could leverage the AI to impersonate a close contact of the target, claiming they are in distress and needing an urgent transfer of funds.

Sophisticated phishing

Phishing emails written by a malicious actor often give themselves away with their poor use of English, but an AI modelled on the communications of a colleague, bank or company could craft more convincing content – and in almost any language. The same goes for malicious links – the website one lands on could appear more genuine than what an actor could build manually.

Automated penetration testing

GenAI-based tools have become available that can automate the penetration testing process. An example of such a tool is PentestGPT[3], which harnesses GPT-4 to guide a user through each step. This

[2] VALL-E (X) (microsoft.com)

[3] GitHub - GreyDGL/PentestGPT: A GPT-empowered penetration testing tool

includes writing phishing emails, generating malicious payloads and running exploits.

Implications for cyber defence

As discussed, threat actors can use GenAI to develop malicious content faster and deploy at scale. In response, cybersecurity vendors are scrambling to adopt it for defensive purposes – for example Microsoft Security Copilot[4] and Crowdstrike's Charlotte AI[5]. Using the same technologies against the aggressors in this way is essentially pitting machine against machine in a rapidly evolving threat landscape.

There are many examples of how GenAI can analyse extensive datasets to assist with cyber defence.

- Simplifying technical issues for non-technical staff for quicker decision making.
- Monitoring networks for malicious indicators or suspicious activity.
- Identifying software and hardware misconfigurations and vulnerabilities.
- Proactive threat hunting – alerting on unusual trends or patterns at an early stage.
- Simulating a hostile environment for training defenders.
- Writing and deploying scripts to automate tasks such as:
    - finding stale active directory accounts and disabling them, or
    - scraping the web for data leaks, brand counterfeiting and phishing campaigns.

When it comes to protecting a corporate network, whether the threat is human or machine shouldn't make any difference to an organisation's existing security posture – the same rules apply. Use this report to serve as a timely reminder to review the basics. CERT NZ's Critical Controls[6] are updated to mitigate most attack types seen in New Zealand. These will work to mitigate most of the malicious actions we will potentially see from GenAI in the future.

Sharing this content

This report is produced by the CERT NZ Insights & Intelligence team and is intended for sharing with CISOs and IT professionals in similar positions.

If you would like to provide feedback or get in contact with CERT NZ, please email: info@cert.govt.nz

[4] Microsoft Security Copilot combines GPT-4 with a security-specific model to assist with cyber defence. Currently in the testing phase, it can correlate threat activity, answer questions, respond to threats, predict attacks, reverse engineer exploits, and learn from user feedback.
Introducing Microsoft Security Copilot: Empowering defenders at the speed of AI - The Official Microsoft Blog

[5] CrowdStrike's Falcon Platform includes Charlotte AI which allows users to ask natural language questions to assist with threat detection and analysis.
CrowdStrike Introduces Charlotte AI to Deliver Generative AI-Powered Cybersecurity

[6] CERT NZ's Critical Controls | CERT NZ