Artificial Intelligence (AI) Powered Cyber Threats

SamCERT Executive Advisory

July 2026

Executive Summary: Managing AI Powered Cyber Threats

Artificial intelligence (AI) is rapidly becoming a fundamental component of our everyday lives, unlocking significant opportunities to improve government services and boost economic productivity. As we harness these positive impacts, we must also prepare for the new security challenges this technology creates.

The Challenge:

Malicious actors are already using publicly available AI tools to attack government systems, steal massive amounts of data, and create highly convincing scams targeting both officials and citizens.

These tools enable even low-skilled criminals to automate attacks, write malicious code, and analyse stolen information with unprecedented speed and efficiency. This means we face the risk of more frequent and more effective attacks on our critical infrastructure, government services, and sensitive national data.

Key Message for Executives:

Crucially, **AI does not create entirely new types of cyber vulnerabilities.** Instead, it excels at finding and exploiting existing weaknesses, such as unpatched software and human error. This presents both a challenge and an opportunity.

Our defence does not require massive investment in complex AI systems. Instead, we must focus on mastering the fundamentals of cybersecurity.

'The Con Artist' vs 'The Burglar': AI Power Threats Targeting People and Systems

A useful way to understand AI-powered threats is to see them through the lens of the attacker's aim. Malicious actors use generative AI capabilities for two distinct purposes, which creates two threat categories: attacks targeting people and attacks targeting systems.

Think of it as the difference between an attacker using AI as a sophisticated con artist to deceive our people, versus using it as a highly efficient burglar to breach our technology.

AI as 'The Con Artist'

Using AI to manipulate human psychology, trust, & behaviour to trick people into making mistakes.

Where we are seeing these threats

Deepfakes & Reply Bots

Deepfakes are making fraud more dangerous by enabling attackers to imitate real people through fake video, audio, and images.

Tailored Phishing Emails

Phishing emails are harder to detect as AI gleans public-facing information to make phishing content less detectable.

Fraud at Scale

Fraudsters use AI to automate personalised, polished attacks at a much larger scale.

Your ministry's primary defence against these threats

Human Resilience

Staff training and awareness to recognise spoofed, suspicious or inappropriate requests before acting.

Stronger Authentication

Use multi-factor authentication to block account takeover, even when AI improves deception.

Email Defences

Use approved email systems with external warning banners and spam and keyword filters configured to flag potentially deceptive materials.

VS.

AI as 'The Burglar'

Using AI to automatically discover & exploit technical vulnerabilities in software, networks, and code

Where we are seeing these threats

Weaponised Vulnerability Discovery

AI models like Mythos are scanning systems to find unpatched security flaws faster than human defenders can close them.

This is a **significant threat** when considered alongside AI malware generation and automated hacking

Your ministry's primary defence against these threats

Technical Resilience

Will required significant uplift cyber security fundamentals – including patching, network segmentation and security monitoring

...recent developments in the government threat landscape make AI as 'The Burglar' an emerging priority risk. *See the case study on the following page.*

Samoa Computer Emergency Response Team (SamCERT)

AI as 'The Burglar' Case Study:
AI Powered Cyber Attacks Targeting Government Ministries

The Situation

In March, the Government of Mexico was targeted by cyber criminals.

These cyber criminals used jailbroken versions of Claude to generate and execute the attack code; whilst ChatGPT was used to comb through the terabytes of data.

Claude + **ChatGPT**

Created and executed the attack code, via hackers overruling its normal safeguards.

Sorted through terabytes of data via a generated python script, also being jailbroken by the hackers.

The Impact

The use of AI-enabled capabilities contributed to a significant large-scale data breach affecting the Government of Mexico, resulting in extensive media scrutiny, erosion of public trust, and the irretrievable compromise of sensitive data that could be used to inflict further harm on government and citizens.

The scale of this attack included:

What's to come:
Purpose built tools for this kind of activity

Notably, this attack used standard generative AI tool to cause significant harm. In April, Anthropic announced a specialised AI model called Mythos; purpose-built for cybersecurity.

This model has only been released to a select group of companies, but it highlights how frontier models will accelerate security flaw detection & enable attacks at scale.

April 2026 Mythos was announced, with a very limited release.

...this is AI as 'The Burglar' in action

271 vulnerabilities Were detected by Mythos in a recent version of Firefox.

Andrew Martin and Carolina Millan, "AI Chatbots Weaponized in Massive Mexican Government Data Breach," Los Angeles Times, March 9, 2026,

The Key Lessons for us

Beyond standard AI tools, cyber-purposed AI models are being built which will raise the capability ceiling and reduce the cost of running large-scale cyber attacks. These capabilities will likely continue to:

- *Increase discovery speed for bad actors*
- *Enable attack preparation & execution*
- *Drastically accelerate analysis of compromised data – enabling flow on harm*

However, the fundamentals remain the same.

Artificial intelligence does not create new vulnerabilities – **it is only revealing old ones.** This means we need to ensure our cyber foundations are strong.

- Proactive Upskilling of ICT Staff
- Constant Patching
- Incident Response Plan and Escalation Paths Defined and Tested
- Layered Defences and Network Segmentation

Samoa Computer Emergency Response Team (SamCERT)

If you require more information, please contact SamCERT on:

samcert@mcit.gov.ws

www.samcert.gov.ws/