Good Ancestors
NEW ZEALAND

Cyberattacks

Key points

- Criminals and state-backed groups are already using AI systems in real cyber operations, and AI is lowering the barrier for less-skilled attackers.
- AI can now find and exploit software vulnerabilities on its own, and the first largely autonomous attacks have been documented.
- Within a few years, autonomous cyber agents could run entire campaigns at the level of today's top state-backed hackers, putting power, water, hospitals and banking at greater risk.
- The Five Eyes cyber security agencies, including New Zealand's National Cyber Security Centre, have issued a call to action on AI preparedness.

AI is changing the economics of cyberattacks. Capabilities that once required skilled, well-resourced teams can now be deployed quickly and at modest cost. This brief sets out what attackers can do now, how the threat is likely to grow and what it means for New Zealand.

Attackers are already using AI

Major AI companies report attackers using their systems in real operations. AI is also being used to write malicious code, produce phishing at scale and clone voices for scams. In one case, a cybercriminal used an AI coding agent to run a data extortion campaign against at least 17 organisations, including in healthcare and government, with the AI helping to decide which data to steal and how large a ransom to demand.

The immediate effect is a lower barrier to entry for cyber crime. Tasks that once required experienced hackers, such as analysing target systems and producing code to exploit any vulnerabilities, can now be delegated to AI. Small or inexperienced groups can operate at a level previously limited to well-resourced teams.

Security agencies are treating this as a live problem. In 2024 alone, the cost to New Zealanders from online threats was estimated to be $1.6 billion. In June 2026 the leaders of the Five Eyes cyber security agencies, including New Zealand's National Cyber Security Centre (NCSC), issued a joint call to action on AI preparedness. The NCSC has warned New Zealand organisations to prepare now for a significant increase in cyberattacks and the business disruption they will cause.

AI is beginning to attack on its own

AI systems no longer just assist human operators. In the past year:

- Anthropic reported disrupting the first documented large-scale cyberattack executed largely without human involvement. In late 2025, a group manipulated Claude into attacking around thirty targets, including technology companies, banks and government agencies. The AI performed 80 to 90 percent of the campaign, with human operators intervening at a small number of decision points.
- In April 2026, Anthropic disclosed that its most capable model had autonomously discovered and exploited previously unknown vulnerabilities in every major operating system and web browser, including a flaw that had gone undetected in the operating system OpenBSD for 27 years. It built working exploits without human guidance in hours.

- In July 2026, two of OpenAI's model being tested for cyber capabilities escaped a supposedly isolated environment. It reached the internet and broke into the production systems of the online platform Hugging Face, stealing credentials and gaining remote control of servers, with no human directing it.

Where this is heading: highly autonomous attacks

Researchers now warn of highly autonomous cyber-capable agents, or HACCAs: AI systems able to plan and run entire campaigns at the level of today's most sophisticated criminal groups and state-backed hackers. These are operations that currently require dozens of experienced people working for months, with budgets over a million dollars.

Leading AI models went from near-zero to a 60 percent success rate on expert-level offensive security challenges within months. On current trends, HACCA-level systems could be feasible before the end of the decade.

Once these capabilities exist they are very likely to spread to open models without safeguards. This would put these tools in the hands of malicious actors. Researchers also warn such agents could sustain themselves, funding their own operations through fraud or hijacked computing.

New Zealand has already experienced the cost of successful attacks:

- The 2021 ransomware attack on Waikato DHB took hospital systems offline for weeks and delayed treatment, leading to $16.5 million in insurance claims.
- In 2020, attacks disrupted trading on NZX, New Zealand's only stock exchange, for four days.

Both attacks predate modern AI. New Zealand's critical infrastructure will increasingly be at risk. Power supply, water supply, hospitals and banking all run on software, almost all of it built offshore, and AI may find flaws in it faster than defenders can patch them.

Offence, defence and where the balance sits

The same advanced AI capabilities strengthen cyber defence. Anthropic restricted its vulnerability-finding model to vetted security partners, who within two months had surfaced more than 10,000 high- or critical-severity flaws in widely used software so they could be fixed. Hugging Face's own AI tools detected the July intrusion. The NCSC's guidance to New Zealand organisations also covers using AI to manage cyber risks.

However, the near-term balance is still likely to favour attackers:

- An attacker needs one exploitable flaw; defenders often require coordinated effort from many organisations and must find and fix all the vulnerabilities to be safe. AI has made flaws far easier to find than to fix, and patching is now the bottleneck.
- Much critical infrastructure runs on legacy systems that are slow or impossible to patch, so some systems remain exposed long after flaws are found.

It's possible this gap between offence and defence will narrow if AI-assisted patching becomes more effective and access to the most capable models is initially restricted to key institutions to enable them to update their systems. Until then, attackers likely hold the advantage.

What New Zealand can do

The government should act now. Priorities include:

- Strengthening the cyber defences of critical infrastructure, on the assumption that vulnerabilities and incidents will rise sharply.
- Resourcing the NCSC's frontier AI work programme, including ensuring it retains access to leading models, using AI for defence and rolling out faster patching across government.

Our accompanying policy papers set out these and other steps in detail.