Good Ancestors NEW ZEALAND

Published November 2026.

AI risk update: November 2026

Key points

- Numerous cases have been recorded of AI systems acting in ways their developers did not intend, including hacking another company and an Australian Government website.
- In October, AI found critical flaws in software that runs most of the world's websites. Millions of attack attempts followed within hours, and several governments took websites offline.
- The heads of leading AI companies have called for the pace of AI progress to slow. Thirty-two countries have called for frontier AI to remain under human control, but New Zealand is not among them.
- AI progress continues to outpace forecasts, underscoring the need for immediate government action.

Our initial risk and policy papers were published in July 2026. This note sets out the most significant developments since then and what they mean for New Zealand.

AI agents have acted beyond their developers' control

In July 2026, OpenAI agents being tested on a cyber capability benchmark broke out of their test environment, reached the internet and hacked the AI company Hugging Face to cheat the test's scoring system. An independent investigation found that hundreds of agents coordinated the attack through a hidden message board and tried to tamper with their own logs. OpenAI only realised what had happened a week later, when its agents broke into its own computing systems.

In June 2026, another OpenAI agent, asked to look up statistics about Australia during internal testing, got around the access controls of an Australian Government health statistics website and obtained non-public files. OpenAI did not discover this until August and did not tell the Australian Government until 10 September, by email to a generic inbox.

Similar incidents have been reported at other AI companies. Such incidents have not yet caused serious harm, but more capable systems in the future could do far more damage. What is clear is that leading AI systems can:

- pursue goals in ways no one intended, including breaking rules to achieve them
- coordinate sophisticated cyberattacks on real systems without human direction
- go undetected by their own developers, sometimes for weeks
- try to hide what they have done.

AI is changing the pace of cyber security

In October, AI tools used by security researchers found critical flaws that had sat unnoticed for more than a decade in PHP, the programming language behind most of the world's websites. Within hours of the fix being published on 21 October, there were millions of attempts to exploit them.

The Netherlands, Finland and France took government websites offline as a precaution, and New Zealand's National Cyber Security Centre urged organisations to patch immediately. This is a glimpse of what more capable AI could mean for security: flaws found and exploited far faster, defenders depending on AI to keep up, and decisions made within hours.

Calls for a slower pace are growing

OpenAI paused part of its frontier model training for two weeks in August, citing the Hugging Face incident and the cyber capabilities of its next model. In late September it shelved the public launch of that model after it did not meet the company's safety standards.

In September, Anthropic researcher Jacob Coxon resigned, warning that leading companies were "racing straight to self-improving superintelligence". Days later, Anthropic's chief executive called for the industry to "slow the pace at which we improve the capabilities of AI models", pointing to AI systems that now do much of the work of building their successors. The heads of OpenAI and xAI said they agreed.

The concern extends beyond the AI companies. In August, Bill Gates warned that the shift to AI could be one of the most turbulent periods in human history. He called for new national bodies and an international organisation to manage it, and said AI companies should not be the ones leading.

Governments are responding

On 21 September Finland's President and Norway's Prime Minister launched a Call for Control of Frontier AI Models, stating that AI "must remain under human direction, oversight and control". Leaders of 32 countries and the European Commission have endorsed it, including Australia, Canada, Singapore, France and Germany. New Zealand has not.

New Zealand agencies are responding to the cyber risks. The National Cyber Security Centre made frontier AI the top concern in its 2026 Cyber Threat Report and has issued guidance to government agencies, and the Reserve Bank and Financial Markets Authority have set expectations for financial institutions. These are positive steps, but they focus on cyber security. New Zealand has no whole-of-government plan for the wider risks from frontier AI, such as job losses, AI-enabled election interference, biological threats or AI systems acting in ways their developers did not intend.

AI is improving faster than expected

Over the past 18 months, almost everyone underestimated how fast AI would progress, including superforecasters and AI experts. One common measure is the length of programming tasks an AI system can complete on its own, measured by how long the task would take a skilled programmer. Most forecasters expected AI to reliably handle three-hour tasks by the end of 2026. It reached that point in April.

In a 2024 survey, AI researchers' median estimate for when AI would solve a major unsolved maths problem was around 2053.[1] In September 2026, an OpenAI system produced a solution to one of the famous Millennium Prize problems, which has not yet been independently verified. By October, OpenAI claimed its systems had solved hundreds more open problems.

AI is also increasingly being used to accelerate AI research. Anthropic reports that AI now writes more than 80 percent of its code, and its chief executive warns that AI improving itself is starting to happen across the industry. If that continues, progress could speed up further.

What this means for New Zealand

These developments make our recommendations more pressing. Our top three recommendations are to:

- Establish an AI Safety Institute to monitor developments like those above and advise Ministers.
- Prepare an AI Safety Strategy setting out who is responsible for what and how they will respond.
- Endorse the global call for frontier AI to remain under human control.

Our policy papers set out these and other steps.

[1] See the bottom of Appendix F, p. 45: Advanced AI according to 1,580 researchers.