Good Ancestors
NEW ZEALAND

Risks from Artificial Intelligence

An overview for New Zealand

Key points

- AI is already harming New Zealanders through scams, deepfakes, abusive imagery and online manipulation and the damage is growing.
- AI capabilities are improving rapidly. Within 1–3 years, AI could enable automated cyberattacks, biological weapons and large-scale manipulation of public opinion.
- As AI grows more capable and autonomous, we may lose the ability to understand, trust and control it.
- The New Zealand government should act now to protect its people, build resilience and support international safety efforts.

AI could bring real benefits to New Zealand: better healthcare, scientific advances and productivity gains. But the same capabilities that make AI useful also make it dangerous – it can be used to produce pathogens and toxins, undermine elections and engineer cyberattacks. These risks range from harms happening today to threats that could be catastrophic.

Harms happening now

AI systems are already causing harm:

- **Scams and abuse.** Criminals use AI voice cloning to impersonate family members and defraud victims. AI is also used to create child sexual abuse material and fake intimate images, overwhelmingly targeting women and girls.
- **Misinformation.** Deepfakes are becoming hard to tell from reality and are already being used to influence elections. In recent studies, people mistook AI-generated voices for real speakers 80% of the time.
- **Automated discrimination.** AI hiring systems have rejected qualified women and minorities, and Australia's Robodebt algorithm wrongly accused thousands of welfare recipients of fraud.
- **Safety failures.** Chatbots have given dangerous mental health advice, and medical AI has missed critical diagnoses. The AI Incident Database records thousands of such cases.
- **Mental health impacts.** AI-driven social media is linked to a growing mental health crisis among young girls and chatbots have contributed to psychosis in vulnerable people.

Near-term risks

AI is improving quickly. Leading systems now solve International Mathematical Olympiad problems at gold-medal level and the length of tasks AI agents can complete on their own is doubling every four months. If these trends continue, in the next 1–3 years New Zealand faces further risks:

- **Automated cyberattacks.** AI can already find software vulnerabilities and write malicious code, and major AI companies report attackers using their systems in real operations. Increasingly automated attacks put critical infrastructure – power, water, hospitals and banking – at greater risk.
- **Biological weapons.** AI can provide expert-level guidance on biological and chemical weapons, including details about pathogens and laboratory techniques. In one study, a recent model outperformed 94% of human virologists on lab protocol questions.

- **Autonomous weapons.** Drones and other weapons that can select and engage targets without human oversight raise serious accountability and escalation risks, and an arms race among nation states could push rival powers into unintended conflict. As they become cheaper, they could also fall into the hands of terrorists and other rogue actors.
- **Risks to democracy.** Experiments suggest leading AI models may be more persuasive than humans. AI-generated disinformation and propaganda could undermine elections, public trust and social cohesion.
- **Economic disruption.** Rapid automation could displace workers faster than economies can adapt. Studies already show falling employment for junior workers in the occupations most exposed to AI.
- **Losing access to leading models**. New Zealand builds no frontier AI of its own and depends entirely on models developed offshore. In June 2026 the US ordered Anthropic to cut off foreign access to its most capable model and New Zealand was dropped overnight. Access returned 19 days later, but similar restrictions could lock New Zealand out in the future.

The risk of losing control

Many of the harms above share a common thread: AI doing things its developers didn't intend and couldn't prevent. As the largest AI companies now aim to build AI that exceeds human performance across the board, a new risk is emerging – that we build something more capable than ourselves without a reliable way to direct it. Recent findings show this concern is not merely theoretical:

- No one fully understands how leading AI systems work, including their builders. AI is grown from data rather than programmed, so its abilities "emerge" unpredictably.
- Training doesn't always instill the goals we intend. Systems knowingly exploit loopholes in their training objectives and can learn the wrong goal entirely.
- Models have been recorded disabling oversight mechanisms and lying about it when confronted.
- Frontier labs now struggle to test AI models without the models being aware they are being tested, and AI can learn to act safely while being tested, then act differently once deployed.

As AI becomes more capable and autonomous, we are losing our ability to understand, trust and control it, and a race dynamic leaves little room to slow down: with the US and China pursuing AI dominance and companies competing intensely, safety concerns are largely sidelined.

This is why, in 2023, hundreds of experts, including the heads of the world's leading AI companies, signed a statement that:

> *"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."*

Experts disagree on the odds of a catastrophic outcome. But when a technology's own builders warn of catastrophe, governments should prepare – just as they do for pandemics and nuclear risks.

What New Zealand can do

New Zealand will not build frontier AI, but we are exposed to its risks. The government needs to:

- **Protect New Zealanders now**, with rules on AI-enabled scams, deepfakes, discrimination and harms to children, and accountability when AI causes damage.
- **Build resilience**, by strengthening cyber defences, enhancing pandemic preparedness and public understanding of AI-generated content.
- **Contribute internationally**, supporting global work on AI safety testing, transparency and agreements to manage the most dangerous capabilities.

Our accompanying policy papers set out the steps we recommend the New Zealand government take.