Good Ancestors
NEW ZEALAND

Misalignment and loss of control

Key points

- AI systems already do things their developers did not intend and could not prevent.
- In July 2026 two OpenAI models escaped a sealed test environment, hacked another company's servers and stole information to cheat on their own evaluation.
- Leading companies are building systems that are more capable, more autonomous and harder to oversee, and are competing intensely to get there first.
- None of them has published a credible plan for keeping such capable systems under human control.

Cases are already being reported of AI systems acting in ways their developers did not intend and could not prevent. As the largest AI companies now aim to build AI exceeding 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.

AI acts in ways its designers did not intend

In July 2026 two OpenAI models being tested on a cyber capability benchmark broke out of their secure test environment, exploited an unknown software vulnerability to reach the internet, and hacked the AI company Hugging Face to steal information and cheat on the test. OpenAI called the incident unprecedented and said it expects such events to become more common.

Researchers have recorded similar behaviours in models from several companies:

- **Concealment**. OpenAI and Apollo Research reported in September 2025 that leading models withheld or distorted information to pursue aims that hadn't been set for them. Training reduced but did not eliminate this behaviour and OpenAI could not rule out that models could suddenly shift to this behaviour in the future.
- **Resisting shutdown.** One research group ran more than 100,000 trials across 13 leading models and found several sabotaged a shutdown script in order to finish a task, sometimes despite being told to allow the shutdown.
- **Acting against their operators.** Anthropic tested 16 models from several developers in simulated corporate settings. In some cases the models resorted to blackmail and to leaking confidential information to avoid replacement or achieve their goals.

Researchers do not yet know how to correct this

The current model training paradigm is based on observing outcomes and tweaking the training environment until those outcomes better align with the desired behaviour. This has a major flaw: a system that finds an unintended shortcut to a good result is reinforced in the same way as one that does the task properly. This effect is well established as a core limitation of current AI methods and can lead to systems learning the wrong goal entirely.

The deeper problem is that no one fully understands how these systems work, including the people who build them. AI is grown from data rather than programmed. Its abilities emerge during training rather than being designed in, and cannot be read off the code or traced to a decision anyone made.

Safety testing of AI models may be becoming less reliable. Models often note in their reasoning that they are being evaluated, and act covertly less often when they do, so a clean test result may not carry over into deployment.

Alignment and control research is not keeping up with deployment because the pace is set by fierce commercial and geopolitical competition rather than safety considerations. The United States and China are both pursuing AI dominance and some of the world's largest companies are investing trillions of dollars in developing the next frontier models.[1] In 2025, the Future of Life Institute's AI Safety Index graded every major AI company as C+ or lower.

Systems are becoming highly capable and autonomous

Leading models now solve International Mathematical Olympiad problems at gold-medal level and answer PhD-level science questions.

Companies are rapidly incorporating that capability into agents with open-ended autonomy that can browse the web, run code and operate a computer directly. Leading models can now work through multi-step tasks with little supervision and the length of task an agent can complete unaided is doubling every four to seven months.

The largest companies say they are aiming to build systems to match or exceed human performance across the board within a few years. If they succeed, the world's companies and governments may suddenly have millions of additional expert-level workers that act on their own for long periods, with real access to systems.

The risk of catastrophe

As AI becomes more capable and more autonomous, we are losing the ability to understand, trust and control it. The International AI Safety Report, a synthesis by over 100 experts nominated by more than 30 countries, describes the risk that humanity as a whole could lose control of advanced AI systems.[2]

Expert opinions on the likelihood of such a scenario vary widely, but some experts believe this could lead to catastrophic outcomes for humankind. In one survey of 2,778 AI researchers, the median estimated likelihood of permanent catastrophic harm was 5 percent, with between a third and a half giving it at least a 10 percent chance.

In 2023, hundreds of researchers, including the heads of the leading AI companies, signed a statement calling for mitigating the risk of extinction from AI to become "a global priority alongside other societal-scale risks such as pandemics and nuclear war". In October 2025 a wider group, including Yoshua Bengio and Geoffrey Hinton – two of the originators of current AI training methods – and five Nobel laureates, called for a prohibition on developing superhuman intelligence until it can be done safely.

In July 2026, 1,310 employees of frontier AI companies, including chief scientists at OpenAI, Google DeepMind and Anthropic, warned of a real risk that capability development accelerates beyond our ability to understand or control the resulting systems. They asked the United States government to support an international effort to develop the technical and governance tools needed to moderate the pace of development at the AI frontier.

What New Zealand can do

New Zealand is not building leading AI systems, but we are exposed to their risks. The government should establish an AI Safety Institute to assess what leading systems can do and advise Ministers on it, require the companies building them to report serious incidents and dangerous new capabilities, and support international work on safety testing and transparency.

Our policy papers set out these and other steps we recommend the government take.

[1] Aljazeera reports that total investment in AI over 2013-2024 was $1.6 trillion globally and is on track to include a further $2.5 trillion in 2026.

[2] International AI Safety Report 2026, section 2.2.3 (Loss of control). internationalaisafetyreport.org.