

Recommendation: Require leading AI companies to report incidents and risks

Key points

- Mitigating the risks from leading AI systems requires up-to-date and reliable information about them. The New Zealand government currently learns of new risks from media coverage and overseas regulators.
- The companies building leading AI models report safety incidents and dangerous new capabilities to governments in California and the European Union, but not to New Zealand.
- We recommend requiring a few leading AI developers to report critical safety incidents and dangerous capability findings to the New Zealand government, using regulatory frameworks they already comply with overseas.

Why it matters

The government cannot manage risks from AI without timely and credible information. Leading AI companies hold it: safety and threat assessments of their most powerful models and records of misuse detected on their own systems.

Voluntary disclosure to the public is limited and in some cases reliant on whistleblowers or investigative reporting. If a model is found to be capable of enabling attacks on critical infrastructure, hacking government data, or helping malicious actors build biological weapons, the government would learn of it only through its security alliances or via the same channels as the public.

The Five Eyes cyber security agencies, including the Government Communications Security Bureau, [warned in June 2026](#) that leading AI models are creating serious new threats requiring urgent government action. Compulsory reporting would let policymakers see such capabilities emerging and act before serious harm occurs.

What we're proposing

We recommend requiring several developers of the world's most capable AI systems to report the following to a designated New Zealand agency:

- **Critical safety incidents:** Serious harm caused, or almost caused, by an AI system, including physical injury or death, damage to critical infrastructure, environmental harm or a cyber security breach.
- **Dangerous capability findings:** Results of safety evaluations showing an AI system has crossed its developer's safety framework thresholds for chemical, biological, nuclear or offensive cyber capabilities.
- **Mitigations:** How the developer prevents misuse, such as alignment techniques, safeguards, monitoring and organisational security, including cases where those measures failed or were not adopted.

The opportunity for New Zealand

New Zealand does not build leading AI systems and has limited capability to assess them. Mandatory reporting is the cheapest way to get early warning of risks that will reach New Zealand, and using requirements in force overseas means implementation and compliance would be quick and low cost.

To ensure New Zealand can act on the information it receives, we also recommend [establishing an AI Safety Institute](#) and [increased AI safety and security funding](#), as set out in our companion briefs.

How it could be done

Mandatory incident reporting is already used in New Zealand to manage major industry risks. Pilots [notify accidents](#) to the Civil Aviation Authority, businesses notify [serious workplace events](#) to WorkSafe, and medicine sponsors report [serious adverse reactions](#) to Medsafe.

The reporting obligation could be limited to a handful of the largest overseas developers by setting thresholds tied to revenue and the computing power used to train a model. Requiring only information those developers already provide to Californian and European regulators would keep compliance costs close to zero. The duty should carry a civil penalty for failure to report and confidentiality protections for sensitive commercial and security material.

It could be implemented as a short standalone Act or an amendment, perhaps modelled on the notifiable privacy breach scheme in Part 6 of the Privacy Act 2020, which defines a trigger by serious harm, requires notification to a named regulator and reaches overseas companies without a local New Zealand presence.

Thresholds and definitions should sit in regulations, so they can be adjusted as overseas requirements and circumstances change. Reports should go to an agency that can act on them, such as the National Cyber Security Centre or an [AI Safety Institute](#). In the future, the responsible organisation could also expand its role to accept voluntary reports from New Zealand businesses and researchers, as recommended by [Stanford's RegLab](#).

International precedent

Industry accepts the principle: in March 2025 the Frontier Model Forum members, including Anthropic, Google DeepMind and OpenAI, agreed to share information on vulnerabilities, threats and capabilities of concern, and the Forum is now [exploring information sharing with government partners](#).

California's [Transparency in Frontier Artificial Intelligence Act](#) has been in force since 1 January 2026. Developers are required to report critical safety incidents to the state's Office of Emergency Services within 15 days, or 24 hours where there is an imminent risk of death or serious injury. Penalties reach \$1 million USD per violation.

Under the European Union's [Code of Practice for General-Purpose AI](#), signatories including OpenAI, Anthropic, Google and xAI report serious incidents to the European AI Office within two days for disruption to critical infrastructure and 15 days for serious harm to health or property.

New York's RAISE Act takes effect on 1 January 2027 with a 72-hour reporting window, and Vietnam's AI Law has required serious incidents to be reported through a national portal since March 2026.

Frequently asked questions

Would overseas companies actually comply? Yes. The Privacy Act 2020 already binds overseas companies operating in New Zealand and the largest AI developers comply with it. They already produce these reports for other regulators, so sending a copy to New Zealand adds almost no additional cost.

Why legislate rather than seek voluntary commitments? Voluntary reporting depends on a firm's goodwill and can be dropped if competitive pressure changes. A statutory duty with a civil penalty would ensure the reliable flow of information that New Zealand decision-makers need.