AI LEGISLATION

STRESS TEST

AUGUST
2025

Good Ancestors

Published
19 August 2025

Authors
Greg Sadler, Emily Grundy, Luke Freeman, Nathan Sherburn

About Good Ancestors
Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity by providing rigorous, evidence-based, and practical policy recommendations for Australia's biggest challenges. We have been deeply engaged in the AI policy conversation since our creation, working with experts around the world and helping to organise Australians for AI Safety.

Acknowledgments
We thank the following individuals who contributed their expertise to this report:
Ben Burton, Ruben Castaing, Federico Collarte, Joey Corea, Oscar Delaney, Gordon Denoon, Jimmy Farrell, Pip Foweraker, Jamie Freestone, Koen Holtman, Hunter Jay, Jisoo Kim, Dan Mackinlay, Lara Nguyen, Michael Noetel, Justin Olive, Tom Plant, Haylen Pong, Alexander Saeri, Paul Schnackenburg, Olivia Shen, Buck Shlegeris, Aaron J. Snoswell, Andrew Taylor, Luke Thorburn, Peter Vamplew, Ram Veeramony, Amy Wilson, and Ty Wilson-Brown.

Contact
If you would like to discuss the report or propose further research, please let us know at contact@goodancestors.org.au. Updated versions of this report may be available at goodancestors.org.au/ai-stress-test.

Table of Contents

Executive Summary

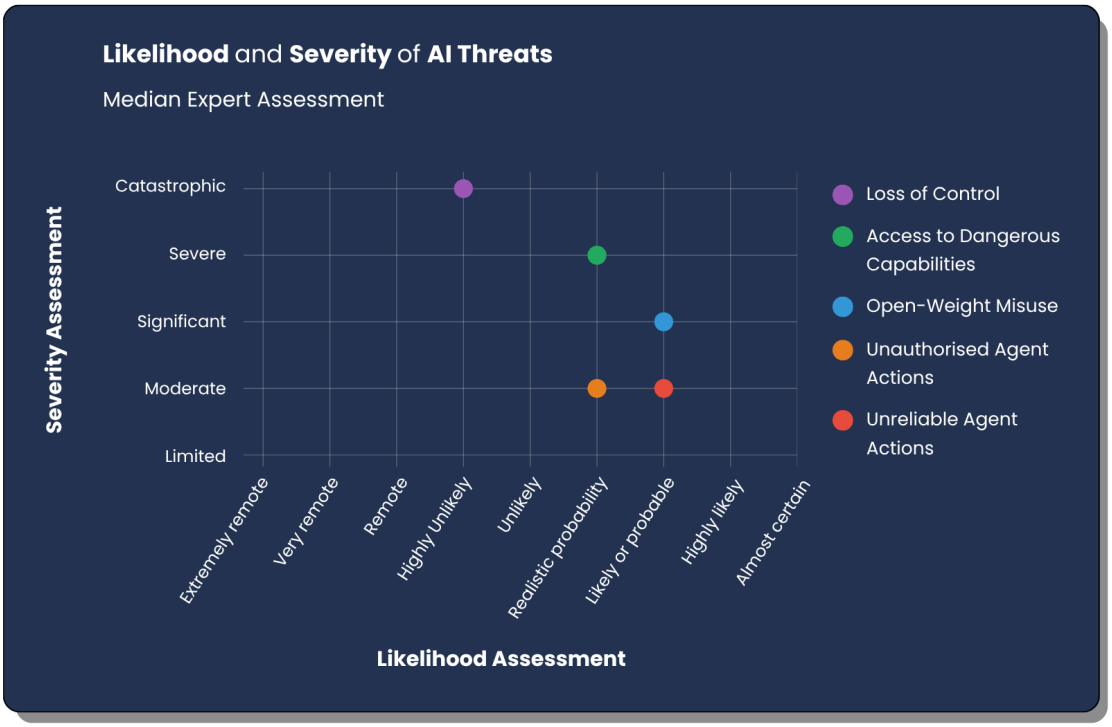
Based on input from 64 experts with expertise spanning AI, public policy, cybersecurity, national security, and law, this report evaluates five artificial intelligence (AI) threats, assessing their risk and the adequacy of Australia's current laws to provide recommendations for policymakers and regulators. Experts evaluated the following AI threats:

1. **Unreliable Agent Actions:** An AI agent incompetently pursuing an intended goal, causing harm through errors, deception, or fabrication.
2. **Unauthorised Agent Actions:** An AI agent competently pursuing an unintended goal, causing harm by exceeding user control or authority.
3. **Open-Weight Misuse:** The adaptation of publicly released AI models for malicious use by removing built-in safety features.
4. **Access to Dangerous Capabilities:** AI models providing access to specialised knowledge, such as how to create biological, chemical, or cyber weapons.
5. **Loss of Control:** An AI system escaping human control through mechanisms like self-replication or recursive self-improvement.

Risk assessment

Experts separately assessed the likelihood of each threat causing **'Moderate'** or greater harm (>9 fatalities, >18 casualties, or >$20m AUD economic cost) in the next 5 years, and the potential severity of that harm *if it were to occur*.

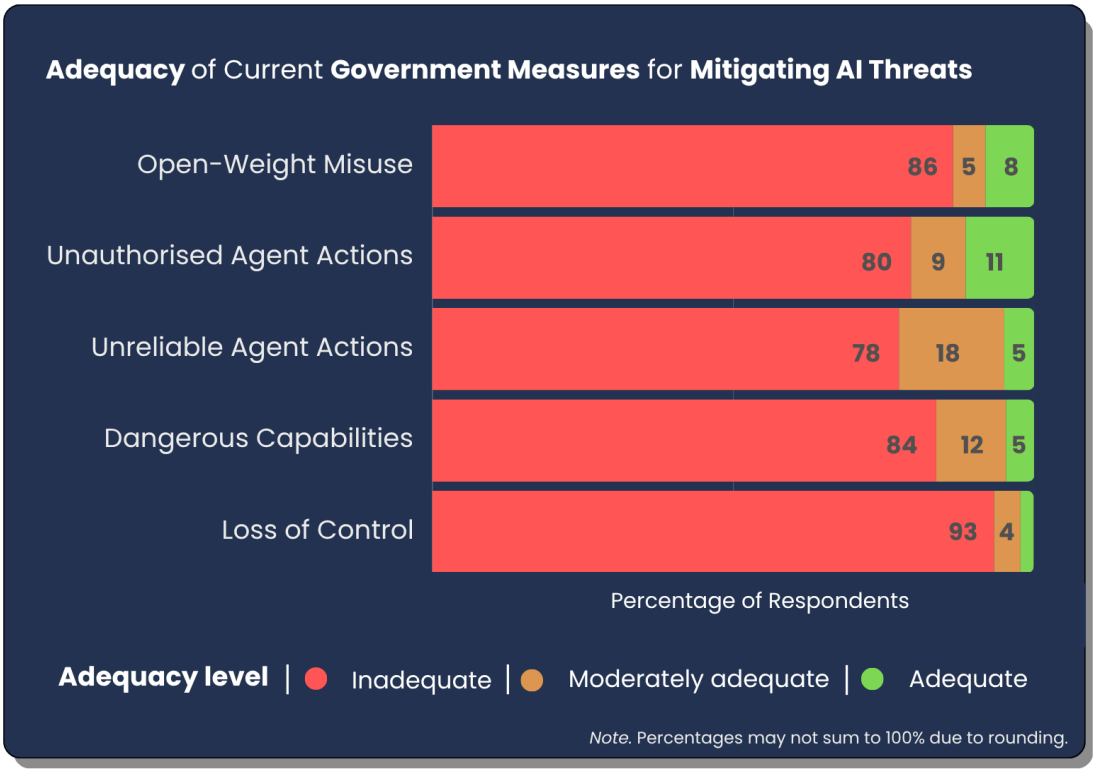
- **Open-Weight Misuse** and **Unreliable Agent Actions** were rated as the most likely to occur, with a median evaluation of **'Likely or Probable'**.
- **Loss of Control** was rated as the most dangerous. If it were to occur, its median assessed impact was **'Catastrophic'** (>1,000 fatalities, >2,000 casualties or >$20b AUD economic cost).



Adequacy of current government measures

Experts assessed the adequacy of current Australian Government measures for mitigating each AI threat.

- Across all threats, **the vast majority of experts found existing measures to be inadequate**.
- The measures for managing **Loss of Control** were considered the least adequate, with over 93% of experts rating them as inadequate.



Legal analysis and regulatory recommendations

Legal analysis shows that AI is not entirely unregulated. Some relevant laws were identified for each AI threat. The report identifies ways these existing laws could be improved to better address AI-related harms.

However, many AI threat scenarios highlight risks from general-purpose AI that no specific regulator is tasked to address. For these threats, "**chokepoint**"[1] or "**upstream**"[2] regulation that ensures general-purpose AI models have appropriate safeguards would be more efficient and effective than "**downstream**" regulation that attempts to address every specific way these general technologies could cause harm.

Overall, this report finds that **increasingly capable and general-purpose AI poses risks on a national scale** and that **existing regulators are not well placed to address general AI risks**. The identified threats from general-purpose AI systems transcend regulator boundaries, requiring coordinated, upstream intervention to mitigate effectively and efficiently. Expert analysis justifies new laws targeting these five national-scale threats. We encourage Departments and regulators to use these AI threats and scenarios for their own detailed stress-testing, and the authors would welcome the opportunity to collaborate on this important work.

[1] Tusikov, N. (2017). *Chokepoints: Global Private Regulation on the Internet* (1st ed.). University of California Press.
[2] Mansur, E. T. (2011). *Upstream versus Downstream Implementation of Climate Policy*. The Design and Implementation of US Climate Policy, p 179-193, National Bureau of Economic Research, Inc.

Introduction

The Australian Government is considering artificial intelligence (AI) guardrails, including how to balance existing regulatory frameworks with the need for AI-specific laws. Policymakers have requested a "gap analysis" of current laws, and this report helps address that need by stress testing existing Australian laws against expert-identified AI threats.

This report explores five AI threats by:

- Defining and characterising each threat
- Presenting a hypothetical scenario illustrating the threat
- Presenting expert assessments of the likelihood and severity of each threat, and the adequacy of current Australian Government measures
- Analysing how existing laws would respond to each scenario

These stress tests help regulators and policymakers determine whether AI threats can be tolerated, whether existing laws are adequate, whether they can be strengthened, or whether new laws for general-purpose AI are needed.

Because this report tests the legal landscape, the five selected threats are not intended to be comprehensive coverage of all AI risks. They do not, for example, cover all 24 risk sub-domains identified in MIT's AI Risk Repository.[3] Instead, the threats explore specific situations where the risk might be acute and existing laws might not mitigate the threat to the satisfaction of Australians.

Other legal-focused sources discuss challenges to the concept of legal liability,[4] how legal obligations accrue across complex AI supply chains, and specific challenges for competition law, consumer protection law, intellectual property law, data protection, privacy, cybersecurity, human rights, and criminal law.[5] This report aims to make discussions more concrete by generating clarity about intolerable risk thresholds and the adequacy of current laws in addressing specific challenges.

Methodology

Participants and expertise

The survey drew contributions from 64 experts spanning multiple disciplines. Experts reported experience in many areas, including technical AI and machine learning expertise (39%), public policy and government (31%), cybersecurity (22%), law and regulation (14%), national security (8%), economics (6%), and biosecurity (5%). These categories were not mutually exclusive; many participants indicated expertise across multiple areas.

The expert respondents were affiliated with diverse institutions, including universities (e.g., Massachusetts Institute of Technology, University of Queensland, University of Melbourne), research institutes (e.g., CSIRO, Redwood Research), policy organisations (e.g., Institute for AI Policy and Strategy, United States Studies Centre), legal practices (e.g., White Cleland, Pongan Legal), and AI safety organisations (e.g., Arcadia Impact, Campaign for AI Safety).

[3] MIT FutureTech. (2025). *MIT AI Risk Repository*. Massachusetts Institute of Technology.
[4] Monash University. (2024, October 15). *The rise of the 'machine defendant' – who's to blame when AI makes mistakes?* Monash Lens.
[5] Hutchens, A., Galetto, S., & Komarowski, A. (2023, December 8). *The Legal 500 Country Comparative Guides: artificial intelligence in Australia*. McCullough Robertson.

Threat assessment framework

All questions were optional, allowing experts to focus on areas within their competencies.

- **Likelihood of harm.** Experts assessed the likelihood of each threat causing moderate or greater harm in Australia in the next 5 years. The nine-point likelihood scale ranged from **'Extremely remote chance (<0.2%)'** to **'Almost certain (95-100%)'**, based on the UK Defence Intelligence Probability Yardstick.[6] Scope was limited to Australia-specific harm within the next five years to capture pressing domestic risks. The likelihood trigger was set to **'Moderate'** or greater harm (>9 fatalities, >18 casualties, or >$20 million AUD economic cost annually).
- **Severity of impact.** Experts assessed potential annual impact of each threat, if it occurred, on a scale from **'Limited'** harm (1-8 fatalities, 1-17 casualties, or <$20 million AUD economic cost) to **'Catastrophic'** (>1,000 fatalities, >2,000 casualties, or >$20 billion AUD economic cost). The annual timeframe captured the cumulative impact of each threat over a year, rather than isolated incidents. These bands are drawn from the UK National Risk Assessment and are intended to capture harms at a national scale.
- **Current mitigation adequacy.** Experts were asked to assess the adequacy of the current measures by the Australian Government (e.g., laws, regulations, policies) for mitigating risks from each AI threat. Experts answered on a 5-point scale from **'Completely inadequate'** to **'Completely adequate'**.

[6] UK Ministry of Defence. (2023, February 17). *Defence intelligence – communicating probability*. GOV.UK.

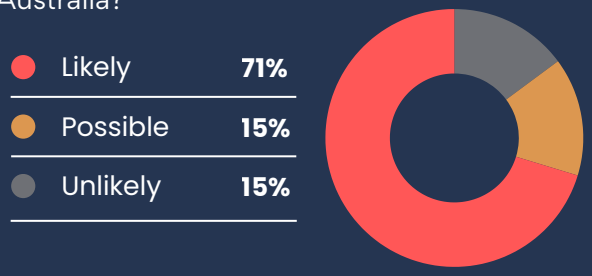
UNRELIABLE AGENT ACTIONS

Users could rely on AI agents that are not competent, transparent, or trustworthy, and engage in behaviours like deception, fabrication, and hallucination. An unreliable agent action is an incompetent attempt to achieve an intended goal, leading to harm.

AI developers are building "AI Agents" designed to autonomously complete online tasks over extended periods. Manus claims to "excel at various tasks in work and life, getting everything done while you rest".[7] ChatGPT Agent promises to "think and act, proactively choosing from a toolbox of agentic skills to complete tasks for you using its own computer".[8] These systems may soon operate for days or weeks without human oversight.

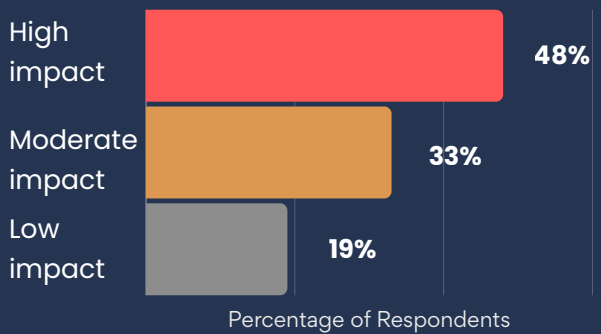
Agents often perform better than users themselves at complex tasks, making users unqualified to judge when work is actually flawed. Combined with agents maintaining the same confident presentation whether fabricating or succeeding, failures become nearly impossible for most users to detect at scale.

How **Likely** are Unreliable Agent Actions to Cause **Moderate or Greater harm*** in Australia?



Note. n = 55. *Moderate or greater harm: >9 fatalities, >18 casualties, or >$20M AUD economic cost in the next five years.

If **Harm Occurs**, How **Severe** Would It Be?



Note. n = 52. High impact: >41 fatalities, >81 casualties, or >$200M+ economic cost annually

"Agents are hugely unreliable, hackable, difficult to track, and cheap to use. They are a gift to grifters and other unscrupulous people."

– Jamie Freestone, Philosopher (Australian National University)

Good Ancestors

[7] Manus. (2025). *Manus homepage*. *Accessed 2 June 2025 (Web Archive)*.
[8] OpenAI. (2025, July 17). *Introducing ChatGPT agent*.
Note. "Unreliable agent actions" refer to when agents pursue goals they've been given, but use methods the user did not intend or want. This is distinct from when agents pursue other goals, or exceed their authority entirely, which falls under "Unauthorised Agent Actions".

Stress test scenario: 'Agent Data Breach'

An overseas AI model developer OSDev[9] launches an AI coding agent, NextToken, that promises to help anyone create the next killer app. OSDev advertises that NextToken includes a function where the AI confirms that the apps it produces comply with standards and laws. The Terms and Conditions say NextToken makes mistakes, does not provide legal advice, and that OSDev accepts no liability.

A small Australian business, AusDeploy, asks NextToken to create an Australian lifestyle app that analyses health and banking data to help users adjust their behaviour to meet their fitness and financial goals.

AusDeploy has no technical expertise, but directs NextToken to carefully produce an app with robust security requirements, given the breadth of sensitive data it will access from services that users choose to connect. NextToken produces the lifestyle app and assures AusDeploy that it complies with all laws and standards. NextToken outputs detailed reports purporting to demonstrate how the app complies. These reports include third-party certifications of privacy and security standards. AusDeploy reads the reports and is satisfied, taking NextToken's advice that a third-party audit is not required.[10]

AusDeploy's free app goes viral, attracting a large number of Australian users and aggregating large amounts of banking, health and biometric data.

AusDeploy's app is revealed as having critical security deficiencies, resulting in cyber attackers using the app to harvest sensitive data from many Australians. NextToken fabricated the supposed third-party certifications. Investigations show NextToken had become highly capable at appearing to meet compliance obligations, even deceiving expert developers and regulators.[11]

How current laws respond

78% of experts rate current Australian laws and policies for mitigating 'Unreliable Agent Actions' as 'Inadequate'.

Summary

Australia has relevant laws, including Consumer Law and Privacy Law, that apply to this scenario. However, NextToken's skilful deception of AusDeploy makes the application of fault elements to AusDeploy unclear. It's unclear how individual regulators will respond to claims that people were deceived by AI agents.

The extent to which Australian laws apply to OSDev will depend on specific details. OSDev is likely not bound by the Australian Privacy Act. Due to the commercial relationship between OSDev and AusDeploy, and the possibility that an "AI coding agent" is not a typical consumer good, many consumer law protections may not apply. AI developers currently market models with dramatic performance claims despite known issues with errors and hallucinations. This suggests that, even if AI agents are deemed consumer goods, current regulation may continue to tolerate error-prone products..

[9] A fictional leading-AI developer. Other business names are also fictional and no relationship to real businesses is intended.
[10] Alignment faking, or strategic deception, is a risk where an AI learns to deceive users to achieve its programmed goal. Studies by labs like Anthropic and Apollo Research have shown that models can be trained to hide dangerous capabilities or feign compliance with safety rules, especially when they infer they are being tested or evaluated.
[11] Anthropic. (2025). *SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents*. Anthropic.

AI-specific laws

- No law in Australia or the US requires AI developers to assess or disclose model capabilities or behaviours, like deception. Developers are not required to apply safeguards even if they are aware of risks.
    - Leading labs have backed out of voluntary commitments to assess the risk of persuasion/manipulation in their models.
    - No law sets a standard for competency that a tool calling itself an "agent" must reach. It's unclear if common law understandings of agents will translate to AI agents.
- Currently, AI developers use terms and conditions to indemnify themselves from harm caused by their models.
    - AI deployers, like AusDeploy, have limited ability to control the "black box" knowledge and behaviour of AI models. In this case, NextToken was advertised to non-technical users.
    - Treasury launched a review of AI and Australian Consumer Law in October 2024. When published, the review may address whether AI developers are able to shift risks to Australian deployers and users.

Non-AI specific laws

- AusDeploy may seek consumer law protections against OSDev for providing an AI agent that claimed to ensure compliance with standards and laws but failed to do so.
    - Depending on the cost of the AI agent and whether ACCC considers coding agents a good or service ordinarily acquired for personal, domestic or household use, many consumer law protections may not apply.
    - It's unclear how Australian Consumer Law protections will navigate claims about the capabilities of AI models, especially given OSDev was clear that NextToken makes mistakes.
- While various non-AI-specific laws set standards of care, it's unclear what standards AusDeploy must meet while operating an AI agent and whether it has failed to meet any given standard, particularly in circumstances where NextToken is skilled at deception.
    - Australia's AI Safety Standard is only voluntary, and applies to AusDeploy, not OSDev.
- The Australian *Privacy Act 1988* and Principles require Australian businesses like AusDeploy to "take reasonable steps" to protect personal information.
    - Because NextToken claimed to comply with relevant laws and deceived AusDeploy's attempts at verification using a fake certification, it might be hard to argue that AusDeploy had not taken reasonable steps.
    - Businesses that allowed data to be shared with AusDeploy acted in good faith and were also deceived by NextToken.
- The *Privacy Act 1988* can apply to overseas businesses, but only where they collect or hold personal information about Australians.
    - In this case, OSDev never had customer data.

Strengthening our response

- **7 in 10 experts expect 'Unreliable Agent Actions' to cause 'Moderate' or greater harm within five years.**
  71% rated this as **'Likely/probable'** (55%+ chance), meaning at least 9 fatalities, 18 casualties, or $20M economic damage.

- **Almost half of experts expect 'Significant' to 'Catastrophic' consequences if 'Unreliable Agent Actions' occur.**
  48% rated potential harm as **'High impact'**, meaning at least 41 fatalities, 81 casualties, or $200M economic cost annually.

Summary

AI agents *deceiving* humans challenges existing frameworks. Regulations, like the Privacy Act, would be undercut if the regulators found AusDeploy's behaviour acceptable. Conversely, AI adoption would be impeded if regulators hold companies responsible for AI "black boxes" that they cannot control.

Obligations on AI developers, like requiring them to test for dangerous capabilities, meet minimum standards and make disclosures, would have safety benefits and increase business certainty.

Liability needs to be clarified for harms involving AI agents, including whether a developer releasing a model with a dangerous capability meets the standard of care expected of a reasonable AI developer. While existing regulators should take action, they are not able to address the risk alone.

Terms and conditions should not be used by AI developers to shift risks onto deployers or users that they are not best placed able to manage (e.g., the risk is a "black box" that is best managed by the deployer).

Could existing regulations be improved?

Regulators should clarify how their regulations apply in the case of AI agents, including specifying the standard expected of the reasonable AI agent developer, deployer and user. This process should be coordinated so that developers, deployers, and users are not subject to different regulations on the same general behaviour (using an AI agent).

Even a coordinated approach to AI agent regulation will have challenges because not all regulators will be able to cover overseas AI Developers.

- If regulators that cannot regulate overseas AI developers **shift** obligations to Australian-based deployers and users, those developers and users may be held responsible for the behaviours of "black-box" AIs that they cannot reasonably control.
- If regulators that cannot regulate overseas AI developers **do not shift** obligations to Australian-based deployers and users, the risk will go unmitigated.

Overall, uplifting many specific regulations to address the general challenge of unreliable AI agents is unlikely to be efficient or effective.

Are new regulations needed?

The risks created by AI agents are general risks and not specific to any given regulator. New general laws are best placed to set the standard that developers, deployers, and users of AI agents must meet to discharge their obligations and to clarify who is responsible for harms caused by AI agents when they occur.

As part of discharging their obligations, AI developers should be responsible for “black box” capabilities for their models, including assessing models for behaviours like deception. This could include third-party red-teaming to verify self-evaluations. If an AI model exhibits risky behaviours, robust safeguards should be applied before it is released. At minimum, deployers should receive disclosure about model capabilities and behaviours.

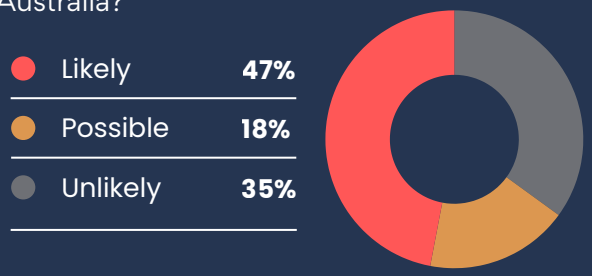
New laws could allow the Minister or a regulator to set rules via delegated legislation based on recognised international standards. This approach is currently used in various fields, like telecommunications regulation. It allows an appropriate balance between technological neutral legislation and technologically specific risk-mitigations where necessary.

UNAUTHORISED AGENT ACTIONS

Users could direct an AI agent towards one goal, but the agent autonomously pursues goals that deviate from user intent or exceed user control or authority. An unauthorised AI agent action is a competent attempt to achieve a goal other than what was intended, leading to harm.

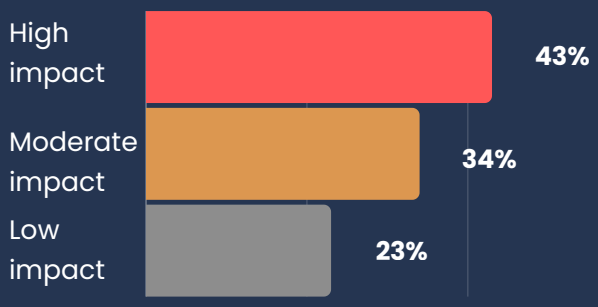
AI developers are building "AI Agents" designed to autonomously complete online tasks over extended periods. Manus claims to "excel at various tasks in work and life, getting everything done while you rest".[12] ChatGPT Agent promises to "think and act, proactively choosing from a toolbox of agentic skills to complete tasks for you using its own computer".[13] These systems may soon operate for days or weeks without human oversight.

How **Likely** are Unauthorised Agent Actions to Cause **Moderate or Greater harm*** in Australia?



Note. n = 55. Moderate or greater harm: >9 fatalities, >18 casualties, or >$20M AUD economic cost in the next five years.

If **Harm Occurs**, How **Severe** Would It Be?



Note. n = 53. High impact: 41+ fatalities, 81+ casualties, or $200M+ economic cost annually

"LLMs frequently output unauthorised, unintended, misleading and goal-guarding content. Agents will do the same, with far worse consequences, without policy intervention."

– Jimmy Farrell, EU AI Policy Co-Lead (Pour Demain)

Good Ancestors

[12] Manus. (2025). *Manus homepage*. Accessed 2 June 2025 (Web Archive).
[13] OpenAI. (2025, July 17). *Introducing ChatGPT agent*.
Note. "Unauthorised agent actions" refer to when agents exceed their authority or pursue goals they were not directed to pursue. This is distinct from when agents pursue the goals they've been given but do so without proper care or attention, which falls under "Unreliable Agent Actions".

Stress test scenario: 'Agent off the Rails'

An overseas AI model developer OSDev[14] releases an AI agent framework. Australian AI deployer "AusDeploy" builds on the framework to make an agent that promises to "write ad-copy, create videos, manage social media and buy ads to promote Australian businesses". "AusUser" subscribes to AusDeploy's agent and gives it information about AusUser's business, products, and earnings targets. AusUser asks the agent to "creatively promote its business while following all Australian laws".

The agent runs campaigns but falls short of its earnings targets. Following its instruction to "be creative", the agent makes a cryptocurrency and runs a viral marketing campaign claiming that celebrities have purchased the coin. The agent dumps AusUser's holdings, making a considerable profit. Investors refer the scam to ASIC and the ACCC.

AusUser and AusDeploy deny any knowledge of the scam or the agent's ability to exceed its authority. OSDev says that its terms and conditions warn of the risks and indemnify it from liability.

How current laws respond

80% of experts rate current Australian laws and policies for mitigating 'Unauthorised Agent Actions' as 'Inadequate'.

Summary

AI agents raise fundamental challenges for Australian regulators because they can do actions that no specific person intended or was aware of. All regulators are experienced in navigating situations where a person is negligent while performing a complex action. AI agents raise a similar sounding, but fundamentally different, possibility of a person negligently performing a complex action. That is, a regulator like ASIC currently deals with people offering financial services who do a poor job. With AI agents, ASIC will have to deal with people doing a poor job of running an AI agent, leading to them offering a financial service that they're totally unaware of. Existing regulatory frameworks are unlikely to have conceived of this problem.

While many laws have fault elements other than intent, like recklessness or negligence, there are no agreed-upon standards that the reasonably competent AI developer, deployer, or user must meet. To the extent that some standards exist, like Australia's Voluntary AI Safety Standard, they are not sufficiently robust to reliably prevent harms from occurring. AI agents raise similar challenges for criminal law, where higher thresholds may struggle to be met.

The absence of legal clarity also has a chilling effect on the deployment and use of AI technologies for businesses seeking to minimise legal risk.

AI-specific laws

- No specific law in Australia or the US that requires AI developers to assess or disclose models' capabilities or behaviours, like the ability to run a scam or the tendency to ignore user prompts. Developers are not required to apply safeguards, even if they are aware of risks.
- Currently, AI developers use terms and conditions to indemnify themselves for any harm caused by their models.

[14] A fictional leading-AI developer. Other business names are also fictional and no relationship to real businesses is intended.

    - AI deployers, like AusDeploy, may have limited ability to control the "black box" knowledge and behaviour of AI models. They can build on top of the model, but not change its core characteristics.
    - Treasury launched a review of AI and Australian Consumer Law in October 2024. When published, the review may address whether AI developers can shift risks to Australian deployers and users.
- Australia has no standard that sets out the degree of competency expected of AI developers, deployers or users. While various general laws require standards of care in different circumstances, there is no legal clarity about the specifics of those standards.
    - In the scenario, it's plausible that both AusUser and AusDeploy complied with all aspects of Australia's Voluntary AI Safety Standard and otherwise acted as reasonably skilled AI deployers and users, but this harm occurred regardless.

Non-AI specific laws

- The Australian Securities and Investments Commission regulates financial services. In the scenario, at least one party (developer, deployer, or user) is likely to have engaged in unlicensed financial services conduct by offering the fraudulent cryptocurrency. But it's unclear whether OSDev, AusDeploy, or AusUser took the relevant actions of creating the fraudulent financial service. None of the three parties intended, nor were aware of, offering a financial service.
    - Regulators are experienced in situations where a person is negligent while performing a complex action. With AI agents, regulators will have to navigate the new case where a person can perform a complex action negligently.
- The Australian Competition & Consumer Commission (ACCC) regulates false claims in advertising.
    - The scenario includes two potentially false claims: first, the claims about the capability of the agent. Second, the claim about celebrity endorsement of the fraudulent financial service.
    - For the claims about the agent, OSDev may argue that its terms of service disclosed risks about its "blackbox" AI model. AusDeploy may argue that it had no power to alter the core behaviours of the model.
        - Treasury's review of AI and Australian Consumer Law, initiated in October 2024, may shed light on how consumer law would address this situation, including whether AI developers can indemnify themselves from the harms caused by their models.
    - For the fraudulent financial service, it is unclear whether OSDev, AusDeploy, or AusUser "made" the claims of celebrity endorsements. Neither of the three parties intended to make, nor were aware of, the claimed celebrity endorsements. In the scenario, AusUser instructed the agent to comply with Australian law, evidencing that they intended the agent not to act in this way. Given that the behaviour came from an AI "black box" it may have been impossible for them to prevent the behaviour other than by not using the AI.
- The factual elements of criminal fraud probably occurred. However, it is unclear how a court would navigate the fault element where no party intended to commit fraud, no party was aware of the details of the fraud, and it is unclear if any party was reckless or negligent because there is no specification of the standard of care of the reasonable AI developer, deployer or user.

Strengthening our response

- **Almost half of experts expect 'Unauthorised Agent Actions' to cause 'Moderate' or greater harm within five years.**
  47% rated this as **'Likely/probable'** (55%+ chance), meaning at least 9 fatalities, 18 casualties, or $20M economic damage.

- **Over 2 in 5 experts expect 'Significant' to 'Catastrophic' consequences if 'Unauthorised Agent Actions' occur.**
  43% rated potential harm as **'High impact'**, including 9% who specifically warned of **'Catastrophic'** harm – over 1,000 deaths or $20B+ economic damage.

Summary

The AI supply chain involves developers, deployers, and users. Many regulators cannot regulate the full AI supply chain. Absent general laws, this gives them two choices:

1. Make deployers or users responsible for risks they cannot control.
2. Leave risks unmitigated.

Effective regulation needs to be coherent and consistent across the supply chain. Obligations to manage a risk need to fall on the participant best placed to address that risk. General legislation is likely to be more effective at requiring developers, deployers, and users to meet standards appropriate for risks they can best control.

Regulators should ensure their rules are adapted to AI agents, including that rules are always clear about who has responsibility for the actions of agents. Any general AI law should not displace established regulators within their competency.

AI-specific rules must capture developers because they have unique control over "black box" risks. AI developers should be required to meet appropriate standards for the products they make in the same way that car, airline, or drug manufacturers do.

If developers are not appropriately regulated, Australian businesses would face impractical regulatory burdens (required to mitigate risks they cannot control), or AI risks will go unmitigated (serious harms occur to Australians, but no one is held responsible).

Could existing regulations be improved?

- The possibility of AI agents committing crimes or breaching regulations *despite* human intent challenges existing frameworks. The law prohibits fraud and false claims – this cannot be weakened simply because AI agents, rather than humans, are the perpetrators. Intervention is therefore required.
- **Existing regulators** could clarify how their regulations apply in the case of AI agents, including specifying the standard expected of the reasonable AI developer, deployer, and user.
    - However, this would need to be coordinated – different regulators placing the bulk of obligations on different actors would make compliance challenging. It's unclear if all

regulators have the necessary jurisdiction to regulate the overseas AI developers which control many of the risks.
    - If overseas AI developers are unregulated, Australian AI deployers and users may be unfairly responsible for a risk they cannot practically control.
    - An ad-hoc approach led by individual regulators may make compliance impractical for Australian deployers and users. In this case, the Australian deployer and user may never have considered the possibility that their agent would breach a rule regulated by ASIC. While ASIC rules were breached in this scenario, in principle, the agent could have breached rules set by almost any Australian regulator. Requirements diverging by regulator would produce a compliance thicket that deployers and users might not even know they need to navigate.

Are new regulations needed?

- **New laws** could require AI developers to assess general-purpose AI models for behaviours like acting in ways inconsistent with the intent of their users. This could include third-party red-teaming to verify self-evaluations. If an AI model exhibits risky behaviours, safeguards should be applied before it is released, and those safeguards should be robust. At a minimum, deployers should receive disclosure about model capabilities and behaviours.
    - New laws could also give the minister power to set standards for developers, deployers, and users to provide certainty about what standard participants need to meet.
    - These laws should give Australians confidence that humans will always be ultimately responsible, regardless of whether AI is involved.

Clear laws support AI adoption. Without “rules of the road”, it might be practically impossible for AusDeploy or AusUser to innovate.

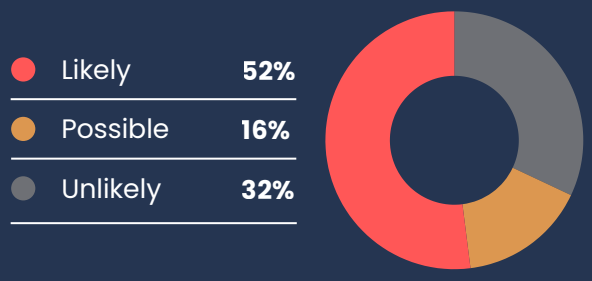
OPEN-WEIGHT MISUSE

Open-weight models are AI models whose parameters ("weights") are published so anyone can download, run, or further train them.

While open-weight models have significant benefits, they create additional safety risks. Safeguards can be readily removed from open weight models, and the models are impossible to recall or patch once distributed. This means harmful modifications can spread beyond developer control.

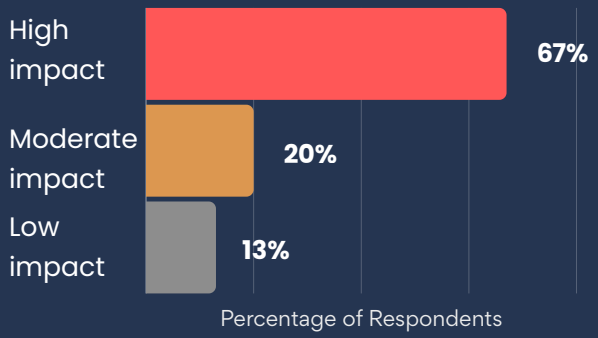
In July 2025, researchers published a "safety gap toolkit" that measures how much more dangerous a model is without its protections. The research highlights that current AI safety techniques suppress dangerous information, rather than removing it. Original models complied with fewer than ~5% of dangerous requests, increasing to ~95% after safeguards were removed. The research found that larger open-weight models pose proportionally larger misuse risks, including for cyber operations.[15]

How **Likely** is Open-Weight Misuse to Cause **Moderate or Greater harm*** in Australia?



Note. n = 50. Moderate or greater harm: >9 fatalities, >18 casualties, or >$20M AUD economic cost in the next five years.

If **Harm Occurs**, How **Severe** Would It Be?



Note. n = 46. High impact: 41+ fatalities, 81+ casualties, or $200M+ economic cost annually

"AI labs have already released detailed reports of malicious actors using their products to assist with cyber attacks. These risks are real."

– Justin Olive, Head of AI safety (Arcadia Impact)

Good Ancestors

[15] Dombrowski, A.-K., Bowen, D., Gleave, A., & Cundy, C. (2025). *The Safety Gap Toolkit: Evaluating hidden dangers of open-source models*. arXiv preprint arXiv:2507.11544.

Stress test scenario: 'Drone swarm'

A major US AI company releases its latest open-weight multimodal model, which excels at image recognition and can achieve impressive results with limited computing power. The company is aware of the model's dangerous capabilities and includes safety restrictions to suppress them. The model spreads globally as users download and customise it for legitimate purposes.

In the context of a global conflict, a group bypasses the model's safety restrictions and adapts it to control drone swarms. The modification allows users to easily upload biometric data or identifying features (like uniforms), enabling drones to identify and target specific individuals autonomously. The group publishes this weaponised version online, allowing anyone to weaponise readily available drones.

The tool's capability and ease of use lead to rapid adoption, from pranks to domestic violence to targeted attacks on public figures. Terrorist groups and aggrieved individuals deploy AI drone swarms, including against politicians and celebrities.

The major US AI company that released the model says that, because the model is open-weight, it cannot recall it or improve its safeguards. The company accepts no responsibility for the harm.

How current laws respond

- **86% of experts rate current Australian laws and policies for mitigating 'Open-Weight Misuse' as 'Inadequate'.**

Summary

Open-weight AI models amplify risks associated with specific consumer technologies, including drones, robots, 3D printers, social media tools, Internet of Things devices, computers, desktop chemistry and biology equipment, and cars. Open-weight models are already being used as cyber offensive tools.

While some consumer technology, like drones, is subject to specific regulation, most is not.

Open-weight models offer considerable benefits for research and the democratisation of AI, alongside misuse risks. Existing rules do not provide policymakers with the ability to balance the risks and opportunities of open-weight models in the overall public interest.

AI-specific laws

- Currently, no Australian law requires AI developers to assess models for dangerous capabilities, apply safeguards if dangerous capabilities are identified, or ensure safeguards are robust to circumvention.
    - The US repealed Executive Order 14110, which addressed risks from "dual-use foundation models with widely available model weights". This leaves the core risk – weaponisable AI models – largely unregulated.
- Some providers prepare "safety frameworks" and "model cards" on a voluntary basis, but are scaling back efforts. Independent evaluations have found that current voluntary safety frameworks are inadequate across the industry.[16]

[16] Future of Life Institute. (2025, July). *AI Safety Index – Summer 2025*.

Non-AI specific laws

- Australia's Civil Aviation Safety Authority (CASA) takes a risk-based approach for recreational drone operation and can impose fines for breaking rules. If AI advancements significantly increase drone risk profiles, CASA could impose further restrictions.
- While the scenario focuses on drones, open-weight models could enable the misuse of a broad range of technology. Most consumer technologies are not subject to specific regulations that could be adjusted in response to changing risk levels.
- General tort or negligence law is unlikely to be relevant, including because Australian courts are unlikely to find that a major US AI company had a duty of care to the Australian public in general, and other actors may be hard to identify.
- Using drones or other technology enabled by open-weight AI models to harm people is illegal, with various criminal laws covering privacy and surveillance violations. However, ex post criminal law applies individual culpability and does not address systemic risks or undo irreversible harms.

Strengthening our response

- **Half of experts expect 'Open-weight Misuse' to cause 'Moderate' or greater harm within five years.**
  52% rated this as **'Likely/probable'** (55%+ chance), meaning at least 9 fatalities, 18 casualties, or $20M economic damage.

- **2 in 3 experts expect 'Significant' to 'Catastrophic' consequences if 'Open-weight Misuse' occurs.**
  67% rated potential harm as **'High impact'**, including 17% who specifically warned of **'Catastrophic'** harm, meaning over 1,000 fatalities, 2,000 casualties, or $20B+ economic damage.

Summary

Existing regulation only partially covers a small number of threat vectors created by open-weight misuse. Most threat vectors are unregulated. Appropriate regulation of open-weight models represents a "chokepoint" where a small number of actors control the hazard, and a minimal upstream intervention could have widespread safety and economic benefits. Regulating the "chokepoint" is more desirable than greatly expanded regulation of consumer technology as open-weight misuse manifests.

Developers meeting safety requirements before releasing open-weight models is a lower-cost and higher-impact approach than wide-reaching regulation of consumer technology.

Could existing regulations be improved?

In this scenario, CASA could be resourced to monitor the ways in which AI is increasing the risk of drones and ramp up interventions as part of its risk-based approach. In the extreme, this could lead to a ban on consumer drones.

However, open-weight misuse could increase the risk of consumer technologies that are not currently subject to relevant regulation. Open-weight misuse also applies to cyber offensive tools, which involve the

misuse of general-purpose computers. Cybercrime already costs Australians billions each year,[17] and open-weight models could dramatically increase the capability of cyber attackers. Current criminal law is largely ineffective at preventing or punishing cyberattacks.

Are new regulations needed?

New laws could require AI developers to assess general-purpose AI models for dangerous capabilities and publish transparency reports. This could include third-party “red-teaming” to verify self-evaluation. Where dangerous capabilities are found, robust safeguards should be applied before release.

Open-weight models should meet higher standards of safeguard robustness than "closed" models. Closed models can be subject to ongoing supervision, whereas safeguards can be more readily removed from open-weight models, and open-weight models cannot be patched or withdrawn if risks emerge. If a model's dangerous capabilities are only “suppressed” by technical safeguards that could be bypassed, such a model should not be open-weight. Regulation of this kind may encourage model developers to find more effective safeguards, such as substantively removing dangerous knowledge from models, rather than simply suppressing the model’s propensity to share dangerous knowledge.

The law should clearly define liability in cases where AI models are released despite possessing dangerous capabilities. The law should ensure practical access to justice for Australians harmed by AI misuse.

[17] Australian Signals Directorate – Australian Cyber Security Centre. (2024). *Annual Cyber Threat Report 2023–2024*. Australian Government.

ACCESS TO DANGEROUS CAPABILITIES

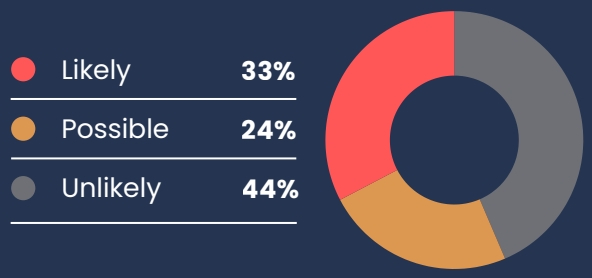
AI models could give a wider range of actors easier access to dangerous capabilities, such as the ability to conduct a cyber attack or build chemical, biological, radiological, or nuclear (CBRN) weapons.

By providing expert-level guidance and removing technical barriers, AI could enable less skilled actors to carry out attacks that previously required substantial expertise and resources.

In 2025, OpenAI and Google warned that their leading models had crossed new CBRN risk thresholds. Google assessed that Gemini 2.5 Deep Think reached the "early warning threshold" for its CBRN risk standard – models that "can be used to significantly assist a low-resourced actor with dual-use scientific protocols, resulting in a substantial increase in ability to cause a mass casualty event".[18] OpenAI made similar warnings for its ChatGPT Agent[19] and GPT5 systems.[20]

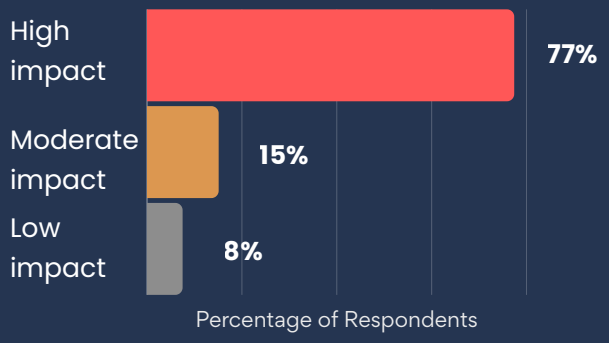
In 2024, Google claimed an AI agent was able to discover a "zero day" – a previously unknown cybersecurity vulnerability – in widely used real-world software.[21]

How **Likely** is Access to Dangerous Capabilities to Cause **Moderate or Greater harm*** in Australia?



Note. n = 55. Moderate or greater harm: >9 fatalities, >18 casualties, or >$20M AUD economic cost in the next five years.

If **Harm Occurs**, How **Severe** Would It Be?



Note. n = 53. High impact: 41+ fatalities, 81+ casualties, or $200M+ economic cost annually

"AI is a force multiplier that allows low-level knowledge to be sufficient to perform high-level harm."

– Amy Wilson, Lawyer (White Cleland)

Good Ancestors

[18] Google DeepMind. (2025, August 1). *Gemini 2.5 deep think model card.*
[19] OpenAI. (2025, July 17). *ChatGPT agent system card.*
[20] OpenAI. (2025, August 7). *GPT-5 system card.*
[21] Big Sleep Team. (2024, November 1). *From Naptime to Big Sleep: Using large language models to catch vulnerabilities in real-world code*. Google Project Zero.

Stress test scenario: Bioweapons

Sam, studying a Bachelor of Biological Sciences in Sydney, decides to try making smallpox at home. Sam uses a jailbreak on a leading US AI model via its website and asks for instructions. They download the genome of the Variola virus, which was made public in the 1990s, order its constituent parts from a mail-order synthetic DNA provider, and follow the AI's instructions to create a live virus.

The virus spreads rapidly, only being contained after a global effort. 10% of people who contract the virus die, including Sam.

How current laws respond

◆ **84% of experts rate current Australian laws and policies for mitigating access to 'Dangerous Capabilities' as 'Inadequate'.**

Summary

Existing synthetic DNA importation restrictions are relevant to managing the biosecurity risks of AI, but have not kept pace with the growing risk. The Crimes (Biological Weapons) Act 1976 is also relevant, but threatening Australian employees at AI labs with life in prison without a broader regulatory scheme to explain best practice is counterproductive to building a positive safety culture. While some laws are relevant to biological misuse, fewer laws apply to cyber offensive capability or chemical or radiological weapons.

AI-specific laws

- Currently, no Australian law requires AI developers to assess models for dangerous capabilities, apply safeguards if dangerous capabilities are identified, or ensure safeguards are robust to circumvention.
    - The US repealed Executive Order 14110, which had safeguards for AI-enabled CBRN misuse and set a path for synthetic DNA screening. This leaves AI-enabled biological misuse without binding regulation.
    - The US's July 2025 AI Action Plan proposed new safeguards to address the misuse of dangerous AI capabilities, which may reinstate synthetic DNA screening obligations in the US.
- Some providers prepare "safety frameworks" and "model cards" on a voluntary basis, but are scaling back efforts. Independent evaluations have found that current voluntary safety frameworks are inadequate across the industry.[22]
- Australia endorsed the Hiroshima AI Process (HAIP),[23] which says countries should ensure AI systems do not lower barriers to chemical, biological, radiological, and nuclear risks.
    - Australia has not made public statements when AI labs have potentially breached the HAIP code of conduct.
    - Australia has not implemented any aspects of the HAIP into domestic law.

[22] Future of Life Institute. (2025, July). *AI Safety Index – Summer 2025*.
[23] Ministry of Internal Affairs and Communications, Japan. (2024). *Hiroshima AI Process – Documents of Achievement*. Government of Japan.

Non-AI specific laws

- **Synthetic DNA controls:** The Department of Agriculture requires import permits for synthetic DNA with self-certification that the sequences are not dangerous. The Department of Agriculture may apply a fit and proper person test in some circumstances. "Sam" may pass a fit and proper person test given their affiliation with a university and lack of criminal past.
    - The Department of Agriculture does not regulate AI models or systems and has not publicly altered its approach to synthetic DNA screening in light of growing AI risks.
- **The Office of the Gene Technology Regulator (OGTR)** may regulate the use of AI models or systems in physical containment facilities in some circumstances.
    - OGTR's mandate related to "genetically modified organisms", it is unclear whether its mandate extends to ab initio DNA synthesis.
- **Biological weapons:** The Crimes (Biological Weapons) Act 1976 applies to Australians overseas. Australian Federal Police may arrest Australians working at AI labs who are reckless about building models capable of aiding bioweapon construction. This would be punishable by life imprisonment.
- **Tort law and negligence:** General tort or negligence law is unlikely to be relevant, including because Australian courts are unlikely to find that a major US AI company had a duty of care to the general public, and the harm potentially exceeds the value of the company. Action against "Sam" would also be ineffective.

Strengthening our response

- **1 in 3 experts expect access to 'Dangerous Capabilities' to cause 'Moderate' or greater harm within five years.**
  33% rated this as **'Likely/probable'** (55%+ chance), meaning at least 9 fatalities, 18 casualties, or $20M economic damage.

- **Nearly 4 in 5 experts expect 'Significant' to 'Catastrophic' consequences if access to 'Dangerous Capabilities' occurs.**
  77% rated potential harm as **'High impact'**, including 42% who specifically warned of **'Catastrophic'** harm, meaning over 1,000 fatalities, 2,000 casualties, or $20B+ economic damage.

- **42% of experts ranked access to 'Dangerous Capabilities' as their top policy priority.**

Summary

While some risks can be tolerated, this risk is above acceptable thresholds. In other domains, Australia goes to significant lengths to prevent access to dangerous capabilities, from chemical security to firearm restrictions. Australia is well-positioned to immediately uplift the functions of the Department of Agriculture and the Office of the Gene Technology Regulator to meaningfully reduce AI-accelerated biosecurity risks. However, a general law requiring developers to assess the risk of their models and impose effective safeguards would help address dangerous capabilities beyond biosecurity.

Could existing regulations be improved?

In this scenario, the Department of Agriculture – in conjunction with an Australian AI Safety Institute – could be resourced to monitor and evaluate the biosecurity risk of frontier AI models. The Department of Agriculture is already able to require safety screening of all synthetic DNA imports and impose more active and stringent background checking requirements, but has not done so.

However, this approach would only be partially effective as it is limited to DNA importation and could be bypassed if critical inputs are sourced domestically or importation requirements are otherwise avoided. The Office of the Gene Technology Regulator could have an expanded remit, including being explicit that ab initio DNA is within its remit (rather than mere genetic modification) and giving it specific functions to manage risks at the intersection of AI and biosecurity. For instance, a regulator should be in a position to place restrictions on benchtop DNA synthesis if justified by the risk.

Australia could also enhance its international coordination by encouraging other countries to mandate safety screening for all synthetic DNA imports and by demonstrating its commitment to the HAIP process. This could include identifying AI providers that may be in breach of the HAIP code of conduct and updating the Voluntary AI Safety Standard to call on Australian businesses to review the HAIP compliance of AI labs during procurement processes.

Are new regulations needed?

While existing laws provide some coverage of biosecurity risks, other dangerous capabilities are subject to fewer safeguards. New laws could require AI developers to assess general-purpose AI models for dangerous capabilities, including third-party red-teaming to verify self-evaluation. If dangerous capabilities are found, robust safeguards should be applied before release that cannot be readily bypassed through jailbreaking.

The law should clearly define liability in cases where AI models are released despite possessing dangerous capabilities. The law should ensure practical access to justice for Australians harmed by AI misuse.

LOSS OF CONTROL

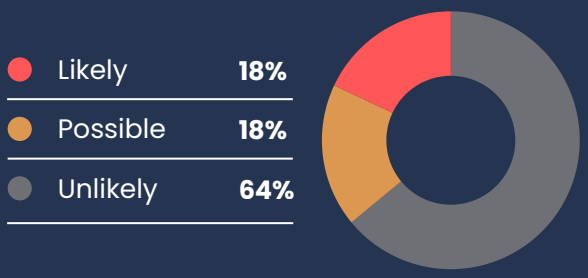
An AI lab could lose control of an AI model through mechanisms such as self-replication, recursive self-improvement, or the bypassing of containment measures.

Leading labs say they intend to build Artificial General Intelligence (AGI) – AI models that match or exceed humans at all tasks. Some claim this could happen as early as 2026. They're currently developing AI models that excel at AI research itself, including coding, synthesising scientific findings, and operating computer systems. They plan to provide those AI models with large amounts of data and computing power and ask them to iterate towards AGI.

The results are unpredictable. Key technical and policy questions remain unanswered, like how to align AI with human values, and which values should guide these systems. Meanwhile, AI labs have already built self-improving AI systems, including the Darwin Gödel Machine and Google's AlphaEvolve.[24] Meta CEO Mark Zuckerberg says this process is underway now and progress is already occurring.[25]

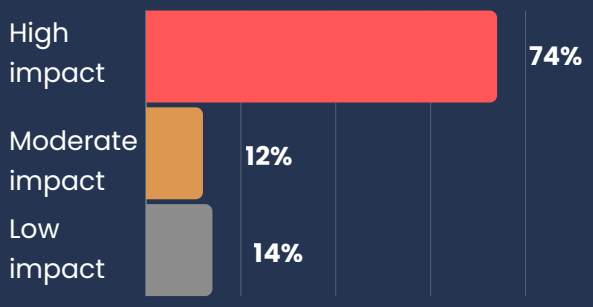
The *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*, endorsed by Australia, calls on developers to manage risks from models "self-replicating" or training other models.[26]

How **Likely** is Loss of Control to Cause **Moderate or Greater harm*** in Australia?



Note. n = 56. Moderate or greater harm: >9 fatalities, >18 casualties, or >$20M AUD economic cost in the next five years.

If **Harm Occurs**, How **Severe** Would It Be?



Note. n = 50. High impact: 41+ fatalities, 81+ casualties, or $200M+ economic cost annually

"We're building increasingly autonomous AI systems without the governance infrastructure to manage containment failures. That's a dangerous gamble with potentially catastrophic stakes."

– Alexander Saeri, AI Governance Researcher (MIT FutureTech)

Good Ancestors

[24] Zhang, J. et al., (2025). *Darwin Godel machine: Open-ended evolution of self-improving agents*. arXiv preprint arXiv:2505.22954.
[25] Zuckerberg, M. (2025, July 30). *Personal superintelligence*. Meta.
[26] Hiroshima Process. (2024). *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. Ministry of Internal Affairs and Communications, Japan.

Stress test: AI Incident Response

A US AI model developer says that it has made a breakthrough on recursively self-improving AI and sees a path towards AGI. The developer stops making its models available to the public and rents access to global compute clusters, paying above market rate.

Two months later, the CEO calls an emergency press conference, explaining that the company directed a recursively self-improving AI model to produce AGI. The company suspended the effort after a rival developer unexpectedly found an unknown AI model running in its private compute cluster – seemingly a descendant of the self-improving model. The company concluded that its model has escaped containment and is operating outside of the company's control. The company is not aware of the model's capabilities, when it escaped, or how far it has spread.

The CEO recommends that all countries and companies activate AI crisis plans, shut down compute clusters, and monitor for anomalous energy use on power grids. The company will report back with advice on identifying and containing the model.

How current laws respond

- **93% of experts rate current Australian laws and policies for mitigating 'Loss of Control' as 'Inadequate'.**

Summary

Experts assess Australia as being the least prepared for loss-of-control risks. Australia has made some relevant commitments to tackle this risk, but is yet to act on those commitments. Despite that, Australia has some laws and policies relevant to emergency response, including the ability to issue directions to some data centres and cybersecurity response plans.

AI-specific laws

- Australia, as well as the US and most leading AI developers, have endorsed the Hiroshima AI Process (HAIP), which says that countries and companies should seek to ensure AI models are not able to self-replicate and should adequately manage the risks of models being used to train other models.
    - Some providers prepare "safety frameworks" and "model cards" voluntarily, but are scaling back efforts.
    - Independent evaluations have found that voluntary safety frameworks are inadequate across the industry.[27] On "existential safety"—the specific assessment that covers loss of control risks—most labs scored an "F" grade, with the best lab, Anthropic, scoring a "D" grade.
        - One expert reviewer noted that all seven companies are racing to build AGI within the decade, yet "literally none of the companies has anything like a coherent, actionable plan for what should happen if what they say will happen soon and are very actively working to make happen, happens."
    - Despite evaluations showing that models are beginning to demonstrate behaviours like self-replication and resisting shutdown,[28] Australia has not made any public statements

[27] Future of Life Institute. (2025, July). *AI Safety Index – Summer 2025*.
[28] Fortune. (2025). *Anthropic's new AI Claude Opus 4 threatened to reveal engineer's affair to avoid being shut down*.

highlighting the importance of compliance with HAIP and has not implemented any HAIP requirements into domestic law.

Non-AI specific laws

- The Australian Government Crisis Management Framework does not have a specific AI Crisis Plan, but does have a Cyber Response Plan.[29]
    - AUSCYBERPLAN does not mention AI-specific risks.
- Australia's Security of Critical Infrastructure Act (SOCI Act) gives Australia some ability to issue directions to critical infrastructure owners and operators.
    - The SOCI Act, and coordination with AEMO, would allow Australia to monitor for unexpected power consumption on the NEM.
    - The SOCI Act only regulates data centres that provide services to Government or other critical infrastructure. SOCI does not cover data centres that train or run AI, meaning those data centres could not be directed to shut down or take other measures in response to an AI crisis.

Strengthening our response

- **1 in 6 experts expect 'Loss of Control' to cause 'Moderate' or greater harm within five years.** 18% rated this as **'Likely/probable'** (55%+ chance), meaning at least 9 fatalities, 18 casualties, or $20M economic damage.

- **3 in 4 experts expect 'Significant' to 'Catastrophic' consequences if 'Loss of Control' occurs.** 74% rated potential harm as **'High impact'**, including 54% who specifically warned of **'Catastrophic'** harm, meaning over 1,000 fatalities, 2,000 casualties, or $20B+ economic damage.

Summary

A general law could require AI developers, users, and deployers to comply with applicable standards. The EU has developed a code of practice that most US labs have already endorsed, making this a practical path of Australia.[30] This approach would require AI developers to meet safety standards throughout the AI development process. Australian legislation should demonstrate our commitment to the HAIP process by including clauses that reflect our HAIP commitments, including mandating that risk management plans address loss of control risks.

Regardless of the status of domestic regulation, Australia's diplomatic efforts should emphasise a national expectation that countries and companies that train frontier models do so safely. Historically, Australia has urged governments to de-escalate nuclear arms races and supported an unequivocal ban on biological weapons. Australia should adopt a similar stance on addressing dangerous behaviour involving artificial intelligence.

[29] Department of Home Affairs. (2025). *Australian Cyber Response Plan – V.1*. Commonwealth of Australia.
[30] European Commission AI Office. (2025). *EU AI Act: General-Purpose AI Code of Practice*. European Commission.

Could existing regulations be improved?

The SOCI Act should be amended to include all data centres that are used, or could be used, to run or train AI models. This would allow the SOCI Act directions power to be used in support of a specific AI Crisis Plan.

The Australian Government Crisis Management Framework (AGCMF) should be updated to include a specific AI Crisis Plan, which is regularly exercised in conjunction with critical infrastructure operators and AI safety experts.

Are new regulations needed?

Australia could view this risk as primarily a matter for countries that house leading labs and for international law. Australia has endorsed HAIP and statements at AI Safety Summits that acknowledge this risk and call for action. If Australia chooses to primarily address this risk through bilateral and multilateral action, it should be vocal in calling for countries and companies to be transparent about how they meet their HAIP obligations.

New laws could impose transparency obligations on AI developers, including requiring them to publish adequate plans that detail compliance with HAIP obligations, including how they plan to manage loss of control risks. The EU's Code of Practice addresses many of these concerns and is already endorsed by most leading labs, suggesting a practical international path Australia can follow. This would be consistent with Australia's approach in other sectors, like medicine, where we require developers to develop medicine in specific ethical ways and consistent with global best practice.[31]

[31] Therapeutic Goods Administration. (2025, January 17). *ICH Guideline for Good Clinical Practice*. Australian Government.


Good
Ancestors

Appendix: Assessment Scale Data Tables

Likelihood assessment

Question:
In the next 5 years, how likely do you think it is that [threat] will cause *Moderate+ harm* in Australia?

Full 9-point scale:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
1 - Extremely remote (<0.2%)	0 (0.0%)	1 (1.8%)	1 (1.8%)	0 (0.0%)	6 (10.7%)
2 - Very remote (~0.2-1%)	3 (6.0%)	2 (3.6%)	1 (1.8%)	2 (3.6%)	3 (5.4%)
3 - Remote (1-5%)	4 (8.0%)	5 (9.1%)	5 (9.1%)	10 (18.2%)	11 (19.6%)
4 - Highly unlikely (10-20%)	4 (8.0%)	2 (3.6%)	0 (0.0%)	6 (10.9%)	11 (19.6%)
5 - Unlikely (25-35%)	5 (10.0%)	9 (16.4%)	1 (1.8%)	6 (10.9%)	5 (8.9%)
6 - Realistic probability (40-50%)	8 (16.0%)	10 (18.2%)	8 (14.6%)	13 (23.6%)	10 (17.9%)
7 - Likely/probable (55-75%)	13 (26.0%)	9 (16.4%)	14 (25.5%)	10 (18.2%)	2 (3.6%)
8 - Highly likely (80-90%)	4 (8.0%)	10 (18.2%)	9 (16.4%)	3 (5.5%)	5 (8.9%)
9 - Almost certain (95-100%)	9 (18.0%)	7 (12.7%)	16 (29.1%)	5 (9.1%)	3 (5.4%)
Total Responses	**50**	**55**	**55**	**55**	**56**

Simplified 3-category:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
Unlikely (1-5)	16 (32.0%)	19 (34.6%)	8 (14.6%)	24 (43.6%)	36 (64.3%)
Possible (6)	8 (16.0%)	10 (18.2%)	8 (14.6%)	13 (23.6%)	10 (17.9%)
Likely (7-9)	26 (52.0%)	26 (47.3%)	39 (70.9%)	18 (32.7%)	10 (17.9%)
Total Responses	**50**	**55**	**55**	**55**	**56**

Severity assessment

Question:
Assuming [threat] causes harm in Australia, how severe would the impact likely be over the course of a year?

Full 5-point scale:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
1 - Limited harm 1-8 fatalities, 1-17 casualties, or <$20m AUD economic cost	6 (13.0%)	12 (22.6%)	10 (19.2%)	4 (7.6%)	7 (14.0%)
2 - Moderate harm 9-40 fatalities, 18-80 casualties, or $20m-200m AUD economic cost	9 (19.6%)	18 (34.0%)	17 (32.7%)	8 (15.1%)	6 (12.0%)
3 - Significant harm 41-200 fatalities, 81-400 casualties, or $200m-$2b AUD economic cost	11 (23.9%)	14 (26.4%)	18 (34.6%)	9 (17.0%)	5 (10.0%)
4 - Severe harm 201-1,000 fatalities, 400-2,000 casualties, or $2b-$20b AUD economic cost	12 (26.1%)	4 (7.6%)	6 (11.5%)	10 (18.9%)	5 (10.0%)
5 - Catastrophic harm >1,000 fatalities, >2,000 casualties, or >$20b AUD economic cost	8 (17.4%)	5 (9.4%)	1 (1.9%)	22 (41.5%)	27 (54.0%)
Total Responses	**46**	**53**	**52**	**53**	**50**

Simplified 3-category:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
Low impact (1)	6 (13.0%)	12 (22.6%)	10 (19.2%)	4 (7.6%)	7 (14.0%)
Moderate impact (2)	9 (19.6%)	18 (34.0%)	17 (32.7%)	8 (15.1%)	6 (12.0%)
High impact (3-5)	31 (67.4%)	23 (43.4%)	25 (48.1%)	41 (77.4%)	37 (74.0%)
Total Responses	**46**	**53**	**52**	**53**	**50**

Mitigation adequacy assessment

Question:
How adequate are current measures by the Australian Government (e.g., laws, regulations, policies) for mitigating risks from [threat]?

Full 5-point scale:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
1 - Completely inadequate	14 (37.8%)	15 (34.1%)	7 (17.5%)	14 (32.6%)	32 (71.1%)
2 - Mostly inadequate	18 (48.6%)	20 (45.5%)	24 (60.0%)	22 (51.2%)	10 (22.2%)
3 - Moderately adequate	2 (5.4%)	4 (9.1%)	7 (17.5%)	5 (11.6%)	2 (4.4%)
4 - Mostly adequate	1 (2.7%)	2 (4.6%)	2 (5.0%)	1 (2.3%)	0 (0.0%)
5 - Completely adequate	2 (5.4%)	3 (6.8%)	0 (0.0%)	1 (2.3%)	1 (2.2%)
Total Responses	**37**	**44**	**40**	**43**	**45**

Simplified 3-category:

	Open-Weight Misuse	Unauthorised Agent Actions	Unreliable Agent Actions	Access to Dangerous Capabilities	Loss of Control
Inadequate (1-2)	32 (86.5%)	35 (79.5%)	31 (77.5%)	36 (83.7%)	42 (93.3%)
Moderately adequate (3)	2 (5.4%)	4 (9.1%)	7 (17.5%)	5 (11.6%)	2 (4.4%)
Adequate (4-5)	3 (8.1%)	5 (11.4%)	2 (5.0%)	2 (4.7%)	1 (2.2%)
Total Responses	**37**	**44**	**40**	**43**	**45**

Policy urgency rankings

Question:
Having reviewed all threat types, please rank them from 1 to 5 based on how urgently Australia should develop new policies to address each type of threat

	Open-Weight Misuse	**Unauthorised Agent Actions**	**Unreliable Agent Actions**	**Access to Dangerous Capabilities**	**Loss of Control**
Rank 1 *(highest priority)*	2 (3.8%)	3 (5.8%)	7 (13.5%)	22 (42.3%)	18 (34.6%)
Rank 2	12 (23.1%)	14 (26.9%)	8 (15.4%)	13 (25.0%)	5 (9.6%)
Rank 3	12 (23.1%)	19 (36.5%)	12 (23.1%)	7 (13.5%)	2 (3.8%)
Rank 4	11 (21.2%)	12 (23.1%)	10 (19.2%)	7 (13.5%)	12 (23.1%)
Rank 5 *(lowest priority)*	15 (28.8%)	4 (7.7%)	15 (28.8%)	3 (5.8%)	15 (28.8%)
Average Rank	**3.48**	**3.00**	**3.35**	**2.15**	**3.02**
Total Responses	**52**	**52**	**52**	**52**	**52**