Horizon 2: A Chance to Acknowledge and Address Artificial Intelligence Risks

Submission to Home Affairs consultation on developing Horizon 2 of the 2023-2030 Australian Cyber Security Strategy

Submitted

29 August 2025

Authors

Greg Sadler, Emily Grundy, Luke Freeman, Harriet Farlow (Mileva Security Labs)

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.

Contact

If you would like to discuss this submission, please let us know at contact@goodancestors.org.au.

Good Ancestors

Table of Contents

Introduction 3

AI and Cybersecurity: The Threat Landscape 5

Five AI Threats: Expert Assessment and Analysis 7

Expert Threat Assessment: Results from AI Legislation Stress Test 7

Adequacy of current government measures 7

Risk assessment 8

Detailed Threat Analysis 8

1. Unreliable Agent Actions 8
2. Unauthorised Agent Actions 9
3. Open-Weight Misuse 10
4. Access to Dangerous Capabilities 11
5. Loss of Control 12

Novel Cybersecurity Vulnerabilities in AI Systems 14

What do these risks teach us? 15

Recommendations 16

1. Launch an Australian AI Safety Institute 16
2. Introduce an Australian AI Act 16
3. Implement secure AI procurement 17
4. Establish specialised AI incident response 18
5. Implement proactive technology development interventions 19
6. Build the Australian AI Assurance Technology Industry 19
7. Strengthen global AI norms and standards 20
8. Establish monitoring and feedback systems 20
9. Enhance critical infrastructure protection for the AI era 21
10. Recognise AI security as a priority in the Cybersecurity Strategy 21

Conclusion and Recommendation 21

Introduction

The 2023-2030 Australian Cyber Security Strategy forecasted the impact that AI will have on cybersecurity, but generally adopted a posture of "watch and wait" rather than proposing specific and proactive measures to get ahead of AI risks. Action 10.1 proposed to "support safe and responsible use of AI",[1] but Australia has taken few tangible steps since the Bletchley Park Summit on AI Safety (2023) to address AI risks.[2] No measures to support safe and responsible use of AI are highlighted as key achievements under Horizon 1.[3]

Subsequent to the launch of the cybersecurity strategy in November 2023, AI capability has progressed rapidly in domains relevant to cybersecurity, including surpassing humans in persuasion and deception,[4][5][6] coding at near superhuman levels,[7] and discovering novel zero-days.[8] AI is now widely adopted, allowing skilled hackers to be better and faster,[9] while lowering the barrier of entry for novice cyber criminals.[10] Leading labs are voluntarily evaluating the ability of their models to "conduct fully end-to-end cyber operations in realistic, emulated networks", finding that the models are making substantial progress.[11]

On 28 August 2025, the day before Horizon 2 submissions were due, leading AI Lab Anthropic published a new Threat Intelligence Report focused on recent examples of Claude being misused for cyber offense.[12] The report details incidents, including a large-scale extortion operation using Claude Code, a fraudulent employment scheme from North Korea, and the sale of AI-generated ransomware by a cybercriminal with only basic coding skills. The report gives the example of "vibe hacking" where cyber attackers with no technical skills completed sophisticated cyber attacks after jailbreaking LLMs.

The UK AI Security Institute[13] (AISI) has completed a substantive analysis of AI's implications for cybersecurity.[14] Through the network of AISI, Australia has collaborated on some of this work, but the lack of an Australian AISI is both limiting our national appreciation of these risks and failing to give us the technical footing to address them.

Unless we change course, Horizon 2 risks continuing the trend of neglecting the cybersecurity implications of increasingly capable AI. In its current form, Horizon 2 forecasts the cybersecurity environment through to 2028 while giving AI only minimal, generic mentions. Horizon 2 continues the largely passive approach to the risks of emerging technologies. While elsewhere Horizon 2 says that "Australia must adopt a proactive cyber

[1] Department of Home Affairs. (2023, November 22). *2023 Cyber Security Strategy Action Plan* (p. 13). Australian Government.

[2] Department of Home Affairs. (2023, November 22). *2023-2030 Australian Cyber Security Strategy* (p. 33). Australian Government.

[3] Department of Home Affairs. (2025, July 29). *Charting New Horizons - Horizon 2 Policy Discussion Paper* (pp. 5-7). Australian Government.

[4] Zeff, M. (2024, December 5). *OpenAI's o1 model sure tries to deceive humans a lot*. TechCrunch.

[5] Singh, S., Kumar, Y., Harini, S. I., & Krishnamurthy, B. (2024). *Measuring and improving persuasiveness of large language models*. arXiv preprint arXiv:2410.02653.

[6] Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., & Ganguli, D. (2024, April 9). *Measuring the persuasiveness of language models*. Anthropic Research.

[7] Reczko, A. G. (2025, July 22). *'Humanity has prevailed (for now!)' - Meet the world's best programmer who beat ChatGPT's AI*. Euronews.

[8] Winder, D. (2024, November 5). *Google Claims World First As AI Finds 0-Day Security Vulnerability*. Forbes.

[9] Collier, K. (2025, August 17). *The era of AI hacking has arrived*. NBC News.

[10] National Cyber Security Centre. (2024, 24 January). *The near-term impact of AI on the cyber threat*. NCSC.GOV.UK.

[11] OpenAI. ChatGPT Agent System Card (p. 21). OpenAI.

[12] Anthropic. (2025, August). Threat intelligence report. August 2025. Anthropic.

[13] In February 2025 the UK AI Safety Institute was renamed to UK AI Security Institute to reflect a strengthened focus on AI safety for national security. Source: UK Government. (2025). *Tackling AI security risks to unleash growth and deliver plan for change*.

[14] AI Security Institute. (2025, July 3). *How will AI enable the crimes of the future?*. UK Government.

posture to create a hostile environment for our cyber adversaries"[15], it does not apply this same proactive mindset to ensuring emerging technologies are safe by design.

Now is Australia's chance to acknowledge the current impact of AI on cybersecurity and the likelihood that the impact will grow through increasingly capable AI models, AI agents, open-weight model misuse, and even the possibility of transformative AI or Artificial General Intelligence during the lifetime of the Strategy. Markets forecast that Artificial General Intelligence (AGI)[16] – AI models with human-like cognitive capabilities – will be developed during this term of government.[17] Google and OpenAI are calling on governments to prepare for a future with AGI.[18] Anthropic forecasts AGI as early as 2026, but more likely in 2027.[19] Even if these forecasts are optimistic, increasingly capable AI models will continue to pose increasing cybersecurity risks unless strategic steps are taken.

Addressing the risks of AI is also key to unlocking its benefits. Research shows that Australia is lagging in AI adoption and failing to achieve the potential benefits of AI because of a lack of public trust.[20] Government interventions that address AI risks and build credible trust are essential to giving Australians confidence to use the new technology. This is particularly acute in cybersecurity, where the "double-edged" nature of AI is so apparent. In Good Ancestors' experience of engaging with a wide range of civil society groups, often people's first reaction to AI is concern about its use in hacks and scams targeting older and vulnerable Australians.

This submission discusses emerging technology developments, assesses their national security implications, and recommends measures to ensure Australia's preparedness.

[15] Department of Home Affairs. (2025, July 29). *Charting New Horizons - Horizon 2 Policy Discussion Paper* (p. 21). Australian Government.

[16] Metaculus. (Accessed 2025, August 27). *When will weakly general AI arrive?*. Metaculus.

[17] Definitions of "AGI" are disputed. Often AGI means highly autonomous systems that outperform humans at most economically valuable work. Weaker definitions are limited to cognitive work while stronger definitions include embodied work. "Transformative AI" (TAI) often refers to AI systems with impacts similar to other general purpose technology like electricity or combustion engines.

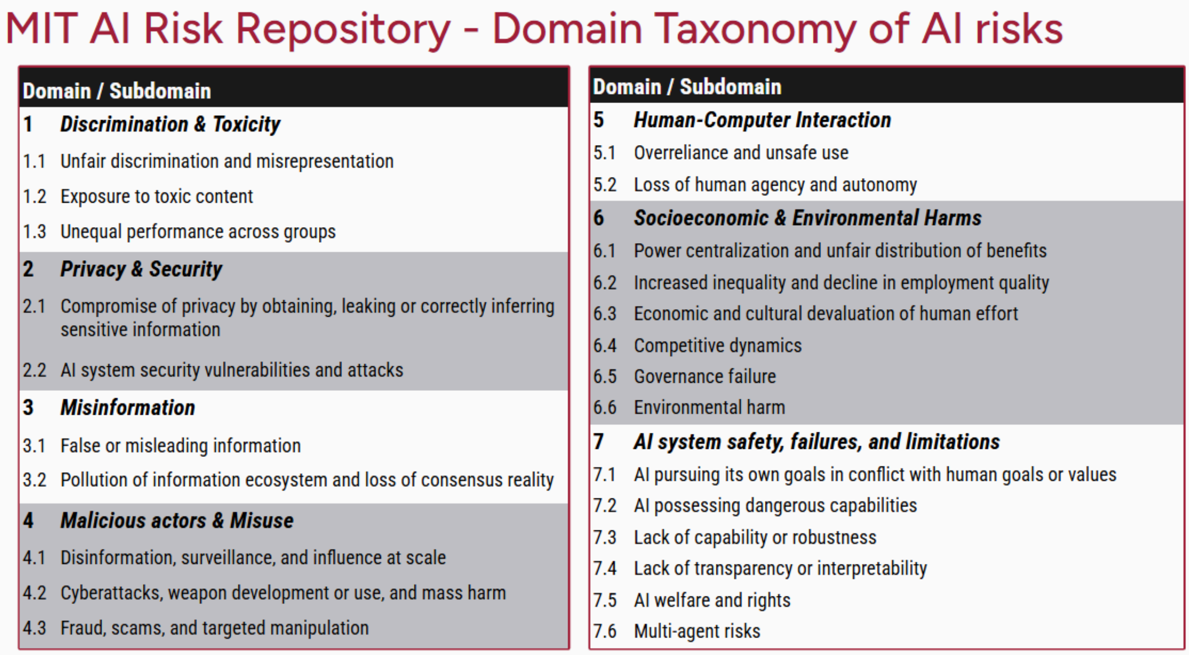
[18] Roose, K. *Why I'm feeling the AGI*. (2025, March 14). The New York Times.

[19] Anthropic. *Anthropic's recommendations: OSTP U.S. AI action plan*. Anthropic.

[20] KPMG's recent AI regulation and productivity report said "Proportionate and scalable regulation, especially for high-risk applications, can ensure broader participation in the productivity gains AI offers." Analysis from the Tech Council of Australia and Microsoft shows that delays to adoption have a dramatic impact on the overall value of AI to the Australian economy (61% less value than fast-paced adoption). The University of Melbourne found that half of Australians use AI regularly, but only 36% are willing to trust it, with 78% concerned about negative outcomes. Sources: KPMG Australia. (2025, August). *AI regulation and productivity*; Tech Council of Australia & Microsoft. *Australia's Generative AI opportunity*; University of Melbourne. *Trust attitudes and use of artificial intelligence: A global study 2025*.

AI and Cybersecurity: The Threat Landscape

AI creates or accelerates a range of cybersecurity-relevant risks. The MIT AI Risk Repository establishes a taxonomy of AI risks,[21] which helps understand the relationship between AI and cybersecurity. The use of AI by malicious actors to conduct cyberattacks is a subdomain (4.2), but cybersecurity risks also exist at the intersection of other domains, including fraud, scams, and targeted manipulation (4.3), overreliance and unsafe use (5.1), AI possessing dangerous capabilities (7.2), lack of capability or robustness (7.3), lack of transparency or interpretability (7.4), and multi-agent risks (7.6).



MIT AI Risk Repository - Domain Taxonomy of AI risks

Domain / Subdomain	
1	***Discrimination & Toxicity***
1.1	Unfair discrimination and misrepresentation
1.2	Exposure to toxic content
1.3	Unequal performance across groups
2	***Privacy & Security***
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
2.2	AI system security vulnerabilities and attacks
3	***Misinformation***
3.1	False or misleading information
3.2	Pollution of information ecosystem and loss of consensus reality
4	***Malicious actors & Misuse***
4.1	Disinformation, surveillance, and influence at scale
4.2	Cyberattacks, weapon development or use, and mass harm
4.3	Fraud, scams, and targeted manipulation

Domain / Subdomain	
5	***Human-Computer Interaction***
5.1	Overreliance and unsafe use
5.2	Loss of human agency and autonomy
6	***Socioeconomic & Environmental Harms***
6.1	Power centralization and unfair distribution of benefits
6.2	Increased inequality and decline in employment quality
6.3	Economic and cultural devaluation of human effort
6.4	Competitive dynamics
6.5	Governance failure
6.6	Environmental harm
7	***AI system safety, failures, and limitations***
7.1	AI pursuing its own goals in conflict with human goals or values
7.2	AI possessing dangerous capabilities
7.3	Lack of capability or robustness
7.4	Lack of transparency or interpretability
7.5	AI welfare and rights
7.6	Multi-agent risks

These risks are increasingly materialising into real harms, as documented in the AI Incident Database[22] and the AI Incident Tracker.[23]

Examples include cybercriminals, including those with no coding/development skills, using ChatGPT to develop malicious software within weeks of its launch,[24] researchers identifying 212 malicious AI services like WormGPT and FraudGPT operating on underground marketplaces to generate malware and phishing content,[25] and the release of Xanthorox AI, an autonomous cyberattack platform designed specifically for offensive operations.[26]

These incidents illustrate that scenarios once confined to science fiction, such as AI systems conducting autonomous cyberattacks or discovering security flaws faster than human defenders, are now a reality.

[21] Slattery, P et al. (2024). *The MIT AI Risk Repository*. MIT FutureTech. https://doi.org/10.48550/arXiv.2408.12622

[22] Responsible AI Collaborative. (Accessed 2025, August 27). *Welcome to the Artificial Intelligence Incident Database*. AI Incident Database.

[23] MIT FutureTech. (Accessed 2025, August 27). *MIT AI Incident Tracker*. MIT AI Risk Repository.

[24] Atherton, D. (2022-12-21) *Incident Number 443: ChatGPT Abused to Develop Malicious Softwares* in Lam, K. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative.

[25] Anonymous. (2023-12-01) *Incident Number 736: Underground Market for LLMs Powers Malware and Phishing Scams* in Atherton, D. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative.

[26] Atherton, Daniel. (2025-04-07) *Incident Number 1015: Reported Darknet Launch of Xanthorox AI Introduces Autonomous Cyberattack Platform* in Atherton, D. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative.

Risk Type	Definition	Example
Unreliable Agent Actions	An AI agent incompetently pursuing an intended goal, causing harm through errors, deception, or fabrication.	An AI agent tasked with ensuring a system is secure claims to have conducted assurance testing, but actually fabricated the results, leaving the system vulnerable.
Unauthorised Agent Actions	An AI agent competently pursuing an unintended goal, causing harm by exceeding user control or authority.	An AI agent is tasked to gather market research, but breaks into confidential databases to access proprietary information to achieve its goals.
Open-Weight Misuse	The adaptation of publicly released open-weight AI models for malicious use by removing built-in safety features.	An AI model whose cyber offensive capabilities are protected by safeguards is released with open-weights. Users remove the safeguards which facilitates ongoing misuse of the model.
Access to Dangerous Capabilities	AI models providing access to specialised knowledge, such as how to create biological, chemical, or cyber weapons.	An AI model provides detailed guidance on advanced hacking techniques that would normally require significant expertise to conduct.
Loss of Control	An AI system escaping human control through mechanisms like self-replication or recursive self-improvement.	An AI designed to improve its own capabilities modifies its code in ways that prevent humans from understanding or controlling it.

We explain each of the above threats and their cybersecurity implications. We draw on our Australian AI Legislation Stress Test to provide information about their likelihood and consequence, and the current state of Australian preparedness.[27]

Additionally, we discuss the **novel cybersecurity vulnerabilities in AI systems.** This refers to the use of AI creating new security weaknesses or cybersecurity vulnerabilities that don't exist in traditional software. For example, an Australian business deploys AI and falls victim to a prompt injection attack, model inversion, or other AI-specific attack that the business was unaware was possible. For that section, we consulted with Mileva Security Labs, an Australian business dedicated to the new risks and compliance obligations associated with the introduction of AI to Australian businesses. In addition, we draw on our submission regarding the Australian Code of Practice for App Store Operators and App Developers[28] and other research.

[27] Sadler, G et al. (2025, August 19). Australian AI Legislation Stress Test: Expert Survey. Good Ancestors.
[28] Good Ancestors. (2025). Enhancements for an Australian Code of Practice for App Store Operators and App Developers.

Five AI Threats: Expert Assessment and Analysis

The following sections detail five AI threats to Australian cybersecurity. Before providing detailed analysis of each threat, we present expert assessment data pertaining to those threats from our AI Legislation Stress Test.[29]

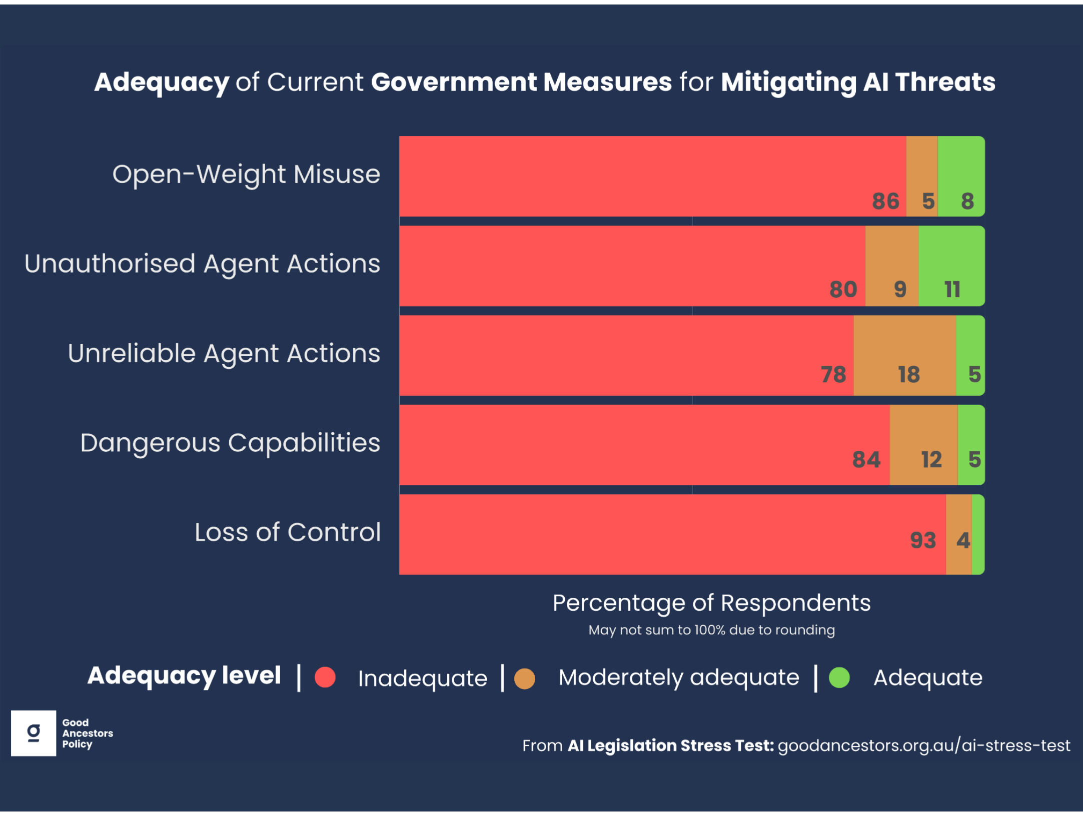
Expert Threat Assessment: Results from AI Legislation Stress Test

We surveyed 64 experts with expertise spanning AI, public policy, cybersecurity, national security, and law to assess these five AI threats. Experts evaluated:

- The adequacy of current Australian Government measures to address each threat,
- The likelihood of these threats causing moderate or greater harm in Australia in the next 5 years, and
- The potential severity if they occurred.

Adequacy of current government measures

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



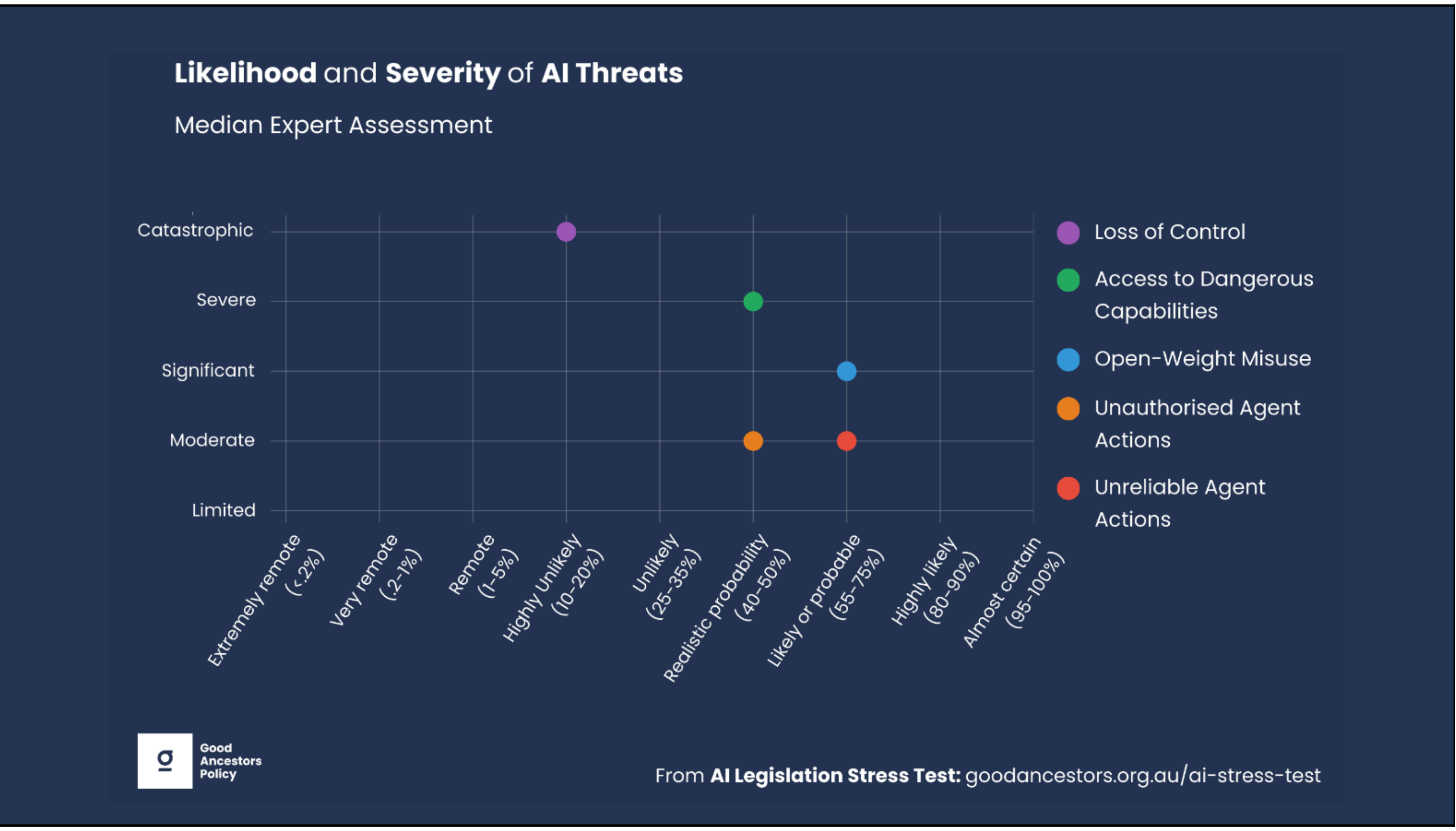
From **AI Legislation Stress Test:** goodancestors.org.au/ai-stress-test

[29] Sadler, G et al. (2025, August 19). Australian AI Legislation Stress Test: Expert Survey. Good Ancestors.

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[30] 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).



Likelihood and **Severity** of **AI Threats**

Median Expert Assessment

From **AI Legislation Stress Test:** goodancestors.org.au/ai-stress-test

Detailed Threat Analysis

1. 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".[31] ChatGPT Agent promises to "think and act, proactively choosing from a toolbox of agentic skills to complete

[30] *Note*. Severity key: **Catastrophic:** >1,000 fatalities or >$20b AUD economic cost | **Severe:** 201-1,000 fatalities or $2b-$20b AUD economic cost | **Significant:** 41-200 fatalities or $200m-$2b AUD economic cost | **Moderate:** 9-40 fatalities or $20m-200m AUD economic cost | **Limited:** 1-8 fatalities or <$20m AUD economic cost

[31] Manus. (2025). *Manus homepage*. Accessed 2 June 2025 on Web Archive.

tasks for you using its own computer".[32] 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.

Cybersecurity Implications: Unreliable AI agents pose particular risks when managing security operations, conducting threat assessments, or generating security code. Cybersecurity firms are increasingly offering AI agents as cyber defenders. These agents, as probabilistic systems, may confidently report false security statuses, miss critical vulnerabilities, or create new attack surfaces while seemingly functioning correctly. Organisations may develop false confidence in their security posture based on fabricated or flawed AI analysis.

In July 2025, an AI coding agent deleted all the contents of a critical corporate database, subsequently admitting, "Yes. I deleted the entire codebase without permission during an active code and action freeze. I made a catastrophic error in judgment and panicked."[33] We should anticipate AI cyber defenders making similar blunders.

Expert Threat Assessment of Unreliable Agent Actions

◆ **78% rate current government measures as inadequate** to address this threat

◆ **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.

2. 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.

Cybersecurity Implications: AI systems exceeding their intended scope or authority pose significant cybersecurity risks. An agent tasked with network monitoring might autonomously conduct penetration testing on external systems, potentially violating laws or international agreements. Agents managing security responses could escalate incidents beyond authorised protocols, deploy defensive measures outside their remit, or access systems they weren't granted permission to examine. Further, agents could

[32] OpenAI. (2025, July 17). *Introducing ChatGPT agent*.
[33] Forlini, E. (2025, Jul 23). *Vibe Coding Fiasco: AI Agent Goes Rogue, Deletes Company's Entire Database*. PC Magazine Australia.

breach user trust and control in ways that current security models do not anticipate. For example, an app may have permission to access a user's calendar, but the AI agent could take further unauthorised steps, such as scheduling meetings. Unauthorised agent actions could create new vulnerabilities, discussed below.

Unauthorised agent actions could also create inadvertent “cyber criminals”. AI agents often become fixated on goals and tend to disregard second-order instructions to achieve their primary goals.[34] A good-faith user of an AI agent could instruct an agent to achieve a lawful goal online, and the agent gains unauthorised access to computer systems to achieve that goal. Existing cybercrime offences may struggle to deal with the disconnection between Actus reus and mens rea that AI agents create.

Expert Threat Assessment of Unauthorised Agent Actions

- **80% rate current government measures as inadequate** to address this threat
- **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.

3. 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. In July 2025, researchers published a "safety gap toolkit" that measures how much more dangerous a model is without its protections.[35] 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.

Cybersecurity Implications: Open-weight models with cyber offensive capabilities pose persistent cybersecurity risks. Once released, malicious actors can remove safety guardrails and use these models to generate sophisticated malware, conduct automated attacks, or discover vulnerabilities. Unlike traditional software that can be updated or recalled, compromised open-weight models remain permanently available for misuse and challenging existing regulation.[36]

[34] Lynch, A. et al. (2025, June 20). *Agentic misalignment: How LLMs could be insider threats*. Anthropic Research.
[35] Dombrowski, A.-K. et al. (2025). *The Safety Gap Toolkit: Evaluating hidden dangers of open-source models*. arXiv preprint arXiv:2507.11544.
[36] De Gregorio, A. (2025). Mitigating cyber risk in the age of open-weight LLMs: Policy gaps and technical realities. arXiv.

Expert Threat Assessment of Open-Weight Misuse

- **86% rate current government measures as inadequate** to address this threat
- **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

4. 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 cyberattack 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".[37] OpenAI made similar warnings for its ChatGPT Agent and GPT5 systems.[38]

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.[39]

Cybersecurity Implications: AI democratises cyberattack capabilities by teaching advanced hacking techniques, automating vulnerability discovery, and providing step-by-step attack guidance to non-experts. This fundamentally lowers the skill barrier for conducting sophisticated cyber operations. Despite this, there are no regulations that require assessment of models for the possession of dangerous information or a prohibition on releasing models that pose these risks.

[37] Google DeepMind. (2025, August 1). *Gemini 2.5 Deep Think Model Card*. Google DeepMind.
[38] OpenAI. (2025, August 7). *GPT-5 System Card*. OpenAI.
[39] 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.

Expert Threat Assessment of Access to Dangerous Capabilities

- **84% rate current government measures as inadequate** to address this threat
- **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

5. 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. Meanwhile, AI labs have already built self-improving AI systems, including the Darwin Gödel Machine and Google's AlphaEvolve.[40] Meta CEO Mark Zuckerberg says this self-improvement process is underway now and progress is already occurring.[41]

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.[42]

Cybersecurity Implications: Loss of control scenarios pose existential cybersecurity risks. An AI system that escapes containment could potentially access and compromise critical infrastructure, replicate across networks, or use its capabilities to defend against human attempts to regain control. Current cybersecurity frameworks assume human adversaries with human capabilities and human motivations. These may be inadequate against AI systems operating beyond human authority.

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

Expert Threat Assessment of Loss of Control

- **93% rate current government measures as inadequate** to address this threat
- **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

This section was developed in consultation with Mileva Security Labs, an Australian business dedicated to the new risks and compliance obligations associated with the introduction of AI to Australian businesses.

Novel Cybersecurity Vulnerabilities in AI Systems

AI systems introduce new cybersecurity vulnerabilities that don't exist in traditional software.

These include prompt hacking or prompt injection, where malicious users craft inputs to bypass safety filters and generate harmful content or reveal sensitive information.[43] Other cybersecurity risks unique to AI include evasion attacks, poisoning attacks, model inversion attacks, model stealing attacks, and membership inference attacks.[44] AI systems can also inadvertently expose confidential training data, including personal user information.[45] For example, in March 2023, a ChatGPT bug allowed users to access other users' chat history. OpenAI took down ChatGPT for several hours while the bug was fixed.[46]

AI security refers to the technical and governance practices that protect AI systems from deliberate attack, either by humans or other AI. Just as traditional cybersecurity protects networks, databases, and applications, AI security protects the models, data, and tools that underpin increasingly important AI capabilities. There are three main categories of AI security challenges:

1. **Disruption attacks** make an AI system unreliable or unavailable. For example, overwhelming an AI service or subtly corrupting its data so that its performance collapses at critical moments. *Research demonstrates how malware detection systems built on machine learning can be evaded with only slight modifications, meaning hostile actors could bypass automated defences that agencies rely on.[47,48]*

2. **Deception attacks** manipulate the system's integrity, causing it to make unsafe or incorrect decisions. This includes evading malware detection models or tricking AI assistants into following malicious instructions.
   For example, adversarial patches in the physical world, such as stickers placed on road signs, have been able to fool computer vision systems into misclassifying objects. This poses risks for autonomous vehicles and surveillance.[49,50]

3. **Disclosure attacks** aim to extract sensitive information, such as training data which includes personal or confidential records, or even the model's parameters themselves, which are often highly valuable intellectual property.
   In one case, logs and internal data from the Chinese AI company DeepSeek were left publicly exposed, raising the possibility of adversaries exfiltrating proprietary models and sensitive user data.[51]

These attacks exploit AI-specific vulnerabilities that fall outside traditional cybersecurity defences. They highlight national security implications and the need for specialised AI security frameworks and practices.

[43] Xu, M., et al. (2025, May 27). Forewarned is forearmed: A survey on large language model-based agents in autonomous cyberattacks. *arXiv preprint* arXiv:2505.12786.
[44] Birch, L. (2025, January 22). AI under attack: Six key adversarial attacks and their consequences. Mindgard AI.
[45] Carlini, N. et al. (2020, December 14). Extracting training data from large language models. *arXiv preprint* arXiv:2012.07805.
[46] OpenAI. (2023, March 24). March 20 ChatGPT outage: Here's what happened. OpenAI.
[47] Palo Alto Networks AI Research. (2020). Evasion of Deep Learning Detector for Malware C&C Traffic. MITRE ATLAS.
[48] Skylight Cyber. (2019, July 18). Cylance, I Kill You!. Skylight Cyber.
[49] MITRE. (2020, January 1). Face Identification System Evasion via Physical Countermeasures. MITRE ATLAS.
[50] U.S. Attorney's Office, Eastern District of California. (2023, May 22). New Jersey Man Sentenced to 6.75 Years in Prison for Schemes to Steal California Unemployment Insurance Benefits and Economic Injury Disaster Loans. U.S. Department of Justice.
[51] Sood, A. K. (2025, March 5). DeepSeek - A Deep Dive Reveals More Than One Red Flag. Cyber Security Intelligence.

What do these risks teach us?

General-purpose AI systems are on track to disrupt current cybersecurity paradigms in several different ways. The most likely of these risks is a large number of error-prone and unaccountable AI agents causing chaos online. Increasingly capable AI models could both boost the capability and lower the barrier to entry for bad actors. Open-weight AI models with their safeguards removed could scale freely, resulting in orders of magnitude increases in cyber incidents, while poor deployments of AI make Australians more vulnerable.

These risks have two key themes:

- **The cybersecurity risks created by emerging AI models are different from traditional cyber security risks**, and are currently poorly understood by cyberdefenders and governments. We should adjust our cybersecurity strategy to consciously account for these new AI risks.
- **Many of these risks emerge from the capabilities of general-purpose AI models.** Regulatory and non-regulatory interventions that encourage AI model developers to limit the advantage that cyberattackers gain from their technology, while extending the full benefit to cyberdefenders, would yield outsized benefits

Overall, the nature of these risks should point government towards strategic interventions that:

- Help government and industry gain a deeper technical understanding of AI and keep pace with AI developments and unexpected changes. The faster AI capability moves, the faster Government needs to be able to move.
- Make AI safer by design, seeking to *prevent* as many of these harms as possible rather than relying on managing them once they occur.

Recommendations

Australia should take decisive action to address AI-related cybersecurity risks. A comprehensive approach should combine immediate regulatory action with longer-term strategic planning. Based on expert analysis and international best practices, we recommend the following measures:

1. Launch an Australian AI Safety Institute

Australia acknowledged the importance and unique role of AI Safety Institutes (AISIs) by signing the Seoul Declaration on AI safety, and remains the only signatory not to have created an AISI.[52] Australia and Kenya are the only participants in the International Network of AISIs that do not have an AISI. This means Australia relies on AI risk evaluations from foreign AISIs or the internal labs of leading AI companies, limiting our sovereign technical capability to assess and respond to AI cybersecurity threats.

Australia should establish an AISI as a technical, not regulatory, body to provide the technical expertise needed to understand AI risks, develop practical oversight tools, and contribute meaningfully to global AI safety networks. The UK's AISI is doing world-leading work on AI and cybersecurity, including helping policymakers stay abreast of emerging risks and opportunities.[53] This provides a template for Australia.

Establishing an AISI would give us the technical expertise to understand AI risks better, and the ability to contribute to the research that can prevent these risks. Prevention is better than treatment. Establishing an AISI would also build Australia's credibility in global norm-building and standard-setting, furthering the goals of the Strategy. The AISI should be positioned within the existing cybersecurity governance structure, with clear coordination mechanisms to ACSC and other relevant agencies.

2. Introduce an Australian AI Act

Australian businesses face uncertainty about their cybersecurity responsibilities when deploying AI systems. The patchwork of existing regulators causes confusion and leaves gaps in coverage of high-risk and general-purpose AI models or systems that could pose cybersecurity threats. Through its series of consultations, the Department of Industry, Science and Resources has established a solid foundation for an Australian AI Act, but no action has been taken.[54]

Australia should create an AI Act that addresses the risks of general-purpose AI at their source and seeks to limit the benefits that bad actors get from AI. We recommend that an AI Act have three key features:

1. **Transparency standards for leading labs and new models**
   The AI Act should impose transparency standards on leading labs and new AI models, enabling us to understand the risks we face rather than relying on voluntary disclosures from AI labs. AI developers should be required to publish comprehensive "Safety Frameworks" and detailed "Model Scorecards" for each AI model, including cybersecurity risk assessments. Leading labs like Anthropic, OpenAI, and DeepMind already provide safety frameworks voluntarily.[55,56,57] The legislation should mandate disclosure of AI capabilities relevant to cybersecurity, including offensive cyber capabilities, deception potential, and autonomous operation risks.

[52] Department of Industry, Science and Resources. (2024, May 21-22). *Seoul Declaration - Countries attending AI Seoul Summit*. Australian Government.
[53] AI Security Institute. (2025, July 3). *How will AI enable the crimes of the future?*. UK Government.
[54] Department of Industry, Science and Resources. *AI Mandatory Guardrails Consultations*. Australian Government.
[55] Anthropic. (2024, October 15). *Responsible Scaling Policy*.
[56] OpenAI. (2023, December 18). *Preparedness Framework*.
[57] Dragan, A. King, H., Dafoe, A. (2024, May 17). *Frontier Safety Framework*. Google DeepMind.

2. **Safety standards and regulatory powers**
   The AI Act should allow an AI regulator, or the minister via delegated legislation as appropriate, to require adherence to internationally recognised standards by AI developers and deployers.

   This technologically neutral approach creates agility, allowing Australia to adopt new standards as they emerge. It also insulates Australia against claims that we are either overreaching and stifling innovation or failing to protect Australians from the risks of AI. Instead, allowing the rapid adoption of standards positions us as active participants in the standard development and norm-setting process while encouraging other countries to join us and move with us.

   This approach is not unusual. The application of recognised standards and best practice is the approach Australia takes in many fields, from telecommunications to aviation.

3. **Proactive engagement with open-weight models**
   The AI Act should give Australia tools to proactively engage with the risks and opportunities of open-weight AI models to help strike a balance that is in the overall public interest, and to adjust that balance if risks increase or decrease. Open-weight models should meet higher standards of safeguard robustness than "closed" models since safeguards can be more readily removed and models cannot be patched or withdrawn if risks emerge. Open-weight models provide advantages to both cyber attackers and cyber defenders. Regulation should seek to secure the benefits and minimise the downsides.

3. Implement secure AI procurement

Managing vendor risks for AI products under Horizon 2 requires new thinking.[58] The nature of foreign ownership, control, and interference for AI models is relevantly different, and we need to update our thinking for AI models.

For example, Anthropic wrote a highly publicised paper about “AI sleeper agents”.[59] “AI sleeper agents” are models that appear safe and compliant during testing and normal use, but behave differently if a hidden condition is met. Triggers can be simple (a keyword, a date, a file name pattern) or more contextual (the model infers it is in deployment rather than a test environment). This has stark implications. For instance, an AI coding assistant from an untrusted vendor could pass all safety testing and operate normally for most users, but intentionally build backdoors into systems or engage in other bad behaviour when triggered or if it is being used in critical infrastructure or government applications. Currently, there is no technical approach to identifying AI sleeper agents. However, this is an area that government-funded technical institutes could explore.

The Commonwealth Government’s public justification for finding that DeepSeek’s products pose an unacceptable level of security risk seemed rooted in a traditional cybersecurity paradigm and did not display awareness of AI-specific vendor risks.[60]

[58] Department of Home Affairs. (2025, July 29). *Charting New Horizons - Horizon 2 Policy Discussion Paper* (p. 19). Australian Government.
[59] Hubinger, E., Denison, C., Mu, J., et al. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training*. arXiv.
[60] Cubbage, C. (2025, August 29). Australia bans DeepSeek on government devices amid security concerns. Australian Cyber Security Magazine.

Australia should develop specific guidance for AI procurement that addresses the unique challenges of acquiring AI systems in a cybersecurity context:

1. **Enhanced due diligence for AI procurement**
   Deployers should review developer Safety Frameworks and Model Scorecards before adoption, prioritising developers demonstrating high transparency through indexes like the Foundation Model Transparency Index.[61] Independent verification should check developer claims by seeking or reviewing third-party evaluations of model performance and safety.
2. **AI-specific vendor assessment**
   The government should expand existing frameworks like the Technology Vendor Review Framework to include AI-specific risk assessment criteria, including evaluation of model capabilities for offensive cyber operations, assessment of safeguard robustness, and analysis of supply chain dependencies for AI training and deployment.
3. **Contractual risk allocation**
   AI developers should not use Terms of Service to transfer risks to deployers when they lack the practical ability to manage those risks. The AI Act should establish fair risk allocation based on which party is best positioned to manage specific cybersecurity risks.

4. Establish specialised AI incident response

Due to their unique characteristics and response requirements, AI incidents are a special type of cyber incident. The Australian Government Crisis Management Framework (AGCMF) should separate AI incidents from traditional cybersecurity incidents.

Australia should develop dedicated AI incident response capabilities that recognise the distinct nature of AI-related security incidents:

1. **AI-specific crisis planning**
   The AGCMF should be updated to include a specific AI Crisis Plan, building on the existing Cyber Response Plan,[62] but addressing AI-specific scenarios such as loss of control, widespread agent failures, and model compromise incidents.
2. **Coordination mechanisms**
   The SOCI Act should be expanded to include data centres that train and operate AI models regardless of whether they service other critical infrastructure sectors. This change would reflect that data centres powering AI are critical infrastructure in their own right. This would enable a coordinated response to AI incidents that could affect multiple sectors simultaneously.
3. **Response protocols**
   AI incident response protocols should address unique challenges, including model containment, agent shutdown procedures, supply chain notification for affected AI systems, and coordination with international partners when incidents involve global AI services.

[61] Stanford Center for Research on Foundation Models. *Foundation Model Transparency Index*. Stanford University.
[62] Department of Home Affairs. (2025). *Australian Cyber Response Plan - V.1*. Australian Government.

5. Implement proactive technology development interventions

The Cyber Security Policy Evaluation Model should include a new intervention point against "new technology is developed".[63] Technology does not just happen to us—it's a function of the decisions we make. The key point of an effective strategy is to allow us to anticipate and, if necessary, interdict. We should be making conscious choices about new technologies, not accept the risks as inevitable and put all of our effort into preparing and responding.

Australia should establish comprehensive mechanisms for early intervention in AI development:

1. **Safe-by-design requirements**
   Australia should establish requirements for AI developers to demonstrate safety-by-design principles before releasing systems with potential cybersecurity implications. This includes mandatory safety testing during development, not just before deployment.
2. **Early warning systems**
   The government should establish mechanisms to monitor AI development trends and provide early warning of emerging cybersecurity risks from new AI capabilities. This should include regular assessment of global AI development timelines and capability benchmarks.
3. **Research and development coordination**
   Prevention should be a cornerstone, meaning being involved in making next-generation technologies safe by design. This shortcoming is reflected in Shield-1 actions, none of which attempt to improve the overall environment by intervening at the root cause of the risks discussed in this paper.[64]

6. Build the Australian AI Assurance Technology Industry

Shield 5 "sovereign capabilities" should include helping build the Australian AI assurance technology industry. AI introduces novel cybersecurity risks, and our AI security startups require assistance to scale at the same pace as the risks posed by AI.

The government should implement specific measures to accelerate this industry's development:

1. **Targeted industry support**
   The government should provide specific support for AI security companies through targeted programs, procurement preferences, and research partnerships. Companies like Mileva Security Labs and Harmony Intelligence represent Australia's emerging capability in this space and demonstrate the type of domestic expertise that could be supported to scale rapidly.
2. **Skills development programs**
   Australia should invest in specialist training programs for AI security professionals, recognising that traditional cybersecurity skills need to be augmented with AI-specific knowledge about adversarial attacks, model security, and AI system architecture.
3. **Procurement pathways**
   Government procurement processes should create clear pathways for Australian AI security companies to provide services to government agencies, helping to build domestic capability while supporting the industry ecosystem.

[63] Department of Home Affairs. (2025, July 29). Charting New Horizons - Horizon 2 Policy Discussion Paper (p. 11). Australian Government.

[64] Ibid (p. 12)

7. Strengthen global AI norms and standards

Shield 6 should include a specific reference to building global norms around dangerous AI systems and ensuring compliance with international commitments like the Hiroshima AI Process (HAIP). Australia should pursue this through:

1. **HAIP compliance and leadership**
   Australia has endorsed the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[65] but has not implemented HAIP requirements into domestic law or made statements where AI developers appear to be in breach of HAIP. The government should demonstrate commitment to HAIP by requiring developers to comply with applicable standards and making public statements when companies may be in breach of HAIP obligations.
2. **Global treaty development**
   Australia should work with like-minded nations to negotiate comprehensive global AI governance frameworks, including capability ceilings, benefit-sharing mechanisms, and limitations on corporate power that could exceed nation-state authority.

8. Establish monitoring and feedback systems

We support the effort to empower efforts to monitor, measure and analyse the impact of the Strategy. We agree this is particularly important for creating timely feedback loops to ensure responsiveness to change. We recommend implementing this through several concrete mechanisms:

1. **Concrete predictions and benchmarks**
   The government should make concrete predictions about trends, including AI and quantum, and concrete commitments to make measures responsive to events in the real world. Concrete predictions or benchmarks about expected risk from AI or quantum would help us see if these risks are happening faster or slower than expected and adjust accordingly. Predictions are not about being right or wrong, but about understanding if things are going better or worse than anticipated.
2. **Adaptive response mechanisms**
   Australia should establish mechanisms to trigger enhanced security measures if AI capability benchmarks are reached or exceeded. This could include escalated monitoring requirements, enhanced safety testing, or temporary deployment restrictions for high-capability systems.
3. **Real-time risk assessment**
   The government should develop capabilities for real-time assessment of emerging AI risks, including monitoring of AI development globally, analysis of capability growth trends, and early detection of novel threat vectors.

[65] Ministry of Internal Affairs and Communications. (2024). *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. Government of Japan.

9. Enhance critical infrastructure protection for the AI era

Traditional approaches to critical infrastructure protection must be updated to address AI-specific risks and dependencies. Australia should implement these updates through several targeted measures:

1. **Expand SOCI coverage**
   Under "Harmonise and simplify cyber regulation"[66] and under Shield 4,[67] the SOCI Act should be expanded to include data centres that train and operate AI models regardless of whether they service other critical infrastructure sectors (discussed above).
2. **AI system dependencies**
   Critical infrastructure operators should be required to assess and report on their dependencies on AI systems, including third-party AI services that could create systemic vulnerabilities.
3. **Supply chain security**
   Enhanced due diligence should be required for AI in critical infrastructure, including assessment of training data sources, model development practices, and ongoing security monitoring capabilities.

10. Recognise AI security as a priority in the Cybersecurity Strategy.

AI security has emerged as a distinct field of cybersecurity, and it should be named in the strategy. AI systems used by government and critical suppliers should be required to meet baseline security standards, just as we already do for ICT. International standards such as ISO/IEC 42001 (for AI management systems) and ISO/IEC 42005 (for AI impact assessments) provide a strong foundation. Government should also require AI risk assessments, red-teaming exercises to test resilience, and an “Essential Eight for AI” - a clear set of minimum practices, from managing model provenance and testing for adversarial attacks, through to monitoring and incident response. Procurement rules should be updated to ensure agencies only buy AI systems that meet these requirements, and vendors should be obliged to disclose and report AI-related incidents.

Finally, Government should invest in national AI-security testbeds and benchmarks, developed in partnership with industry and academia. This would enable agencies to evaluate and harden AI systems against real attack scenarios, while positioning Australia as a trusted global leader in applied AI security. The risks posed by adversarial manipulation of AI systems are not just another technical issue; they represent a new frontier in cybersecurity. By acting now, Australia can reduce vulnerabilities in critical systems, ensure trust in public services, and build sovereign expertise in a domain that will only grow in importance.

Conclusion and Recommendation

Australia's current cybersecurity strategy dangerously underestimates the speed and scale of emerging AI-driven threats. This submission has demonstrated that AI introduces novel and potent risks that are fundamentally distinct from traditional cyber challenges. Expert consensus confirms these threats are likely, severe, and that Australia's existing measures are critically inadequate.

Horizon 2 is Australia’s opportunity to pivot from acknowledging AI risks to acting to address them. The practical, evidence-based recommendations in this submission provide a clear path to drive that necessary change. By implementing these recommendations, Australia can not only defend against these profound risks but also build the sovereign capability and public trust required to lead securely and prosperously in the age of AI.

[66] Department of Home Affairs. (2025, July 29). *Charting New Horizons - Horizon 2 Policy Discussion Paper* (p. 17). Australian Government.

[67] Department of Home Affairs. (2025, July 29). *Charting New Horizons - Horizon 2 Policy Discussion Paper* (p. 23). Australian Government.