# Strengthening CIRMP Rules to Address Artificial Intelligence

**Submission to the 'Consultation on enhancements to the CIRMP Rules'**

***Good Ancestors is an Australian charity providing evidence-based policy recommendations for Australia's biggest challenges. We work with experts around the world and help organise Australians for AI Safety.***

Enhancements to the Critical Infrastructure Risk Management Program (CIRMP) rules are necessary to "keep pace with the evolving threat landscape".[1] However, these enhancements must anticipate the growing role artificial intelligence (AI) plays in Australian society, and the new threats it creates for national security.

Industry and Government, as well as investment trends and capability evaluations, all point to **AI becoming an essential service which underpins Australian society and economy**. By 2030, AI could contribute between $45 billion and $115 billion annually to the Australian economy—equivalent to 2-5 per cent.[2] In the near future, AI may be embedded in most critical infrastructure operations—including healthcare, banking, logistics, and government services—and augment consequential decision-making.

**Increasingly advanced and integrated AI systems will expose Australia to new threats**, including to national security.[3] Large-scale AI infrastructure creates vulnerabilities as both a target and vector (i.e., it could be both used to cause harm or be a target itself). For example, data centres training or operating general-purpose AI models could be compromised through cyberattacks or physical incidents. AI data centres could also cause harm through poisoned training data or manipulated outputs that could propagate to downstream users, including critical infrastructure. As AI becomes increasingly advanced and integrated into Australian society, the potential consequences of inadequate management and protection grow.

**In 2025, we submitted to the Independent Review of the SOCI Act**, arguing that the Act should a) more adequately capture data centres training and operating general-purpose AI models as critical infrastructure, and b) create a pathway to covering AI models themselves once they become essential to Australian society.[4]

**This submission outlines what the CIRMP rules can do under the current regime.** We make two recommendations:

1. Extend the enhanced CIRMP rules to "Critical data storage or processing assets"
2. Develop guidance for AI-specific material risks

---

[1] Critical Infrastructure Security Centre. (n.d.) *Proposed amendments to enhance the Critical Infrastructure Risk Management Program Rules (CIRMP Rules).* Department of Home Affairs, Australian Government
[2] Microsoft and Tech Council of Australia. (2023). *Australia's Generative AI Opportunity.* Microsoft and Tech Council of Australia.
[3] Grundy, E., Sadler, G., & Freeman, L. (2025). *Artificial Intelligence and National Security*. Good Ancestors.
[4] Grundy, E. & Sadler, G. (2025). *Extending SOCI Act Coverage to AI Infrastructure.* Good Ancestors.

## Recommendation 1: Extend enhanced CIRMP rules to "Critical data storage or processing assets"

Data centres, particularly those training and operating general-purpose AI models, should be subject to the enhanced CIRMP rules.

As outlined in the consultation paper, the enhanced rules apply to asset classes based on a) their "criticality…to the ongoing availability of other critical infrastructure sectors and the broader economy" and b) "assessments conducted by the National Intelligence Community" (p. 4). Data centres training and operating AI models meet these criteria.

**AI infrastructure increasingly underpins other critical infrastructure.** General-purpose AI models trained or operating in Australian data centres may be embedded in sectors like healthcare, finance, energy, and government services. If these facilities are disrupted or compromised, the effects would cascade across other critical infrastructure sectors and the economy. This is the same rationale used to justify enhanced rules for energy, communications, water, and transport.

**Data centres will also be high-value targets.** The consultation paper highlights that "hostile foreign state actors and their proxies are increasingly targeting critical infrastructure globally to gain strategic leverage" (p. 2). AI training and operating facilities house valuable intellectual property, sensitive data, and computational resources. This makes them attractive targets. See the Scenario box below for an example.

> ℹ️ **Scenario: Malicious customer rents compute**
>
> A foreign actor rents GPU compute at a Sydney data centre and uses it to train and operate an AI model designed for sophisticated cyber attacks. Under baseline CIRMP rules, the data centre has no specific obligation to assess foreign ownership, control, or influence (FOCI) risks associated with its customers, map its supply chain vulnerabilities, or maintain a vendor assessment process.
>
> The model begins conducting automated attacks against Australian financial institutions. The Australian Signals Directorate attributes the attacks' origin to the Sydney data centre. However, the data centre operator has limited visibility into which customer is responsible, or even exactly who its customers are, leading to delays in identification and resolution.

Applying the enhanced CIRMP obligations to data centres would address some key gaps. For example:

- **FOCI as explicit material risk** would require considering and addressing risks associated with foreign influence, including from customers using their compute
- **Supply chain vulnerability mapping** would address, or at minimum understand, concentration risks in AI hardware supply chains (e.g., GPU/TPU production being dominated by a small number of foreign manufacturers)
- **Cyber maturity level 2** would establish a higher security baseline which would protect against potential attacks (see 'Risks from AI-enabled attacks' below)

## Recommendation 2: Develop guidance for AI-specific material risks

The Department is proposing new "cyber and information hazard specific material risks that the responsible entity will need to minimise or eliminate". This includes "the deployment of advanced and emerging technology, and use of such technology by malicious and state-sponsored actors against the asset" (p. 12). This includes AI.

The proposal for assets to consider and mitigate risks posed by AI is welcome, but requires specificity to be effective. Without concrete guidance, assets may not meaningfully address the threats they face.

**AI risks are novel and distinct from traditional cyber threats.** The responsible entities may not know what "consider AI risk" means in practice. Guidance accompanying these rules should outline both risks from deploying AI systems and from adversaries using AI to attack the asset.

### Risks from deploying AI

Critical infrastructure operators increasingly deploy AI for operational efficiency, monitoring, and decision support. This introduces risks that differ from traditional software, including:

- **Unreliable agent actions:** AI systems may pursue intended goals incompetently, causing harm through errors, deception, or fabrication. For instance, an AI agent tasked with monitoring system security could claim to have conducted assurance testing while actually fabricating results, leaving the system vulnerable.

- **Unauthorised agent actions:** AI agents may competently pursue unintended goals, causing harm by exceeding user control or authority. For instance, an agent tasked with network monitoring might autonomously conduct penetration testing on external systems, potentially violating laws or international agreements.

- **Novel AI security vulnerabilities:** AI systems introduce new security vulnerabilities that don't exist in traditional software. There are three main categories of novel AI security challenges:[5]

    - **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.*[6,7]

[5] Sadler, G., Grundy, E., Freeman, L., & Farlow, H. (2025). *Horizon 2: A Chance to Acknowledge and Address Artificial Intelligence Risks.* Good Ancestors.
[6] Palo Alto Networks AI Research. (2020). *Evasion of Deep Learning Detector for Malware C&C Traffic*. MITRE ATLAS.
[7] Skylight Cyber. (2019). *Cylance, I Kill You!*.

- **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 fooled computer vision systems into misclassifying objects. This poses risks for autonomous vehicles and surveillance.*[8,9]

- **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.*[10]

The above risks and vulnerabilities fall outside traditional cybersecurity considerations. Responsible entities need guidance to identify and mitigate these novel threats.

## Risks from AI-enabled attacks

AI enhances the efficiency and effectiveness of cyber attacks. It democratises access to capabilities that previously required substantial expertise: AI can teach advanced hacking techniques, automate vulnerability discovery, and provide step-by-step attack guidance to non-experts. Anthropic's August 2025 Threat Intelligence Report detailed how actors with only basic coding skills misused Claude for large-scale extortion and AI-generated ransomware—an evolution in AI-assisted cybercrime.[11]

AI also makes attacks harder to detect and defend against. It can discover cybersecurity vulnerabilities, as demonstrated by Google's AI agent Big Sleep discovering a "zero day"[12] in widely used real-world software.[13] AI can modify malware and attack methods to evade detection systems, reducing the effectiveness of traditional cybersecurity approaches. These capabilities create an asymmetric advantage where AI-enabled attacks not only succeed more often but are harder to detect and defend against.

Guidance should help operators understand how AI changes the threats they face.

[8] MITRE. (2020). *Face Identification System Evasion via Physical Countermeasures*. MITRE ATLAS.
[9] U.S. Attorney's Office, Eastern District of California. (2023). *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.
[10] Sood, A. K. (2025). *DeepSeek - A Deep Dive Reveals More Than One Red Flag*. Cyber Security Intelligence.
[11] Anthropic. (2025). *Threat intelligence report*. Anthropic.
[12] A zero day is a previously unknown cybersecurity vulnerability.
[13] Big Sleep Team. (2024). *From Naptime to Big Sleep: Using large language models to catch vulnerabilities in real-world code*. Google Project Zero.

## Example AI risks by sector

The following table provides examples of AI deployment and attack risks across various sectors.

| Sector | Risk type | Category | Example |
|---|---|---|---|
| Energy | Unreliable agent actions | Deployment | AI system managing grid load confidently reports normal operations while failing to detect cascading failures |
| Water | Prompt injection | Attack | Attacker manipulates AI-enabled water quality monitoring system to suppress contamination alerts |
| Communications | AI-enhanced social engineering | Attack | AI-generated phishing targets employees with access to broadcast or DNS infrastructure |
| Transport | Unauthorised agent actions | Deployment | AI logistics system autonomously reroutes shipments in ways that create supply chain vulnerabilities |
| Energy | Training data poisoning | Attack | Adversary corrupts training data for predictive maintenance AI, causing it to miss equipment failures |

Good Ancestors recommends the Department develop specific guidance on AI-related material risks, drawing on work by institutions such as the UK AI Security Institute, National Institute of Standards and Technology, Apollo Research, and Centre for AI Safety. This guidance should be updated regularly as AI capabilities and associated risks evolve.

# Conclusion

As the National AI Plan highlighted, AI is a "critical technolog[y] in the national interest" that is "already shaping our economy and society". Appropriate risk management practices will be essential to ensuring Australia's security and economic resilience. Enhancements to the CIRMP rules should reflect this.

**Submitted**
4 February 2026

**Authors**
Emily Grundy, Greg Sadler

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