AI Transparency and Incident Reporting to Support Ministerial Directions Powers

Submission to the consultation on proposed amendments to the Ministerial Directions Powers in Part 3 of the SOCI Act

30 April 2026

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.

Good Ancestors supports the proposed amendments to the ministerial directions powers. The reforms address practical barriers that have limited Government's ability to respond to national security threats. Artificial Intelligence (AI) is one such threat, and two of the measures are particularly relevant to AI-related risks: Measure 1 (reforming the s32 directions power) and Measure 3 (high-risk vendor restrictions).

We support both measures and make two recommendations to strengthen Measure 3. For the Minister to exercise the vendor-restriction power against AI-related risks, Government needs access to information it does not currently have. We recommend the government require transparency from AI developers and establish an AI incident reporting mechanism.

Measure 1: Reforming the s32 directions power supports faster response to AI risks

The proposed reforms to s32 — replacing the formal Adverse Security Assessment requirement with a flexible 'obtain and consider' model, and softening the regulatory exhaustion precondition — are sensible. Good Ancestors supports these changes.

AI-related national security risks may emerge faster than traditional threats. A compromised AI model embedded across multiple critical infrastructure sectors could cause cascading harm within hours. The proposed reforms would allow Government to respond at a speed closer to that at which AI risks can materialise.

Measure 3: Restrictions on high-risk vendors, products, or services requires information about AI risks and incidents

Good Ancestors supports the proposed vendor-risk direction power. The ability to issue directions across an entire asset class or sector matches how AI will be increasingly embedded throughout society. A single general-purpose AI model from a single developer could be integrated into energy, water, communications, and transport systems and operations. If that model is compromised, or if its developer is subject to foreign laws that compel data access or enable interference, the risk is systemic.

The AI Legislation Stress Test, produced by Good Ancestors with input from 64 experts, found that 78–93% of experts rated current government measures for mitigating AI threats as inadequate.[1] AI models trained with hidden behaviours that activate under specific conditions ('sleeper agents')[2] and models whose safety measures can be readily removed[3] represent the kinds of risks this power will hopefully be able to address.

However, for the Minister to determine that something "poses a material risk that is prejudicial to national security" and that vendor-risk direction is warranted, Government needs to understand the nature and extent of the risk. Two recommendations could address this in the case of AI.

Recommendation 1: Require, or at minimum heavily weight, transparency from frontier AI developers

The Minister cannot identify when an AI model or service poses a national security risk without information about the risks that model carries. The same gap affects critical infrastructure entities, who are required under the Security of Critical Infrastructure Act 2018 (SOCI) to identify and manage material risks to their assets, including supply chain risks. Where entities are integrating AI models and systems, meaningful risk assessment requires transparency from the AI developer about capabilities, known vulnerabilities, and the results of safety testing. Without a safety framework, system card, or independent evaluation results, neither the Minister nor the entity can properly identify the risks the product carries.

Transparency also enables independent evaluation. In other technology domains, devices like telecommunications or SCADA equipment are built to specifications and standards. This serves two purposes: Government and industry can assess whether a given standard is appropriate for a particular use case, and they can independently test whether specific equipment meets that standard. No equivalent currently exists for AI models. There is no standardised set of evaluations, performance thresholds, or certification claims against which a critical infrastructure entity can judge whether a model is fit for its use case. Absent such claims, there is also no defined test for Government, the Australian AI Safety Institute (AISI), or third parties to replicate or verify. As Australia develops AI model evaluation capability, including through the AISI, that evaluation benefits from being conducted against a developer's attestations about a model's design, capabilities, and known risks. The more transparency developers provide, the more practical it becomes for industry, the AISI, and intelligence agencies to assess AI risks to critical infrastructure.

Good Ancestors has recommended transparency requirements across multiple submissions and reports to Government, including that frontier AI developers publish Safety Frameworks and Model Scorecards detailing each model's capabilities, limitations, and testing results.[4] Mandatory transparency is the appropriate response. Voluntary disclosure is unreliable:[5] independent evaluations rate current voluntary safety frameworks as inadequate across the industry, red-teaming and model evaluation methodologies are not standardised, and results are often not published.[6] Even the strongest voluntary disclosure regimes fall short of what risk assessment requires.

---

[1] Sadler, G., Grundy, E., Freeman, L., & Sherburn, N. (2025). *Australian AI Legislation Stress Test*. Good Ancestors.
[2] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv.
[3] Dombrowski, A.-K., Bowen, D., Gleave, A., & Cundy, C. (2025). *The Safety Gap Toolkit: Evaluating Hidden Dangers of Open-Source Models*. arXiv.
[4] Good Ancestors. (2025). *Australia AI Policy, 2025–2028*.
[5] Future of Life Institute. (2025). *AI Safety Index – Summer 2025*.
[6] Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., ... & Mindermann, S. (2026). *International AI Safety report 2026*.

Absent a mandatory regime, the quality of transparency should itself be a factor in risk assessment. Critical infrastructure entities entering into contracts with AI vendors should actively seek detailed information about model capabilities, limitations, safety testing, and known vulnerabilities — and should expect a level of disclosure that exceeds current industry practice. Transparency that merely matches current practice should be a concern. Transparency that falls below it — for example, the absence of a detailed system card made available to the AISI and other third parties for independent evaluation — almost certainly represents an unacceptable risk for use in critical infrastructure. The practical operation of SOCI should put upward pressure on transparency expectations for AI used in critical infrastructure. AI vendors should be expected to meet rising expectations if they want their products integrated into critical infrastructure systems.

Recommendation 2: Establish an AI incident monitoring and reporting mechanism

Government needs a mechanism to systematically track AI failures, security incidents, or near-misses affecting critical infrastructure. This could include mandatory reporting from AI companies and voluntary reporting from other users and businesses.

Australia operates analogous systems in other sectors. The Australian Transport Safety Bureau requires mandatory notifications of accidents and serious incidents affecting aircraft safety. The Therapeutic Goods Administration monitors adverse events for medical devices. The ACCC can enforce safety standards and initiate product recalls for consumer goods. Applying similar principles to AI would give Government early warning of systemic vulnerabilities. The Minister needs this kind of information to determine whether a vendor-risk direction is warranted.

The EU AI Act requires providers and deployers of high-risk AI systems to report serious incidents, including serious disruption of critical infrastructure.[7] California requires AI companies training models above a compute threshold to report safety incidents.[8] Australia would be aligning with international partners by establishing a comparable capability.

An AI incident reporting mechanism could start small — voluntary reporting housed within an existing agency — and scale to mandatory reporting as the framework matures.

Conclusion

The proposed amendments to Part 3 of the SOCI Act are a welcome step. Measures 1 and 3 in particular will strengthen Government's ability to respond to AI-related national security risks. To make the vendor-restriction power effective for AI, it should be supported by transparency obligations on frontier AI developers and a national AI incident reporting mechanism.

[7] European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 (AI Act)*, Article 73.
[8] California State Legislature. (2025). *Senate Bill 53: Artificial Intelligence Models: Large Developers*.