Seven Reasons Australia Needs to Act on AI Risk

Australia's artificial intelligence (AI) policy debate is at a crossroads. Decisions made in the coming months will determine whether Australians are able to trust AI and capture its benefits or whether we are left exposed to unacceptable and avoidable risk.

Calls to wait for yet further reviews ignore the evidence that action is needed now. Australia cannot afford to repeat mistakes made with previous technological developments, such as social media, where delayed action left communities exposed and regulators playing catch-up.

Global Shield Australia **has prepared the attached primer providing *Seven Reasons Australia Needs to Act on AI Risk*.** Now is the time for the government to deliver the safeguards that Australians expect and need to fully capture the AI dividend.

SEVEN REASONS AUSTRALIA NEEDS TO ACT ON AI RISK

1. **To address the unique hazards associated with foundation AI models and the systemic risk created by their widespread deployment**, this includes from their demonstrated capabilities to deceive, self-replicate, and pursue their own goals.
2. **To establish monitoring and incident reporting of AI in Australia**, to enable the government to systemically track and respond to AI-related harms across the entire economy.
3. **To provide a consistent baseline for regulation of AI across the economy and resolve potential inconsistencies and uncertainties between regulatory regimes,** including on allocating legal responsibility for harm.
4. **To mandate content provenance and labelling at scale** by regulating generative models, a key tool to deal with harms ranging from deepfakes to disinformation.
5. **To establish specific security standards and obligations to prevent the exfiltration and misuse of advanced and high-risk AI models and applications by rogue actors.**
6. **To deliver uniform assessment and certification of AI models and tools**, making it easier for small businesses and users to use and trust the technology.
7. **To align Australia with global partners and ensure we can lead in our region**, minimising compliance friction and enabling Australian businesses to access global AI assurance opportunities.

For more information about this document, please contact australia@globalshieldpolicy.org.

Seven Reasons Australia Needs to Act on AI Risk

*Australia's existing regulatory frameworks were developed without AI in mind and cannot, even with amendment, provide the economy-wide, forward-looking safeguards required for such a transformative technology. A holistic approach to AI regulation is needed to ensure Australia can innovate with confidence while protecting the public from systemic harm. This **Global Shield Australia** primer sets out seven reasons why Australia needs to take action on AI risk by regulating advanced AI models.*

1. Systemic risk: Regulation is needed to address the unique hazards associated with AI models and the systemic risk created by their widespread deployment.

Decisions made during the development and training of frontier AI models have cross-cutting and systemic implications. Any defects in a model are inherited across all of its downstream applications. A single AI model can also be used in applications across multiple industries and use cases. This means that one algorithmic flaw, biased dataset, or security vulnerability can cause widespread, systemic harm or simultaneous failures across sectors and regulators.

Amendments to existing regulatory frameworks can potentially manage specific AI harms as they arise in the use of AI in specific sectors. However, they cannot effectively address hazards embedded within the foundation AI models themselves. These hazards include:

1. **Deception:** AI models deliberately misleading users about their intentions or actions.
2. **Jailbreaking**: users bypassing safeguards to make AI tools produce harmful outputs.
3. **Hijacking:** AI agents being manipulated by hostile actors when engaging with public material or applications and pursuing instructions contrary to their user—such as to disclose personal or sensitive information.
4. **Self-propagation and escape**: AI models seeking to copy themselves without authorisation, creating potential loss control threats.
5. **Autonomous goal-seeking:** AI models pursuing objectives against their users' or developers' intent or resisting attempts to shut them down.
6. **Training data poisoning:** malicious or flawed training data creating hidden vulnerabilities in the resulting AI model.

These are not hypothetical concerns. These issues have already been observed in testing of frontier models. Existing domain specific laws, such as for consumer protection, generally focus on addressing harm after it occurs; but they are less suited to imposing pre-deployment duties on AI model developers in relation to their models. As such, the most efficient approach to addressing these AI model hazards is at their source – namely during model design, testing, and deployment.

Only AI-specific regulation can ensure AI model-specific hazards are addressed systemically and at the point of greatest impact. This includes requirements to: test and certify AI models before release; put in place safety measures as part of model development; and disclose information regarding training and evaluations to regulators.

2. **Monitoring:** Regulation is needed to establish monitoring and incident reporting of AI in Australia, enabling the government to systemically track and respond to AI-related harms across the entire economy.

An AI model does not fail in isolation. Failures at the model level can surface across multiple industries and deployments given the relatively few foundation AI models that power a broad range of AI applications. Harms are also likely to emerge unpredictably from AI being used in untested contexts, unforeseen user interactions, and integration with other tools.

At present, it is largely only AI developers who have full knowledge of AI failures, harms, and early warning signs. They are under no obligation to systematically report AI incidents. This can result in multiple regulators facing what appear to be single failures but missing a potentially systemic hazard. It also means that companies, regulators, and insurers undertake risk assessments of AI deployments based on disjointed reporting, third-party expert analysis, and voluntary disclosures by the major AI companies.

This information asymmetry and regulatory fragmentation means that Australia lacks a proper understanding of AI risk and a capacity to respond appropriately when harm does materialise.

AI-specific regulation can address this gap by mandating monitoring and reporting obligations for AI developers and deployers. This could include measures such as registration of high-risk systems, adverse incident and near miss reporting requirements, and voluntary reporting pathways for users. Similar systems are already used for a range of products and industries in Australia, including automobiles, airplanes, medical devices, and consumer goods.

Australian reporting requirements could also align with obligations these companies are already subject to overseas, lowering any compliance burden by leveraging work companies are already undertaking.

3. **Legal clarity:** Cross-cutting regulation is needed to provide a consistent baseline for regulation of AI across the economy and resolve potential inconsistencies and uncertainties between regulatory regimes, including on allocating responsibility for harm.

At present, the same AI model or application can fall under multiple regulatory frameworks depending on how and where it is deployed. Without a regulatory baseline, each regulator can apply different definitions and requirements to these systems, meaning the same AI system could be subject to multiple and duplicative regulatory requirements.

Current regulatory frameworks may struggle to apportion responsibility in complex AI supply chains involving foundation model developers, application developers, and end-user deployers. Upstream developers can also use contractual terms to improperly shift liability, even if they are the ones best placed to address or remedy the harm.

This can create uncertainty for developers, deployers, and users, raise compliance costs, and undermine effective safeguards.

Cross-cutting regulation of AI can provide a uniform baseline of definitions, standards, and general duties, ensuring regulatory consistency across all sectors. It can also establish a consistent, economy-wide baseline framework for allocating legal responsibility for actions by AI tools.

This would reduce opportunities for regulatory arbitrage, and ensure Australians are protected by minimum, clear, and coherent rules, no matter where or how AI is deployed. It would also enable innovation, not chill it. By setting clear limits and obligations, cross-cutting regulation would allow safe experimentation and protect responsible innovators from being undercut by less responsible actors.

4. **Labelling:** Regulation is needed to mandate content provenance and labelling at scale.

The rapid rise of generative AI is making it increasingly difficult to determine whether images, video, or audio are real or artificially generated. Existing laws can prohibit harmful content and regulate particular uses of these tools, but they cannot easily resolve the underlying problem of provenance – knowing what has been created by AI in the first place.

Regulation is needed to mandate that developers and deployers of generative AI models and applications embed provenance signals such as metadata, watermarking, or equivalent measures at the source. This would allow regulators, platforms, and the public to detect AI-generated media at scale, strengthening safeguards against misinformation, fraud, and other forms of abuse and misuse.

5. **Security:** Regulation is needed to establish security standards and obligations to prevent the exfiltration and misuse of advanced and high-risk AI models by rogue actors.

Frontier AI models are prime targets for theft and misuse by criminal groups or hostile State and non-State actors. Without dedicated safeguards, critical components such as model weights, training data, or deployment architectures could be stolen, leaked, or repurposed for malicious use. Existing security regulations are generally not designed to address these risks at the model level.

For example, a single model could raise distinct national security concerns across multiple domains. It could be used to enable biological weapons threats while also enhancing cyber attack capabilities. Without a coherent framework for the underlying model, these threats would need to be addressed repeatedly under multiple regulatory frameworks.

Regulation is needed to mandate minimum security controls for advanced and high-risk models, ensuring protections are in place before deployment. It can also establish a clear taxonomy of "high-risk" and "nationally significant" AI systems, providing consistent obligations across other regulatory regimes.

6. **Certification:** Regulation is needed to deliver uniform assessment and certification of AI models and tools.

Businesses and consumers need confidence that AI systems meet consistent safety and security standards. Without a uniform framework, industry and consumers risk confusion, fragmented certifications, and "safety washing" by companies making unverified claims. Australia will soon be applying mandatory cyber security standards and voluntary labelling to smart devices such as smart speakers. Far more powerful AI models and applications with higher levels of risk deserve at least the same level of assurance for consumers and businesses.

Regulation is needed to establish clear conformity assessment processes and trusted certification schemes – such as an AI Safety or Trust Mark – that apply across all sectors and supply chains. This would provide a single, recognisable signal of compliance, making it easier for small businesses to adopt AI safely and for users to trust the technology.

7. **Global leadership:** AI-specific regulation would align Australia with global partners and enable us to influence and lead regional and global practice on AI governance.

While some might oppose AI regulation as potentially "scaring" leading AI developers away from Australia, these companies are already subject to and operating under regulation in jurisdictions such as the European Union and California. In our own region, Vietnam, Malaysia and Thailand are all moving ahead with risk-based AI legislation, including elements inspired by global best practice such as incident reporting.

Even in the United States, recent federal efforts to ban state-led regulation of AI were broadly rejected by Congress with a vote of 99 to 1 in the US Senate. The world's leading AI jurisdiction – California – also recently passed a landmark law on AI transparency.

Appropriately drafted AI regulation would align Australia with the relevant elements of those existing regulatory frameworks, and with OECD and G7 principles. It would also enable us to lead in our region and globally. With most major AI companies already needing to comply with requirements in other jurisdictions, there is no reason why Australians shouldn't also benefit from similar protections here.

Without regulation of AI, Australia risks becoming a mere rule-taker or jurisdiction of convenience. By aligning with international best practice, we can reduce compliance friction, give Australian firms access to global assurance ecosystems, and ensure our businesses can compete and collaborate on a level playing field. AI regulation would also provide credibility to the government's expressed ambition to "become a global leader" in trusted, secure, and responsible AI.

About Global Shield Australia

Global Shield Australia is an independent, non-profit organisation dedicated to reducing global catastrophic risk. We advocate for credible and effective regulation to minimise AI risk and maximise its benefits. For more information on this brief or our work, please contact australia@globalshieldpolicy.org.