Building Trust: The Case for Mandatory AI Safety Testing and Standards

The Policy Challenge

Artificial intelligence (AI) presents extraordinary opportunities for Australia. But realising those opportunities requires addressing some key challenges:

1. **A lack of public trust inhibits options.** High-profile failures and the opaque nature of many AI products make it difficult for consumers to believe they are safe, fair, or reliable.
2. **Businesses (particularly small and medium enterprises) lack access to authoritative evaluations of the safety and security of AI products.** This stifles innovation and investment, and risks Australia becoming a dumping ground for lower-quality, untested AI systems.
3. **Receiving our share of the 'AI dividend' requires supporting the development of Australian AI ancillary industries and services**, in areas we have a competitive advantage.
4. **The global nature of AI development and deployment means that domestic action alone is insufficient**. A 'race to the bottom' overseas could undermine protections Australia puts in place domestically.

The Proposed Solution

Australia should invest in a robust AI assurance ecosystem. This will support safe adoption and lift domestic capability. It will also build on our existing expertise to create new export opportunities to markets with strong assurance requirements (such as the European Union).

Testing and standards are proven ways of ensuring safety and building public trust. It is also common across mature, high-impact industries. For AI regulation, this should include:

1. **A robust National AI Safety and Security Standard (NAISSS)** that builds on the existing Voluntary AI Safety Standard and is developed in collaboration with Standards Australia, the National AI Centre, industry and experts. This should be aligned with international best practice and provide a clear benchmark for what Australia considers to be safe and responsible AI.
2. **A National AI Assurance Framework** that includes mandatory third-party testing and evaluation of high-risk AI systems, combined with formal accreditation of AI auditing and evaluation solutions to foster a domestic assurance industry.
3. **Active engagement in bilateral, regional and multilateral collaborations to shape global norms and standards**, including through contributions to international AI safety institute summits and building on our extensive existing international engagements.

Box 1. Why mandate and promote third-party assurance for AI systems in Australia?

Complying with standards and third-party testing is common across industries and builds trust:

"For many years, we've had systems that aim to ensure that products and services meet basic safety, reliability and other legal standards. There is an increasing need to apply similar processes to the deployment of AI, so people can know they are trustworthy and reliable." (Human Technology Institute)

"This type of industry-wide testing approach isn't unusual – most important sectors of the economy are regulated via product safety standards and testing regimes, including food, medicine, automobiles, and aerospace." (Anthropic)

AI assurance is an economic opportunity:

"By the year 2030, we estimate the global [AI Assurance Technology] *market could reach approximately USD $276 billion."* (Juniper Ventures)

"...our research shows that adopting [Responsible AI] *practices can enhance business competitiveness. 79% of organisations see responsible AI as providing either a slight or significant competitive advantage. This perception is even stronger among more mature organisations, with 91% of the Leading segment recognising RAI's competitive benefits."* (Fifth Quadrant)

Australia has the expertise required to grow an AI assurance industry:

"We have the largest group of responsible AI experts in the world." (CSIRO)

"The AI Assurance Lab is validating AI technologies with respect to: Quality (accuracy, robustness, interpretability, usability and bias); Safety (response to attacks or manipulation by adversaries); Privacy; Reliability (sensitivity to perturbations in inputs and the environment)." (University of Melbourne AI Assurance Lab)

Industry supports and is investing in assurance:

Major AI developers (including Amazon, Anthropic, Google, Meta, Microsoft, and OpenAI) are funding capability assessments and organisations such as the Frontier Models Forum to identify best practice and support standards development for AI safety and security.

"*We believe that the AI sector needs effective third-party testing for frontier AI systems. Developing a testing regime and associated policy interventions based on the insights of industry, government, and academia is the best way to avoid societal harm—whether deliberate or accidental—from AI systems."* (Anthropic)

About Global Shield Australia

Global Shield Australia is an international advocacy organization dedicated to reducing global catastrophic risk. We advocate for credible and effective regulation of artificial intelligence to reduce its potential for harm and thus ensure that its opportunities can be fully realised.

For more information on this proposal or our work, please contact Devon Whittle at devon.whittle@globalshieldpolicy.org or on 0458 980 372.