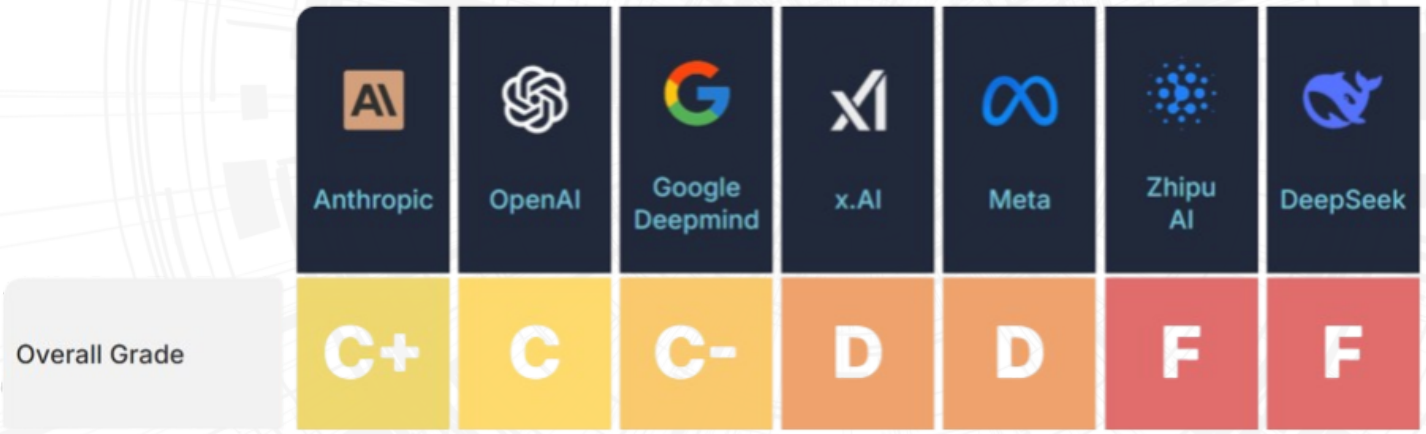
Building Trust to Secure Australia's AI Future

Australia stands at a critical juncture. Highly capable AI systems may soon bring profound economic transformations. The Tech Council of Australia estimates AI could contribute between $45 billion and $115 billion annually by 2030. This estimate is conservative compared to Google ($280 billion), Kingston AI Group ($200 billion), and McKinsey (up to $600 billion).

However, these economic benefits are not guaranteed. The main factor influencing Australia's path is trust.

Australians do not trust AI. Ipsos found that 69% of Australians are apprehensive about AI—the least trusting anywhere in the world. The University of Queensland found 80% of Australians fear negative outcomes, particularly catastrophic risks.

Australians are justified in not trusting AI. AI is already causing real-world harms, and experts agree that highly capable AI systems could even pose an existential risk. Despite this, AI labs are trying to “move fast and break things”, with independent watchdogs finding that "*none of the companies has anything like a coherent, actionable plan*" to manage safety concerns.



In addition to real risks of harm, the trust deficit has economic implications. The gap between the Tech Council’s slow ($45 billion) and fast ($115 billion) adoption scenarios shows that **low trust in AI could cost Australia more than $70 billion annually**.

Trust-building must be prioritised, and trust-building requires credible government action.

Recommendation 1: Establish a Well-Funded Australian AI Safety Institute

The UK, US, Japan, EU, Singapore, Israel, India, France, South Korea, Canada, Germany, and Brazil have established AI Safety Institutes. **Australia and Kenya are the only members of the network of AI Safety Institutes that don't have an AI Safety Institute.**

Establishing an adequately funded AI Safety Institute is crucial for:

- Demonstrating credible government action on AI safety to directly address Australians' concerns.
- Building sovereign technical capability to evaluate AI risks independently of foreign or commercial interests.
- Catalysing a domestic AI assurance sector, projected globally to reach USD $276 billion by 2030.
- Proactively preventing AI failures, reducing harm and fostering confidence.

The investment required—approximately $15 million annually—is modest compared to the tens of billions in economic benefits achievable through enhanced public trust.

Recommendation 2: Introduce Mandatory Guardrails Focused on Developer Responsibility

Australia needs binding AI safety regulations placing responsibility on those best equipped to manage risks: AI developers. Australian businesses currently face significant uncertainty because developers supply "black box" systems and transfer liability through Terms of Service agreements, exposing deployers to risks they cannot effectively manage.

An analogy with aviation safety illustrates this principle: aircraft manufacturers must meet rigorous safety standards before planes reach operators. Similarly, **AI developers should bear primary responsibility for mitigating AI risks before deployment.**

Mandatory guardrails should:

- Define and prohibit unacceptable AI risks, such as models that enable the development of bioweapons or sophisticated cyberattacks.
- Impose clear obligations on developers, preventing "risk-shifting" via contract terms.
- Mandate transparency reporting and independent evaluation (red-teaming) of high-risk AI systems.
- Create mandatory incident reporting mechanisms to systematically capture AI failures and improve regulatory responses.

By clearly delineating responsibilities, these guardrails will enhance trust without significantly impeding domestic AI adoption.

Learn more at https://www.goodancestors.org.au/ai-safety