sparke HELMORE LAWYERS

Voluntary AI Safety Standard

A set of voluntary principles to guide safe and responsible use of Artificial Intelligence (AI) in Australia

Published by the Department of Industry, Science and Resources in September 2024, the Voluntary AI Safety Standard (**Standard**) comprises 10 guardrails (**Voluntary Guardrails**) that apply to all organisations throughout the AI supply chain.

The intention is that by adopting the Voluntary Guardrails, organisations can use and innovate with AI in a consistent, safe, and responsible way.

The Voluntary Guardrails

- Implement and publish an **accountability process including** governance, and regulatory compliance strategies
- Establish and implement a **risk management process** to identify and mitigate risk
- Protect AI Systems and implement **data governance** measures to manage data quality and provenance
- **Test AI models** and systems to evaluate model performance and monitor the system once deployed
- **Enable human control** or intervention in an AI system to achieve meaningful human oversight
- **Inform end users** regarding AI-enable decision, interactions with AI and AI-generated content
- Establish processes for people impacted by AI systems to **challenge use or outcomes**
- **Be transparent** with other organisations across the AI supply chain about data, models, and systems to help effectively address risk
- **Keep and maintain records** to allow third parties to assess compliance with guardrails
- **Engage your stakeholders** and evaluate their needs and circumstances with a focus on safety, diversity, inclusion, and fairness

Human Centred Approach

The Standard aligns with Australia's AI Ethics Principles.

It aims to protect people's rights, promote diversity, inclusion, and fairness, prioritise human-centred design and encourage deployment of trustworthy AI systems to support social licence.

Risk-based Approach

The Standard promotes a risk-based approach to AI harm prevention, emphasising proactive measures to identify and mitigate risk.

This includes risk assessments, management frameworks and prohibition on certain activities with unacceptable risks.

Global Interoperability

The Standard is consistent with ISO/IEC 42001:2023 and the US standard on AI risk management, NIST AI RMF 1.0.

Key Takeaways

Best Practice

Although the Standard is *voluntary*, compliance is viewed by many as best practice. Organisations should consider taking the following steps to implement the Standard:

- refreshing internal governance frameworks and policies;
- ensuring contractual terms with suppliers address AI use and mitigate risk;
- ensuring compliance with existing laws when using AI (e.g. product safety, financial services regulation, and privacy laws); and
- updating communications with end users to disclose AI use

Focus on Deployers

The initial focus is on **deployers of AI** (i.e., organisations that supply or use an AI system to provide a product or service). A future version will focus on developers of AI.

Procurement Guidance

The Standard provides guidance for engaging with suppliers. However, we anticipate further guidance to be released towards the end of the year.