Proposed Mandatory Guardrails

Proposed Mandatory Guardrails for high-risk Artificial Intelligence (AI) Systems

Published by the Department of Industry, Science and Resources for consultation in September 2024, the Mandatory Guardrails (**Mandatory Guardrails**) would apply to deployers, and developers of AI used in **high-risk** settings.

The Mandatory Guardrails require developers and deployers of high-risk AI to take proactive measures to ensure their products are safe and to reduce the likelihood of harm arising.

The Mandatory Guardrails

1. Implement and publish an **accountability process including** governance, and regulatory compliance strategies
2. Establish and implement a **risk management process** to identify and mitigate risk
3. Protect AI Systems and implement **data governance** measures to manage data quality and provenance
4. **Test AI models** and systems to evaluate model performance and monitor the system once deployed
5. **Enable human control** or intervention in an AI system to achieve meaningful human oversight
6. **Inform end users** regarding AI-enable decision, interactions with AI and AI-generated content
7. Establish processes for people impacted by AI systems to **challenge use or outcomes**
8. **Be transparent** with other organisations across the AI supply chain about data, models, and systems to help effectively address risk
9. **Keep and maintain records** to allow third parties to assess compliance with guardrails
10. **Undertake conformity assessments** to demonstrate and certify compliance with the guardrails

What is High-Risk AI?

The definition of 'high-risk' is under consultation however, the Government has proposed that it covers two main categories of AI:

Category 1 – Foreseeable and Intended use

When assessing if an AI system is high-risk, regard should be had to the risk of adverse impacts on:

- Human rights
- Physical and mental health
- Legal effects, defamation, or similar effects
- Collective rights of a group
- Societal risk
- The severity and extent of the above

Category 2 – General Purpose AI

AI models that are capable of being used for a variety of purposes, both for direct use as well as for integration in other systems (e.g., ChatGPT and DALL-E).

General purpose AI also includes Agentic AI, which operates autonomously and can be used to send emails, instructions, or act as virtual web-browsing assistant.

In addition to having the potential for causing foreseeable risks, general purpose AI also poses unforeseeable risk because it can be applied in contexts it was not originally designed for.

The first 9 Mandatory Guardrails are the same as the first 9 Voluntary Guardrails. The 10th Mandatory Guardrail is different and emphasises the importance of undertaking conformity assessments to certify compliance with the Mandatory Guardrails.

How will the Australian Government implement the Mandatory Guardrails?

The Australian Government is exploring ways to implement the Mandatory Guardrails

- **Domain Specific Approach** – adapting existing regulatory frameworks to incorporate the Mandatory Guardrails
- **Framework Approach** – introducing framework legislation that will amend existing laws to align with the framework
- **Whole of Economy Approach** – introducing a new cross-economy AI Act to apply across all economic sectors and an independent AI regulator

What's next following the initial consultation?

Responses to the initial consultation closed on 4 October 2024. Since then, industry feedback has been published but no indication of next steps has been announced. Notably, the Government has indicated that they will continue to strengthen and clarify existing laws, to improve their applicability to AI models and systems.

In preparation, organisations should consider taking steps to implement the Mandatory Guardrails, including developing internal AI governance frameworks and policies, and ensuring contractual terms with suppliers / customers / users address AI use and risks.