**Table of Contents**

[Executive Summary 2](#_Toc209692259)

[1.0 Introduction 3](#_Toc209692260)

[1.1 Scope 3](#_Toc209692261)

[1.2 Methodology 3](#_Toc209692262)

[2.0 Findings 4](#_Toc209692263)

[2.1 Critical Analysis of Australian High-Risk AI Principles 4](#_Toc209692264)

[2.1.1 Individual Principle Evaluation 4](#_Toc209692265)

[2.1.2 Framework Strengths and Critical Gaps 5](#_Toc209692266)

[2.2 Framework Adequacy Assessment 6](#_Toc209692267)

[2.2.1 Are the Six Principles Adequate? 6](#_Toc209692268)

[2.2.2 International Comparison Analysis 7](#_Toc209692269)

[2.3 Specific Framework Enhancement Recommendations 7](#_Toc209692270)

[2.3.1 Principles That Must Be Added 7](#_Toc209692271)

[2.3.2 Principles Requiring Substantial Revision 8](#_Toc209692272)

[2.3.3 Principles Requiring Consolidation 8](#_Toc209692273)

[2.4 Effects and Benefits of the Enhanced Australian Framework 8](#_Toc209692274)

[2.4.1 Improved Risk Detection and Prevention 8](#_Toc209692275)

[2.4.2 Enhanced Protection for Vulnerable Communities 9](#_Toc209692276)

[2.4.3 Stronger International Competitiveness 9](#_Toc209692277)

[2.4.4 Democratic and Institutional Resilience 9](#_Toc209692278)

[2.4.5 Economic and Innovation Benefits 10](#_Toc209692279)

[3.0 Conclusion 10](#_Toc209692280)

[4.0 Acknowledgements 10](#_Toc209692281)

[5.0 Reference 11](#_Toc209692282)

# Executive Summary

The rapid development of Artificial Intelligence (AI) creates an urgent need for robust frameworks to identify high-risk applications requiring regulatory oversight. This report evaluates Australia's proposed six principles for high-risk AI identification, comparing them with Malaysia's approach and different perspectives from academic literature. The Australian framework establishes comprehensive risk assessment criteria across human rights, health, legal effects, cultural impacts, systemic effects, and impact severity.

The analysis reveals significant strengths in the Australian framework's technology-agnostic design, international alignment, and multi-dimensional risk assessment approach. However, critical limitations include definitional ambiguity, reliance on fragmented human rights architecture, and insufficient attention to emerging AI capabilities. Therefore, recommendations include enhanced definitional clarity, dynamic assessment mechanisms, strengthened cultural protections, and comprehensive implementation support, which will be proposed to enhance the Australian framework for effective governance of high-risk AI systems.

# Introduction

Artificial Intelligence (AI) has rapidly evolved from an emerging research field into one of the most transformative technologies of the 21st century. It is expected to contribute over USD 13 trillion to the global economy by 2030, creating new industries [7]. However, alongside these benefits, AI also brings significant risks including biased algorithms, accountability gaps, the misuse of generative systems, and even the potential emergence of superintelligence [4]. Such concerns highlight the importance of establishing robust principles to capture these high-risk AI.

A central concern in current discussions of AI governance is the concept of high-risk AI. High-risk AI refers to systems that, if misused or poorly regulated, have the potential to cause serious harm to individuals or society in terms of human rights, public safety, or mental health [3]. Clearly identifying and categorising high-risk AI is critical because it enables governments and organisations to apply stricter oversight and ensure that AI is designed, deployed as well as monitored in ways that minimise potential harms. Without appropriate principles, the rapid expansion of AI could easily outpace the capacity of regulators and institutions to protect society.

The need to capture and regulate high-risk AI is urgent for several reasons. First, AI has already become part of everyday life: people make use of it for learning, assessment, and decision-making, where mistakes or bias in it can harm vulnerable individuals. Second, the scalability of AI means that even small design flaws can produce widespread harm when deployed at scale. Third, because AI is developed and adopted globally, differences in national standards and regulations can create inconsistencies and accountability gaps. Together, these challenges underscore the importance of assessing whether existing governance frameworks are sufficient or whether new principles are needed.

## Scope

The scope of this report is limited to a critical examination of principles designed to identify and categorise high-risk AI systems, specifically evaluating their comprehensiveness and effectiveness in capturing contemporary AI risks. The primary focus will be on evaluating the six principles proposed by Australian Government, Department of Industry, Science and Resources [3], with a comparative analysis of Malaysia's National Guidelines on AI Governance and Ethics [7]. The report will not attempt to review all global frameworks but will instead narrow its focus to principles that specifically address the risks associaed with high-risk AI. Special attention will be given to whether these principles sufficiently address issues of human rights, societal well-being as well as long-term impacts.

## Methodology

This research adopts a qualitative approach, drawing primarily on policy document analysis and comparison with academic literature. The Australian and Malaysian governance frameworks will be reviewed based on ethical considerations highlighted by Burton et al. [4], as well as other relevant scholarly and policy sources. The aim is to evaluate whether the proposed principles are adequate for mitigating the challenges of high-risk AI, to determine if revisions are necessary, and to consider whether additional principles should be introduced or existing ones removed.

# Findings

## Critical Analysis of Australian High-Risk AI Principles

### Individual Principle Evaluation

The Australian Department of Industry, Science and Resources has proposed six principles to identify high-risk AI requiring mandatory guardrails [3]. Each principle demonstrates specific strengths and critical limitations that affect the framework's overall adequacy.

#### Principle A: Human Rights Protection

This principle targets "adverse impacts to individual rights recognised in Australian human rights law without justification." While aligning with international approaches, this principle contains a notable limitation that the document acknowledges: Australia's fragmented human rights architecture. Australia lacks comprehensive federal human rights legislation, attempting to compensate by referencing international obligations [5] unlike the EU Charter of Fundamental Rights or the Canadian Charter. This approach creates uncertainty about which rights take precedence when conflicts arise. The principle's narrow focus on individual discrimination cases, such as age, disability, race, and sex, reflects existing legislation but is less comprehensive than international equivalents. Furthermore, while providing examples of discriminatory AI systems in employment, criminal justice, and facial recognition, the principle fails to address emerging algorithmic rights concerns like transparency requirements and automated decision-making protections increasingly recognised in frameworks like the EU AI Act.

#### Principle B: Health and Safety

This principle addresses "adverse impacts to physical or mental health or safety," appropriately encompassing both traditional harms and contemporary mental health impacts. However, the principle inadequately addresses population-level health effects, such as cumulative psychological impacts from social media algorithms or workplace surveillance systems. Contemporary health risks extend beyond medical misdiagnosis to include chronic stress from AI monitoring systems and psychological impacts of AI-generated content. The principle also lacks consideration of long-term health implications from prolonged AI interaction, including potential addiction to AI-driven platforms, erosion of human decision-making capabilities, and impacts on cognitive development in children exposed to AI systems from early ages.

#### Principle C: Legal Effects

This principle focuses on "adverse legal effects, defamation or similarly significant effects." The phrase "similarly significant effects" creates substantial enforcement challenges due to definitional vagueness. The principle fails to address emerging legal concerns including AI-generated evidence in court proceedings or the cumulative effects of multiple AI systems affecting the same individual, where combined impacts may exceed individual system significance. Additionally, the principle does not adequately consider how AI systems might undermine procedural justice through automated decision-making that lacks transparency or appeal mechanisms, potentially violating fundamental principles of natural justice embedded in Australian legal traditions.

#### Principle D: Impact on Group

This principle addresses "adverse impacts to groups or collective cultural rights" and represents progressive inclusion of collective harm assessment. However, it remains significantly underdeveloped, lacking specific guidance for measuring collective harms and failing to address Indigenous data sovereignty, which is a critical omission given Australia's substantial Indigenous population [10]. The principle inadequately recognises that effective Indigenous protection requires acknowledgment of data sovereignty and self-determination principles. Moreover, the principle fails to address how AI systems might discriminate against people with multiple disadvantaged identities, such as elderly ethnic minorities or disabled women, creating compounded discrimination that is worse than the sum of individual biases.

#### Principle E: Systemic Impacts

This principle demonstrates sophisticated understanding of societal-level threats through comprehensive analysis of democratic risks. However, its excessive breadth potentially captures numerous AI applications posing minimal systemic risk while lacking specific metrics for measuring impact severity. The principle also inadequately addresses cumulative effects of multiple AI systems operating simultaneously across society. The framework's analysis of democratic threats, while comprehensive, fails to consider how AI systems might reshape social structures, labour markets, and educational systems in ways that fundamentally alter Australian society's character and values.

#### Principle F: Severity and Extent of Impact

This principle functions as a meta-principle requiring impact severity assessment. While providing useful analytical structure, it essentially operates as general risk methodology rather than AI-specific guidance, lacking concrete thresholds for determining high-risk classification. The principle provides insufficient guidance on how to weigh different types of harms against each other or assess risks that may be low probability but catastrophic in consequence, such as AI systems that might contribute to democratic breakdown or large-scale social manipulation.

### Framework Strengths and Critical Gaps

The framework's technology-agnostic design provides flexibility for emerging AI capabilities while maintaining consistent assessment standards. International alignment with harm-based approaches supports interoperability and reduces compliance burdens, reflecting global convergence around core AI governance principles [1][5]. The comprehensive scope attempting to address individual, collective, and systemic impacts represents sophisticated understanding of AI risk complexity.

However, three critical gaps undermine adequacy: definitional inadequacy across multiple principles creates interpretation challenges that could enable regulatory arbitrage; structural omissions exclude emerging high-risk categories such as environmental impacts and algorithmic transparency requirements that are increasingly central to international AI governance; implementation assumptions about organisational capacity may prove unrealistic for smaller enterprises, potentially creating a two-tier system where only large organisations can effectively navigate compliance requirements.

## Framework Adequacy Assessment

### Are the Six Principles Adequate?

The proposed principles are fundamentally inadequate for capturing contemporary high-risk AI systems, based on four key deficiencies:

#### Incomplete Risk Coverage:

The framework omits several recognised high-risk categories. Environmental impacts from energy-intensive AI systems present significant concerns due to AI's growing carbon footprint, particularly given large language models' energy consumption equivalent to small cities, yet these impacts are entirely absent from the framework. Algorithmic transparency requirements, fundamental to accountability, lack dedicated assessment criteria despite their recognition as essential for democratic oversight. Cross-jurisdictional governance challenges remain unaddressed, creating potential gaps where multinational AI deployments might exploit regulatory arbitrage.

#### Implementation Limitations:

The framework assumes sophisticated risk assessment capabilities that many organisations lack. Research indicates smaller enterprises often lack technical capacity for comprehensive risk evaluation [6]. Without graduated approaches, the framework may prove inaccessible to significant portions of the AI ecosystem, potentially stifling innovation in smaller companies while advantaging large technology corporations with extensive compliance resources.

#### Cultural Inadequacy:

While mentioning collective rights, the framework inadequately addresses Indigenous data sovereignty and minority community protection. Effective AI governance for Indigenous communities requires explicit recognition of collective data rights largely absent from the current framework [10]. Research on algorithmic fairness demonstrates that discrimination can be embedded in seemingly neutral technical processes, requiring explicit attention to bias prevention and mitigation strategies [8]. The individualistic orientation may prove fundamentally incompatible with Indigenous concepts of collective responsibility and community-controlled development.

#### Dynamic Assessment Gaps:

Contemporary AI systems create evolving risk profiles based on deployment context and user interaction [9]. The framework's static assessment approach inadequately addresses these dynamic capabilities, particularly concerning for foundation models that can be rapidly adapted to new applications with unpredictable risk implications.

### International Comparison Analysis

Malaysia's National Guidelines emphasise community consultation and cultural sensitivity [7], revealing gaps in Australia's community protection mechanisms. The Malaysian approach includes specific provisions for religious considerations and mandatory community engagement, demonstrating how cultural values can be integrated into AI governance frameworks. The EU AI Act's prohibited practices provide clearer guidance than Australia's principles-based approach, explicitly banning certain AI applications rather than relying solely on risk assessment. Canada's impact assessment requirements offer more structured implementation pathways with specific procedural requirements and accountability mechanisms.

Academic literature supports harm-based approaches but emphasises implementation challenges. Mittelstadt's analysis highlights the gap between ethical intentions and practical outcomes [2], which the Australian framework inadequately addresses through specific implementation guidance or enforcement mechanisms.

## Specific Framework Enhancement Recommendations

### Principles That Must Be Added

#### New Principle G: Environmental and Sustainability Assessment

The framework must include explicit evaluation of environmental impacts from AI systems, addressing critical gaps in sustainability assessment and aligning with Australia's international climate commitments. The principle should assess: (a) direct energy consumption and carbon emissions from AI training, deployment, and inference operations; (b) indirect environmental impacts through supply chain effects, including semiconductor production and data centre infrastructure; and (c) lifecycle environmental costs including hardware production, transportation, and disposal phases. This principle becomes increasingly urgent as AI systems scale, with some foundation models reportedly consuming electricity equivalent to entire cities during training.

#### New Principle H: Algorithmic Transparency and Explainability

A dedicated principle must require assessment of AI system transparency capabilities, addressing the "black box" problem undermining accountability and democratic oversight. The principle should evaluate: (a) availability of meaningful explanations for AI decisions that affected individuals can understand and use for challenge purposes; (b) accessibility of explanation mechanisms to affected individuals, including provision in appropriate languages and formats; and (c) technical documentation sufficient for regulatory oversight, academic scrutiny, and public accountability processes. This principle is essential for maintaining human agency and democratic control over AI systems.

#### New Principle I: Cross-Jurisdictional Governance and Data Sovereignty

The framework must address risks from cross-border data flows and jurisdictional governance gaps, critical for multinational AI deployments. The principle should assess: (a) cross-border data transfer implications and compliance with multiple regulatory regimes; (b) jurisdictional conflicts that may create enforcement gaps or competitive disadvantages; and (c) data sovereignty considerations for affected communities, particularly Indigenous peoples whose data may be used across borders without appropriate governance mechanisms.

### Principles Requiring Substantial Revision

#### Principle A Enhancement:

Current: "Risk of adverse impacts to individual rights recognised in Australian human rights law without justification" can be enhanced to "Risk of violations of internationally recognised human rights standards, incorporating Australia's domestic protections and international treaty obligations, with particular attention to collective rights and Indigenous sovereignty"

This addresses the fragmented domestic rights architecture by anchoring assessment in comprehensive international standards while explicitly recognising collective and Indigenous rights often overlooked in individualistic frameworks. The revision incorporates Australia's obligations under international treaties including the UN Declaration on the Rights of Indigenous Peoples.

#### Principle D Expansion:

The principle should explicitly incorporate Indigenous data sovereignty requirements, including: (a) mandatory consultation processes with affected Indigenous communities that respect traditional decision-making processes; (b) recognition of collective data rights under frameworks like the CARE Principles (Collaborative, Accessible, Respectful, and Ethical) for Indigenous Data Governance; and (c) specific protection mechanisms for cultural knowledge and traditional governance systems that AI systems might affect, including sacred site information and traditional ecological knowledge.

### Principles Requiring Consolidation

#### Merge Principles C and F:

The legal effects and severity assessment principles demonstrate significant overlap in impact measurement focus. A consolidated principle addressing "Legal Effects and Impact Severity Assessment" could reduce framework complexity while improving coherence by integrating legal consequence evaluation with systematic impact assessment methodology.

## Effects and Benefits of the Enhanced Australian Framework

### Improved Risk Detection and Prevention

The enhanced framework with three additional principles would significantly strengthen Australia's capacity to identify and prevent AI-related harms before they occur. Environmental assessment requirements would prevent deployment of unsustainable AI systems, supporting Australia's climate commitments while encouraging innovation in energy-efficient AI technologies that could become export opportunities. Algorithmic transparency mandates would enable earlier detection of biased or flawed AI systems, reducing the likelihood of discriminatory outcomes in critical sectors like healthcare, employment, and criminal justice while maintaining public trust in AI deployment.

Cross-jurisdictional governance provisions would close regulatory gaps that currently allow harmful AI systems to exploit jurisdictional arbitrage, ensuring consistent protection for Australians regardless of where AI systems are developed or hosted. This comprehensive coverage would position Australia as a global leader in responsible AI governance while protecting citizens from emerging AI risks and potentially attracting international investment in responsible AI development.

### Enhanced Protection for Vulnerable Communities

The revised framework would provide substantially stronger protection for Indigenous Australians and minority communities currently underserved by existing principles. Mandatory Indigenous consultation processes would ensure that AI systems affecting Indigenous communities respect traditional governance structures and data sovereignty principles, preventing digital colonisation and supporting Indigenous self-determination. Collective rights recognition would address systemic discrimination that disproportionately affects marginalised groups through algorithmic bias, providing mechanisms for community-level harm assessment and remediation.

Community-controlled oversight mechanisms would empower local communities to monitor AI deployment in their contexts, ensuring that AI systems serve community interests rather than imposing external technological solutions that may undermine local values or governance systems.

### Stronger International Competitiveness

The enhanced framework would align Australia with leading international AI governance approaches while maintaining distinctive national characteristics. Harmonisation with EU and Canadian frameworks would reduce compliance costs for multinational companies operating in Australia while ensuring Australian organisations can compete globally with recognised responsible AI credentials and access to international markets requiring robust governance frameworks.

Comprehensive risk assessment capabilities would attract international investment in responsible AI development, positioning Australia as a preferred location for companies seeking robust governance frameworks while avoiding regulatory arbitrage that undermines responsible development practices.

### Democratic and Institutional Resilience

The reinforced framework would significantly strengthen Australia's democratic institutions against AI-related threats. Enhanced systemic risk assessment would provide early warning systems for AI systems that threaten electoral integrity, media plurality, or public discourse quality, supporting democratic resilience in an era of increasing information manipulation. Transparency requirements would enable public scrutiny of AI systems affecting democratic processes, from political advertising algorithms to government decision-making systems, maintaining democratic accountability.

Citizen participation mechanisms would ensure that AI governance remains accountable to democratic processes rather than being captured by technical experts or industry interests, preserving democratic legitimacy while enabling informed public discourse about AI development directions.

### Economic and Innovation Benefits

Rather than constraining innovation, the enhanced framework would drive sustainable AI development that creates long-term economic value. Clear regulatory requirements would provide industry certainty, enabling companies to invest confidently in responsible AI development while avoiding the compliance uncertainties that currently hamper investment decisions. Graduated implementation approaches would support smaller enterprises in adopting AI technologies while maintaining appropriate safeguards.

Environmental sustainability requirements would encourage development of energy-efficient AI systems, potentially creating export opportunities for Australian clean AI technologies. Cultural sensitivity provisions would drive innovation in inclusive AI design, creating competitive advantages in diverse global markets and supporting Australia's multicultural values.

# Conclusion

This analysis demonstrates that Australia's proposed six principles for identifying high-risk AI systems are fundamentally inadequate for addressing contemporary AI governance challenges. The assessment reveals four primary deficiencies: incomplete risk coverage that omits environmental impacts, algorithmic transparency requirements, and cross-jurisdictional governance challenges; implementation limitations that assume unrealistic organisational capacity; cultural inadequacy in addressing Indigenous data sovereignty; and dynamic assessment gaps that fail to account for evolving AI capabilities. These deficiencies create significant regulatory gaps that may expose Australians to emerging AI risks while disadvantaging smaller enterprises unable to navigate complex compliance requirements.

Looking at the fundamental question that drives this research, it becomes clear that the proposed principles do not adequately capture high-risk AI. The framework requires substantial enhancement through three additional principles addressing environmental sustainability, algorithmic transparency, and cross-jurisdictional governance, alongside significant revisions strengthening human rights protections and Indigenous data sovereignty recognition. The enhanced framework would transform Australia's approach from reactive harm mitigation to proactive risk prevention, positioning the nation as a global leader in responsible AI governance while strengthening protection for vulnerable communities, improving international competitiveness, and driving sustainable innovation that serves Australian values and community needs.

# Acknowledgements

A screenshot of a computer

AI-generated content may be incorrect.

Make use of AI to understand the three given documents and deepen my understanding as well.

A screenshot of a phone

AI-generated content may be incorrect.

Consult with AI to organise the structure of the research and correct the grammar.

# Reference

1. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019.
2. B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501-507, 2019.
3. Department of Industry, Science and Resources, *Safe and Responsible AI in Australia: Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings*. Canberra, Australia: Commonwealth of Australia, 2024.
4. E. Burton, J. Goldsmith, S. Koenig, B. Kuipers, N. Mattei, and T. Walsh, “Ethical considerations in artificial intelligence courses,” AI Magazine, vol. 38, no. 2, pp. 22–34, 2017.
5. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An Ethical Framework for a Good AI Society," Minds and Machines, vol. 28, no. 4, pp. 689-707, 2018.
6. M. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," Science, vol. 358, no. 6370, pp. 1530-1534, 2017.
7. Ministry of Science, Technology and Innovation (MOSTI), *The National Guidelines on AI Governance & Ethics*. Putrajaya, Malaysia: MASTIC, 2024.
8. S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2019.
9. S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105-114, 2015.
10. T. Kukutai and S. Taylor, *Indigenous Data Sovereignty: Toward an Agenda*. Canberra, Australia: ANU Press, 2016.