A black and red sign with white text

AI-generated content may be incorrect.

**COS30019**

**Introduction to Artificial Intelligence**

**Assignment 1: Research About AI Ethics**

**Topic Chosen: Do the PPs Adequately Capture High-Risk AI? Should Any Principles Be Added or Removed?**

Name: Andrew Teck Foon YII

Student ID: 104386568

Tutorial Group: Group 1

**Table of Contents**

[Executive Summary 2](#_Toc210844816)

[1.0 Introduction 3](#_Toc210844817)

[1.1 Scope 3](#_Toc210844818)

[1.2 Methodology 3](#_Toc210844819)

[2.0 Findings 4](#_Toc210844820)

[2.1 Evaluation of 6 Proposed Principles in Australia’s Framework 4](#_Toc210844821)

[2.2 International Comparison and Australia’s Framework Adequacy 5](#_Toc210844822)

[2.2.1 International Comparison Analysis 5](#_Toc210844823)

[2.2.2 Framework Inadequacy Assessment 6](#_Toc210844824)

[2.3 Specific Framework Enhancement Recommendations 7](#_Toc210844825)

[2.3.1 Three Principles That Must Be Added 7](#_Toc210844826)

[3.0 Conclusion 9](#_Toc210844827)

[4.0 Reference 10](#_Toc210844828)

[5.0 Appendix 11](#_Toc210844829)

# Executive Summary

The rapid development of Artificial Intelligence (AI) has created an urgent need for robust frameworks to identify high-risk applications requiring regulatory oversight. This research critically evaluates Australia's proposed six principles for high-risk AI identification, comparing them with international approaches and examining their alignment with academic literature. Although the Australia’s framework shows considerable strengths, particularly its technology-agnostic design and multi-dimensional risk assessment approach, it still has the limitations in three critical areas of environmental sustainability, algorithmic transparency, as well as indigenous data sovereignty.

This research focuses primarily on proposing three new principles based on the comparisons with international frameworks and research evidence. The enhanced framework presented here would better position Australia as a leader in responsible AI governance while providing more comprehensive protection against emerging AI risks.

# 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 as 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

This research focuses specifically on the critical examination of principles designed to identify and categorise high-risk AI systems. The primary analysis centres on evaluating the six principles proposed by the Australian Government's Department of Industry, Science and Resources [3], with comparative analysis from international frameworks including the European Union (EU) AI Act, Canada's Algorithmic Impact Assessment, and Malaysia's National Guidelines [7]. This comparative approach makes a thorough assessment of whether Australia's proposed principles are adequate or require enhancement.

## 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

## Evaluation of 6 Proposed Principles in Australia’s Framework

The Australian Department of Industry, Science and Resources has proposed six principles to identify high-risk AI requiring mandatory guardrails [3]. However, the Australian framework has some significant inadequacies.

**Principle A (Human Rights):**

This principle addresses how AI systems might impact individual rights under Australian human rights law, with recognition of international obligations including the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR). The framework emphasises AI discrimination risks based on age, disability, race, and sex, citing documented cases in recruitment, criminal justice, and biometric systems. However, the principle faces structural challenges. Unlike jurisdictions with centralised human rights charters, Australia's fragmented legislation creates uncertainty when rights conflict. The framework also overlooks emerging algorithmic rights such as automated decision-making protections increasingly recognised internationally.

**Principle B (Health and Safety):**

This principle covers risks to physical and mental health from AI systems. It appropriately recognises both traditional safety concerns and contemporary mental health issues. The framework emphasises healthcare applications, showing risks from biased medical AI such as screening tools trained on non-representative data. For instance, pulse oximeters have been found to overestimate blood oxygen levels in patients with darker skin, leading to undertreatment. However, the principle shows limitations in addressing population-level health effects and long-term risks. The narrow focus on individual medical harms overlooks cumulative psychological impacts and systemic health risks that may emerge as AI becomes more embedded in daily life.

**Principle C (Legal Effects):**

From this principle, it is shown that AI systems that produce adverse legal effects, defamation, or similarly significant impacts on individuals. The framework emphasises protecting legal rights where people cannot avoid using the system, particularly regarding access to essential services such as law enforcement, housing, and finance. The principle reflects approaches in other jurisdictions like the EU's General Data Protection Regulation (GDPR), which restricts automated decisions with legal effects. It covers situations like accessing health services and job recruitment done entirely by AI. Nevertheless, the principle overlooks emerging legal concerns such as AI-generated evidence in court proceedings or cumulative effects when multiple AI systems affect the same individual.

**Principle D (Impacts on Groups):**

This principle addresses adverse impacts AI systems may have on groups or collective cultural rights, which represents a progressive recognition of collective harm assessment. Yet the principle remains underdeveloped, lacking specific guidance for measuring collective harms. More concerning is its failure to address Indigenous data sovereignty, a notable omission given Australia's substantial Indigenous population and the unique relationship between Indigenous peoples and data concerning their communities, lands, as well as cultural heritage [9].

**Principle E (Systemic Effects):**

This principle highlights systemic risks AI systems pose to the Australian economy, society, environment, and rule of law. It appropriately recognises threats to democratic processes, including AI-generated disinformation, deepfakes, and manipulation of public opinion that undermine electoral integrity and social cohesion. Environmental impacts from poorly designed automated systems also receive attention. Yet the principle's broad scope creates challenges. Without specific metrics for measuring impact severity, it could capture many AI applications that pose minimal systemic risk. Cumulative effects from multiple AI systems operating simultaneously across society remain inadequately addressed.

**Principle F (Severity and Extent of Impact):**

This principle serves as a meta-principle for assessing the severity and extent of impacts from AI systems identified under principles A through E. The framework requires consideration of who experiences impacts, the scale and intensity of harms, the likelihood of adverse impacts occurring, and the effectiveness of mitigation measures. While this assessment structure proves useful, it operates more as general risk methodology than AI-specific guidance. The principle lacks concrete thresholds for determining when an AI system crosses into high-risk classification, creating uncertainty about practical application.

## International Comparison and Australia’s Framework Adequacy

### International Comparison Analysis

Comparing Australia's framework with international approaches reveals several important gaps. Research shows that AI ethics frameworks around the world generally share common principles such as transparency, accountability, and fairness [1]. However, different countries implement these principles in quite different ways. Some jurisdictions create specific rules that explicitly prohibit the certain uses of AI, while Australia has chosen a broader, more flexible principles-based approach. While this flexibility offers adaptability, it also means Australia's framework lacks the strict implementation guidance and specific requirements found in other countries. This gap becomes particularly clear when the comparison is made.

Malaysia's National Guidelines show how cultural values can be integrated into AI governance [7]. The Malaysian framework includes seven principles which are fairness, reliability and safety, privacy and security, inclusiveness, transparency, accountability, and the pursuit of human benefits and happiness. Malaysia's framework makes transparency a core requirement, asking AI companies to openly show how they handle personal data. The framework also emphasises inclusiveness by considering marginalised groups like rural farmers, grounding its approach in the moral responsibility to promote human well-being.

The European Union (EU) and Canada take different approaches to implementing AI governance [4]. While Australia's principles offer flexibility, this creates uncertainty about how to actually apply them in practice. Without clear procedural steps and accountability measures, organisations may find it hard to know what counts as proper compliance.

Academic research also shows important gaps in Australia's approach. Studies by Mittelstadt show that principles alone don't guarantee ethical AI without proper enforcement and accountability [2]. Research on AI's environmental impact shows that training large models produces huge carbon emissions [5], yet Australia's framework largely ignores sustainability. Additionally, according to Kukutai and Taylor, Indigenous communities need explicit recognition of their collective data rights and self-determination [9]. Their work on Indigenous Data Governance offers frameworks Australia could use, but the current principles miss these opportunities despite Australia's significant Indigenous population.

### Framework Inadequacy Assessment

After comparing with international framework and literature review, Australia's proposed principles show inadequacies across three key dimensions that collectively undermine their effectiveness for identifying high-risk AI systems.

#### Incomplete Risk Coverage:

The framework fails to consider environmental impacts, which is problematic because AI systems require enormous amounts of energy. As highlighted in the literature review, training large language models produces substantial carbon emissions that contradict Australia's climate goals [5]. Similarly, the framework lacks dedicated criteria for algorithmic transparency and explainability, which is a principle that Malaysia successfully incorporates [7] and that international frameworks widely recognise [1]. This limits accountability in AI decision-making processes. Cross-jurisdictional governance challenges also remain unaddressed, as AI systems frequently operate across national boundaries while the framework assumes purely domestic deployment.

#### Cultural Inadequacy:

The framework mentions collective rights but fails to properly address Indigenous data sovereignty and protection for minority communities. As Kukutai and Taylor show, Indigenous communities need explicit recognition of collective data rights and self-determination mechanisms [9]. Malaysia's framework shows how cultural values can be substantively integrated into governance structures [7], yet Australia's current principles overlook these opportunities despite the country's significant Indigenous population and international obligations. Research on algorithmic fairness has also demonstrated that discrimination can be embedded in seemingly neutral technical processes [8], but the framework provides insufficient mechanisms for assessing intersectional discrimination.

#### Implementation Limitations:

The framework assumes organisations have the capacity for sophisticated risk assessment, but many businesses simply don't have these resources. As noted in the comparison above, the European Union and Canada provide structured implementation pathways with clear standard operating procedures [4], while Australia's principles offer flexibility without concrete guidance. Studies show that smaller organisations often lack the technical resources needed for thorough evaluation [6]. These risk burdening smaller companies while giving advantages to large technology corporations.

## Specific Framework Enhancement Recommendations

### Three Principles That Must Be Added

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

Australia’s framework is unable to address environmental impacts is particularly concerning given Australia's climate commitments. Research by Strubell shows that training large AI models can generate carbon emissions exceeding the lifetime output of multiple cars [5]. It also shows that the high energy consumption and environmental costs of AI systems create real policy challenges that governance frameworks need to address [5]. These environmental impacts sit awkwardly alongside Australia's international climate obligations and net-zero targets, yet the current principles completely ignore sustainability.

International frameworks are increasingly recognising environmental sustainability as a governance issue. According to Jobin, it is found that global AI ethics guidelines are starting to incorporate environmental considerations [1], showing that other countries have successfully integrated environmental assessment into their AI oversight. This emphasises that adding environmental criteria is both feasible and increasingly expected in modern AI governance frameworks.

**Proposed Principle:**

"*Risk of adverse environmental impacts from energy consumption, carbon emissions, resource depletion, and electronic waste across the AI system lifecycle*.”

This added principle would identify AI systems as high-risk when their environmental footprint reaches significant levels that could undermine Australia's climate commitments. It would capture systems that consume substantial energy during training or deployment, generate large carbon emissions from data center operations, deplete natural resources through hardware manufacturing, or create environmental harm through electronic waste. The principle would require organisations to assess environmental impacts across the entire AI lifecycle from initial model development and training through ongoing deployment and eventual hardware disposal. This creates accountability for AI developers to consider sustainability alongside other risk factors, ensuring that environmental costs don't get overlooked in the rush to deploy powerful AI systems.

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

Algorithmic transparency is also the aspect the Australia's framework lacks. AI systems now make critical decisions in areas like criminal justice, welfare eligibility, and medical diagnosis, yet they often work like black boxes. When people can't understand how AI reaches its conclusions, they have no way to question or contest those decisions. Transparency matters because it gives people the power to push back against unfair AI decisions. As Mittelstadt points out, good principles for AI on paper mean nothing without practical ways to verify and enforce them [2].

As the evidence, the COMPAS system used in criminal justice to predict whether someone might reoffend. The system showed racial bias, but this only came to light after investigative journalists dug into it [8]. By then, COMPAS had already influenced countless bail and sentencing decisions. Other countries have learned from these problems and built transparency requirements into their frameworks. The EU AI Act requires that high-risk systems be "sufficiently transparent to enable users to interpret the system's output and use it appropriately" and demands comprehensive technical documentation. Canada's Algorithmic Impact Assessment includes explanation requirements that scale with how serious the impact is. Malaysia's National Guidelines make transparency a core requirement [7], asking AI companies to openly show how they handle data and make decisions. Even the OECD AI Principles, which 42 countries have endorsed, explicitly include transparency and explainability [1].

**Proposed Principle:**

"*Risk of decisions or assessments that significantly affect individuals or communities being made through algorithmic processes that lack adequate transparency, explainability, or mechanisms for meaningful human review and challenge."*

This proposed principle would flag AI systems as high-risk when they make important decisions about people's lives without being transparent about how they work. It covers systems used in criminal justice, welfare decisions, medical diagnosis, employment, and credit assessments. These are all situations where people need to understand why the AI reached a particular conclusion and have a way to challenge it if they think it's wrong. Under this principle, organisations would need to provide explanations that ordinary people can actually understand, keep detailed technical records that regulators can review, and set up clear processes for human oversight and appeals. This stops AI from operating as a "black box" in high-stakes situations, giving people the right to understand and contest decisions that significantly affect their lives.

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

AI systems often operate across national borders, but governance frameworks usually stay within their own countries. This mismatch creates gaps that let companies avoid regulations and raises serious concerns about Indigenous data sovereignty. When AI systems can simply move operations to countries with weaker rules, it undermines everyone's protections.

The problem becomes especially acute for Indigenous communities. Indigenous peoples' data gets collected, stored, and processed across borders without proper consent or governance [9]. The UN Declaration on the Rights of Indigenous Peoples recognises that Indigenous peoples have rights to maintain and control their cultural heritage and traditional knowledge, yet AI systems routinely access and process this information without appropriate consent or benefit-sharing arrangements. This perpetuates colonial patterns of taking Indigenous knowledge without permission.

Other countries have shown how to address these issues. The EU's AI Act includes extraterritorial provisions, meaning EU standards apply to AI systems affecting EU residents no matter where the system was developed or hosted. This prevents companies from dodging regulations by moving operations overseas. New Zealand's approach to Māori data sovereignty through Te Mana Raraunga provides a proven framework that respects Indigenous rights. The CARE Principles for Indigenous Data Governance offer internationally recognised standards developed by and for Indigenous peoples [9]. Canada's OCAP principles demonstrate another workable approach. These examples show Australia can adopt similar protections.

**Proposed Principle:**

"*Risk of governance gaps, jurisdictional conflicts, or violations of data sovereignty principles arising from cross-border AI operations, particularly affecting Indigenous peoples' rights to data self-determination and collective cultural protection.*"

Addition of this principle would identify AI systems as high-risk when they operate across borders in ways that create governance gaps or violate Indigenous data sovereignty. It would capture systems that exploit differences between countries' regulations, process Indigenous data without proper consent from Indigenous communities, or operate in ways that undermine Indigenous peoples' right to control information about their communities, lands, and cultural heritage. The principle would require organisations to respect Indigenous data sovereignty regardless of where systems operate, establish clear governance for cross-border AI deployments, and ensure Indigenous communities have authority over their own data. This prevents regulatory arbitrage while protecting Indigenous rights to self-determination and cultural protection in the AI era.

# Conclusion

Australia's proposed six principles for identifying high-risk AI systems demonstrate inadequacies across multiple dimensions that collectively undermine their effectiveness for contemporary AI governance challenges. The framework exhibits four primary deficiencies which are incomplete risk coverage omitting environmental impacts, algorithmic transparency requirements, and cross-jurisdictional governance challenges; implementation limitations assuming unrealistic organisational capacity; cultural inadequacy in addressing Indigenous data sovereignty; and dynamic assessment gaps failing to account for evolving AI capabilities. These deficiencies create regulatory gaps potentially exposing Australians to emerging AI risks while disadvantaging smaller enterprises unable to navigate complex compliance requirements.

Addressing the fundamental research question, the proposed principles do not adequately capture high-risk AI systems. 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. These enhancements would transform Australia's approach from reactive harm mitigation to proactive risk prevention, positioning the nation as a leader in responsible AI governance while providing robust protection for all Australians, particularly vulnerable communities whose needs the current framework inadequately addresses.

# 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. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, June 2019, pp. 1 -6.
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. T. Kukutai and S. Taylor, *Indigenous Data Sovereignty: Toward an Agenda*. Canberra, Australia: ANU Press, 2016.

# Appendix

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

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 computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Utilise AI to make sure my idea aligning with the topic chosen, avoid the out of topic problem.

A screenshot of a cell phone

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Consult with AI to consolidate my understanding

A screenshot of a computer

AI-generated content may be incorrect.

A paper with text on it

AI-generated content may be incorrect.

A screenshot of a questionnaire

AI-generated content may be incorrect.

A screenshot of a document

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Discuss with AI about the planed structure of research report

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Use AI to fasten the research paper searching process

A screenshot of a computer

AI-generated content may be incorrect.

Write the report with help of Grammarly AI

Link for the AI Chat: <https://chatgpt.com/share/68e64c02-6a80-8003-82c0-24c6da12e674>