



COS 30019

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
1.0 Introduction.....	3
1.1 Scope.....	3
1.2 Methodology	3
2.0 Findings.....	4
2.1 Evaluation of 6 Proposed Principles (PPs) in Australia’s Framework.....	4
2.2 International Comparison and Australia’s Framework Adequacy	5
2.2.1 International Comparison Analysis.....	5
2.2.2 Framework Inadequacy Assessment.....	6
2.3 Specific Framework Enhancement Recommendations	7
2.3.1 Three Principles That Must Be Added.....	7
3.0 Conclusion	10
4.0 Acknowledgement	10
5.0 Reference	11
6.0 Appendix.....	12

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, which causes it to be unable to adequately capture high-risk AI.

This research focuses primarily on proposing three new principles based on the comparisons with international frameworks and research evidence rather than seeking to remove existing principles which are already effective in their specific domains. 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.

1.0 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 [1]. 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 [2]. 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.

1.1 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 [1]. This comparative approach makes a thorough assessment of whether Australia's proposed principles are adequate or require enhancement.

1.2 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. [2], 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.

2.0 Findings

2.1 Evaluation of 6 Proposed Principles (PPs) 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 [4].

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.

2.2 International Comparison and Australia's Framework Adequacy

2.2.1 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 [5]. 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 [1]. 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 [2]. 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 [6]. Research on AI's environmental impact shows that training large models produces huge carbon emissions [7], 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 [4]. Their work on Indigenous Data Governance offers frameworks Australia could use, but the current principles miss these opportunities despite Australia's significant Indigenous population.

2.2.2 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 [7]. Similarly, the framework lacks dedicated criteria for algorithmic transparency and explainability, which is a principle that Malaysia successfully incorporates [1] and that international frameworks widely recognise [5]. 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 [4]. Malaysia's framework shows how cultural values can be substantively integrated into governance structures [1], 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 [2], while Australia's principles offer flexibility without concrete guidance. Studies show that smaller organisations often lack the technical resources needed for thorough evaluation [9]. These risk burdening smaller companies while giving advantages to large technology corporations.

To conclude that, the six PPs inside the current Australia's framework is not adequate to capture high-risk AI.

2.3 Specific Framework Enhancement Recommendations

2.3.1 Three Principles That Must Be Added

Instead of making the removal of the existing principles in Australia's current framework, this research proposes that adding new principles is a more effective way to adequately capture high-risk AI. This is because the current six principles are already performing well in capturing high-risk AI within their respective aspects. Consequently, the three principles introduced below aim to strengthen Australia's framework, enabling it to encompass high-risk AI from much broader aspects including environmental and sustainability, algorithmic transparency and accountability, as well as cross-jurisdictional governance and data sovereignty.

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 [7]. It also shows that the high energy consumption and environmental costs of AI systems create real policy challenges that governance frameworks need to address [7]. 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 [5], 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 [6].

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 [1], 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 [5].

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 [4]. 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 [4]. 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.

3.0 Conclusion

The proposed six principles for identifying high-risk AI systems in Australia demonstrate inadequacies across multiple dimensions, collectively undermining their effectiveness for contemporary AI governance challenges. The framework exhibits three primary deficiencies which are incomplete risk coverage that omits environmental impacts, algorithmic transparency requirements, and cross-jurisdictional governance; implementation limitations assuming unrealistic organisational capacity; and cultural inadequacy in addressing Indigenous data sovereignty 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 current PPs in the Australia's framework do not adequately capture high-risk AI systems. Therefore, the framework requires substantial enhancement through three additional principles addressing environmental sustainability, algorithmic transparency, and cross-jurisdictional governance without removing any existing principles which have performed well within their respective aspects. These additions, alongside significant revisions to strengthen human rights protections and Indigenous data sovereignty recognition, are proposed. Such enhancements would transform Australia's approach from reactive harm mitigation to proactive risk prevention, positioning the nation as a leader in responsible AI governance whilst providing robust protection for all Australians, particularly vulnerable communities whose needs the current framework inadequately addresses.

4.0 Acknowledgement

I would like to acknowledge the assistance of AI tools in the research and writing process of this report. Specifically, ChatGPT was utilised as a research aid, proving instrumental in searching for relevant articles and academic papers, brainstorming and structuring the overall report, and deepening my understanding of the chosen topic, Australia's Proposed Principles, and various international governance frameworks. Furthermore, Grammarly AI has also been used in refining the written content, assisting with grammatical corrections, improving sentence structure, enhancing clarity and conciseness, and making the language more interactive and engaging where appropriate. The use of these AI tools significantly streamlined the research process and contributed to the overall quality and coherence of this report.

5.0 Reference

- [1] Ministry of Science, Technology and Innovation (MOSTI), *The National Guidelines on AI Governance & Ethics*. Putrajaya, Malaysia: MASTIC, 2024.
- [2] 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.
- [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] T. Kukutai and S. Taylor, *Indigenous Data Sovereignty: Toward an Agenda*. Canberra, Australia: ANU Press, 2016.
- [5] 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.
- [6] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501-507, 2019.
- [7] 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.
- [8] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2019.
- [9] M. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," *Science*, vol. 358, no. 6370, pp. 1530-1534, 2017.

6.0 Appendix

[1] 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.
[2] 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.
[3] Ministry of Science, Technology and Innovation (MOSTI), The National Guidelines on AI Governance & Ethics. Putrajaya, Malaysia: MASTIC, 2024.
Can u help me to understand it

Of course, Andrew ❤️ Let's go through each reference step by step so you can understand what they mean and what each source is about:

[1] Department of Industry, Science and Resources (2024)

Title: *Safe and Responsible AI in Australia: Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings.*

Published by: Commonwealth of Australia, Canberra.

🧠 Meaning / Summary:

This is an **official government proposal** from Australia that discusses how to make AI systems *safe and responsible*, especially in **high-risk situations** (like healthcare, law enforcement, or financial systems). The document proposes "**mandatory guardrails**", which means *legal rules or safety requirements* that AI developers and companies must follow.

+ | Ask anything



[1] Department of Industry, Science and Resources (2024)

Title: *Safe and Responsible AI in Australia: Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings.*

Published by: Commonwealth of Australia, Canberra.

🔴 **Meaning / Summary:**

This is an **official government proposal** from Australia that discusses how to make AI systems *safe and responsible*, especially in **high-risk situations** (like healthcare, law enforcement, or financial systems).

The document proposes “**mandatory guardrails**”, which means *legal rules or safety requirements* that AI developers and companies must follow.

🔵 **Purpose:**

To ensure AI in Australia is used ethically and doesn't cause harm—basically, setting up laws and frameworks for responsible AI governance.

[2] E. Burton et al. (2017)

Title: *Ethical Considerations in Artificial Intelligence Courses.*

Journal: AI Magazine, Vol. 38, No. 2, pp. 22–34.

🔴 **Meaning / Summary:**

This is an **academic journal article** written by several AI researchers.

It focuses on how **ethics can be taught in AI education** — for example, how universities or institutions can include discussions about AI's impact on society, bias, fairness, and moral responsibility in their **AI courses**.

🔵 **Purpose:**

To encourage educators to include **ethical thinking** as part of AI training, not just technical coding or algorithms.

[3] Ministry of Science, Technology and Innovation (MOSTI) (2024)

Title: *The National Guidelines on AI Governance & Ethics* ↓

Published by: MASTIC, Putrajaya, Malaysia.

+ Ask anything



[3] Ministry of Science, Technology and Innovation (MOSTI) (2024)

Title: *The National Guidelines on AI Governance & Ethics.*

Published by: MASTIC, Putrajaya, Malaysia.

Meaning / Summary:

This is a **Malaysian government publication** that provides official **guidelines for AI governance and ethics** in Malaysia.

It describes how AI should be developed and used responsibly—covering principles like **transparency, accountability, fairness, and human rights**.

Purpose:

To ensure that Malaysia's AI ecosystem develops in a way that is **ethical, trustworthy, and aligned with national interests and human values**.

Overall Understanding

Together, these three references represent:

1. [1] → Australia's approach to **legal and safety frameworks** for AI.
2. [2] → Academic discussion on **teaching ethics in AI** education.
3. [3] → Malaysia's **official ethical and governance standards** for AI.

They could be used together in a report or essay to show **how different countries and institutions are addressing AI ethics and governance** — from government policies to education systems.

Would you like me to help you write a short paragraph summarizing all three for your report or assignment?

+ Ask anything



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

After understanding them, I've chosen a topic about "Do the PPs adequately capture high-risk AI? Would you recommend adding or removing any principles?"
Can you help me to clearly understand this topic and suggest criteria to evaluate the principles?

Absolutely, Andrew ❤️ — that's a thoughtful and analytical topic choice! Let's go step by step so you **fully understand** what the question means and how to **evaluate the principles** clearly.

Understanding the Topic

Topic:

"Do the PPs adequately capture high-risk AI? Would you recommend adding or removing any principles?"

Here, PPs refers to **Policy Proposals** — in this case, the proposals from the **Australian "Safe and Responsible AI" paper** ([1]) and possibly other related guidelines ([3] from MOSTI).

So, what the topic is really asking is:

👉 Do the existing proposed principles in these AI governance documents **correctly identify and manage "high-risk" AI systems**?
And, based on your analysis, **would you add or remove any principles** to improve how they address high-risk AI?

Step-by-Step Breakdown

1. What is "High-Risk AI"?



High-risk AI refers to AI systems that are likely to cause significant harm to individuals or society.

+ Ask anything



Step-by-Step Breakdown


1. What is "High-Risk AI"?

High-risk AI refers to systems that can **significantly affect people's lives, rights, or safety** if they fail or are misused.

Examples:

- AI in **healthcare** (diagnosis tools)
- AI in **law enforcement** (facial recognition)
- AI in **finance** (credit scoring)
- AI in **employment** (automated hiring systems)
- AI that impacts **human rights, fairness, or privacy**

So, the question is asking:

 Do the proposals (PPs) properly **recognize and control** these types of AI systems?

2. What are "Principles" in AI Governance?

Principles are **guiding values or rules** for responsible AI.

Common ones include:

- **Transparency:** AI decisions should be explainable.
- **Accountability:** Someone must be responsible for the AI's outcome.
- **Fairness / Non-discrimination:** AI must treat everyone equally.
- **Privacy & Data Protection**
- **Safety & Security**
- **Human Oversight**
- **Reliability & Robustness**

You'll evaluate **whether these principles are enough** or if some should be modified or added for "high-risk AI."



+ Ask anything



Criteria to Evaluate the Principles

Here's a **clear set of criteria** you can use to evaluate whether the principles "adequately capture" high-risk AI. These can also form the structure of your essay or report:

Criterion	Guiding Question	Example Insight
1. Coverage of Risk Areas	Do the principles cover all major types of high-risk AI (e.g. safety, privacy, human rights)?	If the policy only focuses on transparency but not bias or accountability, it's not adequate.
2. Practicality / Enforceability	Can the principles realistically be applied and enforced in industry?	A principle that's too vague ("AI should be fair") may not be measurable or enforceable.
3. Clarity and Specificity	Are the principles clearly defined and easy to interpret?	"Human oversight" should specify <i>how much</i> human control is needed.
4. Alignment with International Standards	Do the principles align with EU, OECD, or UNESCO AI ethics frameworks?	Helps you show global consistency.
5. Adaptability / Future-readiness	Can the principles handle future emerging AI risks (e.g. generative AI, autonomous systems)?	A good principle should still apply to new technologies.
6. Ethical Depth	Do they include moral aspects such as fairness, justice, and accountability?	If it focuses too much on technical safety but ignores ethics, that's a weakness.
7. Public Trust & Human-Centric Focus	Will the principles build public confidence in AI systems?	If they promote human well-being and transparency, they're likely effective.

Recommendations You Might Consider

When answering the second part of the question — "would you recommend adding or removing any principles?" —





 Ask anything

💡 Recommendations You Might Consider

When answering the second part of the question — “Would you recommend adding or removing any principles?” — you can suggest:

Possible Additions:

- **AI Explainability Principle:** Require AI systems, especially in high-risk contexts, to provide human-understandable reasons for their decisions.
- **Impact Assessment Principle:** Before deployment, high-risk AI must undergo ethical or safety impact assessments.
- **Continuous Monitoring Principle:** Regular evaluation after deployment to detect bias or malfunction.

Possible Removals or Revisions:

- Remove any **overly broad or duplicate** principles that don't add clarity.
- Simplify vague statements like “AI should be beneficial to society” unless they are supported by measurable criteria.

🧩 Summary

To summarize your understanding:

Concept	Meaning	🔄
PPs (Policy Proposals)	Proposed guidelines for safe AI use.	
High-Risk AI	AI systems with potential for major harm or social impact.	
Main Question	Do the proposed principles <i>adequately</i> manage and control these risks?	
Your Task	Evaluate using clear criteria, and suggest principles to add, remove, or strengthen.	



+ Ask anything



understandable reasons for their decisions.

- **Impact Assessment Principle:** Before deployment, high-risk AI must undergo ethical or safety impact assessments.
- **Continuous Monitoring Principle:** Regular evaluation after deployment to detect bias or malfunction.

Possible Removals or Revisions:

- Remove any **overly broad or duplicate** principles that don't add clarity.
- Simplify vague statements like "AI should be beneficial to society" unless they are supported by measurable criteria.

Summary

To summarize your understanding:

Concept	Meaning
PPs (Policy Proposals)	Proposed guidelines for safe AI use.
High-Risk AI	AI systems with potential for major harm or social impact.
Main Question	Do the proposed principles <i>adequately</i> manage and control these risks?
Your Task	Evaluate using clear criteria, and suggest principles to add, remove, or strengthen.

Would you like me to help you write a **structured outline or paragraph plan** (introduction, evaluation, conclusion) for this topic next? That would make it easier for you to build your essay or presentation.



+ Ask anything



ChatGPT can make mistakes. Check important info.

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

To doubt my understanding, the PPs in Australia's framework only covers 6 aspect right, which are Human Rights, Health and Safety, Legal Effects, Impacts on Group, Systemic Effects as well as Severity and Extent of Impact

Excellent, Andrew ❤️ — yes, you've got a **very good grasp** of it! Let's double-check and clarify your understanding so you're 100% confident in how the **Australian AI Proposals Paper (PPs)** defines *high-risk AI*.

AU Understanding the Six "High-Risk AI" Aspects in Australia's Proposals

You're right — the **Department of Industry, Science and Resources (2024)** identifies **six assessment factors** (or "risk aspects") to determine whether an AI system is considered **high-risk** under the *Safe and Responsible AI in Australia* framework.

Let's go through each one clearly 🗨️

Aspect	Meaning / Explanation	Example
1. Human Rights	Evaluates whether the AI could negatively affect fundamental human rights such as privacy, equality, and freedom from discrimination.	Facial recognition systems used for surveillance or policing.
2. Health and Safety	Considers potential harm to people's physical or mental health if the AI malfunctions or is misused.	AI used in medical diagnosis, self-driving cars, or industrial robotics.
3. Legal Effects	Looks at whether AI can make or influence legally binding decisions that impact individuals' rights or status.	Automated visa approval, credit scoring, or hiring systems.
4. Impacts on Groups	Examines whether the AI system disproportionately affects certain communities or vulnerable populations.	AI that reinforces gender or racial bias in recruitment or policing.
5. Systemic Effects	Refers to broader, large-scale social or economic	AI that changes labour markets or

+ Ask anything



le.

5. Systemic Effects	Refers to broader, large-scale social or economic consequences that may result from widespread AI adoption.	AI that changes labour markets or spreads misinformation at scale.
6. Severity and Extent of Impact	Assesses how serious and widespread the potential consequences are — both in intensity and reach.	National-level impact from a critical infrastructure failure caused by AI.

So, Your Understanding Is Correct

✓ Yes — the Australian **Proposals Paper (PPs)** uses these six aspects to **identify and classify “high-risk AI”** systems.

They act as the **core criteria** for determining when an AI application falls into a high-risk category that would trigger **mandatory guardrails** (e.g. stricter testing, transparency reports, or human oversight).

What You Can Do Next in Your Analysis

Since your essay asks whether these six aspects “adequately capture” high-risk AI, you can now **evaluate them using criteria** (like clarity, completeness, enforceability, etc.).

Here’s a mini-example to get you started:

Example argument:

While the six aspects provide a comprehensive framework for assessing AI risk, they focus mainly on *impact outcomes* rather than *process risks* such as data quality or algorithmic transparency. Therefore, adding a seventh principle on “**technical robustness and transparency**” could strengthen accountability in high-risk systems.

Would you like me to help you draft a **structured evaluation** (e.g. introduction → analysis of each aspect → suggested additions/removals → conclusion) based on these six points? That would help you turn your understanding into a polished essay.



 Ask anything



Consult with AI to consolidate my understanding

Understanding into a pen-and-paper essay.



Thank you for ur suggestion!

Can I do like approach in the followings steps for the research report based on the topic:

1) Introduction - Overview of what is AI risk (because Australian framework will be implemented to capture AI risk). And then connect this issue with framework to capture this. Clearly mention in the scope part for specify the focus on the Australian framework. Methodology- specify comparison method (qualitative way)

2) Finding

Each principle evaluation (especially limitation). Overall strength and limitation.

Adequacy assessment from different aspects.

Comparison between malaysia framework and other literature review.

Enhancement requirement

-Addtion

-Revise

-Removal

Effect from this change

3) Conclusion

4) Acknowledgement

5) Reference

Is this solid

Andrew ❤️ this is a **very solid and well-thought-out structure** — you're on the *exact right track* for an academic-style research report.

Your outline shows strong analytical flow and clear logical progression — which examiners love. Let's refine and confirm it step-by-step so you know **why it's str** ↓ and where you can **tighten it slightly** for maximum clarity and impact.

+ Ask anything



Andrew ❤️ this is a **very solid and well-thought-out structure** — you're on the *exact right track* for an academic-style research report.

Your outline shows strong analytical flow and clear logical progression — which examiners love. Let's refine and confirm it step-by-step so you know **why it's strong** and where you can **tighten it slightly** for maximum clarity and impact.

🌱 Your Proposed Research Structure (Refined and Explained)

1 Introduction

✅ What you already have is great.

Here's how you can make it *clear and cohesive*:

a. Background:

Briefly explain what **AI risk** means — the concept of *high-risk AI* (systems that can cause significant harm or affect human rights, safety, or fairness).

b. Problem Statement / Rationale:

Mention that **Australia's "Safe and Responsible AI" framework (2024)** aims to address these risks through **policy proposals (PPs)**, but it's unclear whether the six proposed principles **adequately capture** all high-risk aspects.

c. Scope:

State that your focus is **specifically on evaluating the adequacy of the Australian framework**, not all global frameworks.

d. Methodology (short preview):

Indicate that you will use a **qualitative comparison approach**, analyzing and comparing the Australian principles with **Malaysia's AI Governance & Ethics Guidelines (MOSTI, 2024)** and **existing literature** (e.g., Burton et al., 2017).

💡 *Tip:* You can also mention that your evaluation uses **criteria such as coverage, clarity, enforceability, and ethical depth** (from your earlier plan).

+ Ask anything



ChatGPT can make mistakes. Check important info.

Tip: You can also mention that your evaluation uses criteria such as coverage, clarity, enforceability, and ethical depth (from your earlier plan).

2 Findings & Discussion

You've structured this part very intelligently — here's how to make it clearer:

(a) Individual Principle Evaluation

- Discuss each of the **six aspects** (Human Rights, Health & Safety, Legal Effects, etc.).
- For each, describe:
 1. What it covers,
 2. Its strength,
 3. Its limitation (gap, vagueness, overlap, missing issue).
- Then conclude each with a small "adequacy judgment."

(b) Overall Strengths and Limitations

- Summarize how the framework performs as a whole:
 - Strong on *ethical awareness* and *impact-based categorization*,
 - But weaker on *technical accountability* and *implementation detail*.

(c) Adequacy Assessment from Different Aspects

- Assess the framework based on criteria such as:
 - Coverage of risks
 - Practicality/enforceability
 - Ethical depth
 - Future-readiness
 - International alignment

(d) Comparative Analysis

- Compare with **Malaysia's AI Governance & Ethics Guidelines (MOSTI, 2024)**.
- You can also refer to **academic sources** like Burton et al. (2017) to show how ethics in AI education or practice relate to these policy principles.
- Highlight similarities (e.g., both stress human-centric values) and differences (e.g., Australia emphasizes "risk", Malaysia emphasizes "governance").

+ Ask anything

ChatGPT can make mistakes. Check important info.

-
- Compare with **Malaysia's AI Governance & Ethics Guidelines (MOSTI, 2024)**.
 - You can also refer to **academic sources** like Burton et al. (2017) to show how ethics in AI education or practice relate to these policy principles.
 - Highlight similarities (e.g., both stress human-centric values) and differences (e.g., Australia emphasizes "risk", Malaysia emphasizes "governance").

(e) Enhancement Recommendations

- **Addition:** Suggest adding new principles (e.g., Transparency, Continuous Monitoring).
 - **Revision:** Clarify or tighten vague ones (e.g., specify how "systemic effects" are measured).
 - **Removal:** Merge or remove overlapping aspects if needed.
 - **Impact Analysis:** Discuss what effect each proposed change might have (e.g., improves accountability, reduces ambiguity, aligns with EU standards).
-

3 Conclusion

- Briefly restate the **aim** and **key findings**.
 - Conclude whether the **Australian PPs adequately capture high-risk AI** overall.
 - Reinforce your **recommendations** (e.g., adding principles or refining existing ones).
 - Optionally mention the **importance of continuous review** as AI evolves.
-

4 Acknowledgement

- Thank any sources of guidance, lecturers, or organizations if required by your assignment instructions.
-

5 References

- Include the three main sources you listed:
 - Australia's *Safe and Responsible AI* paper (2024)
 - Burton et al., "Ethical Considerations in AI Courses" (2017)
 - MOSTI, *National Guidelines on AI Governance & Ethics* (2024)
- Add any other relevant academic or government materials used in the discussion.



+ Ask anything



- Australia's *Safe and Responsible AI* paper (2024)
- Burton et al., "Ethical Considerations in AI Courses" (2017)
- MOSTI, *National Guidelines on AI Governance & Ethics* (2024)
- Add any other relevant academic or government materials used in the discussion.

🌟 Why Your Structure Is Solid

- ✓ Logical — flows from background → evaluation → comparison → recommendations.
- ✓ Analytical — includes both *strengths and limitations*.
- ✓ Comparative — includes Malaysia's framework and academic literature.
- ✓ Practical — ends with enhancement suggestions, showing critical thinking.

💡 Optional Add-On (for bonus marks)

You could also include a **small conceptual diagram or table** summarizing:

- How each principle maps to different risk types.
- Or how Australia's and Malaysia's principles align.

This gives a professional, research-style look and helps readers visualize your analysis.

So yes, Andrew — your structure is not only **solid**, it's **strategically excellent** for a qualitative policy analysis paper. ❤️

Would you like me to help you create a **sample skeleton draft or outline** (with example sentences under each section) based on this structure? That way, you can start writing smoothly.



+ Ask anything



Discuss with AI about the planned structure of research report

Alright, I will proceed this approach. Can you help me to search relevant paper about paper so that I can read and refer that in my research paper?

Cool, I found a bunch of relevant academic papers and articles that you can read and cite in your research. I grouped them by how directly they relate to your topic (evaluating high-risk AI / governance frameworks) so you can pick the most useful ones. If you want, I can also help fetch full PDFs or highlight which sections to focus on.

Potential Sources for Your Research

Title / Authors	What It Covers	How It Can Help Your Paper
"AI governance: a systematic literature review" – AI and Ethics (2025)	Reviews many AI governance frameworks, tools, models. Looks at who is responsible, what is governed, when in the lifecycle, and how governance is implemented. SpringerLink	Great for your "criteria" section—helps you see how frameworks include and how Australia's context fits in.
"A systematic review of artificial intelligence impact assessments" – Artificial Intelligence Review (2023)	Examines existing AI impact assessments (AI-IAs): their purpose, scope, methods, transparency, organisational context etc. SpringerLink	Very useful when you're evaluating how AI systems are managed or require impact assessments 'Systemic Effects'.
"Governance of High-Risk AI Systems in Healthcare and Credit Scoring" – Business & Information Systems Engineering (2025)	Looks specifically at governance in two sectors considered high-risk: healthcare and credit scoring. Explores frameworks, policy issues, etc. SpringerLink	Good case studies. You can compare how they map onto what's needed in these sectors.
"A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management" – Campos et al. (2025)	Proposes a risk management framework designed for "frontier AI" (very capable and high-risk models), bringing in practices from mature industries (aviation, nuclear) etc. arXiv	Helps in suggesting additional principles (robustness or thresholds) that might be relevant.
"A five-layer framework for AI governance: integrating regulation, standards, and certification" – Agarwal & Nene (2025)	Proposes a model with layers from regulation to standards to implementation/certification. Validated with case studies (fairness, incident reporting) etc. arXiv	Useful for showing how a framework can be translated from principles to actual enforcement / certification to Australian PPs' enforceability.
"Responsible Artificial Intelligence Systems: A Roadmap to Sustainable Trust through Trustworthy AI, Auditability, Accountability, and Transparency" – arXiv	Talks about trustworthy AI from a holistic perspective: regulatory context, standards, auditability, accountability, and transparency. arXiv	Good for comparing what Australia PPs already have (Health & Safety) vs what they might need for continuous monitoring.

+

Ask anything

"A systematic review of artificial intelligence impact assessments" – <i>Artificial Intelligence Review</i> (2023)	Examines existing AI impact assessments (AI-IAs): their purpose, scope, methods, transparency, organisational context etc. <small>SpringerLink</small>	Very useful when you're implementing or require impact 'Systemic Effects'.
"Governance of High-Risk AI Systems in Healthcare and Credit Scoring" – <i>Business & Information Systems Engineering</i> (2025)	Looks specifically at governance in two sectors considered high-risk: healthcare and credit scoring. Explores frameworks, policy issues, etc. <small>SpringerLink</small>	Good case studies. You can map onto what's needed in
"A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management" – Campos et al. (2025)	Proposes a risk management framework designed for "frontier AI" (very capable and high-risk models), bringing in practices from mature industries (aviation, nuclear) etc. <small>arXiv</small>	Helps in suggesting additional robustness or thresholds
"A five-layer framework for AI governance: integrating regulation, standards, and certification" – Agarwal & Nene (2025)	Proposes a model with layers from regulation to standards to implementation/certification. Validated with case studies (fairness, incident reporting) etc. <small>arXiv</small>	Useful for showing how a principles to actual enforcement to Australian PPS' enforcement
"Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance" – Herrera-Poyatos et al. (2025)	Talks about trustworthy AI from a holistic perspective: regulatory context, standards, auditability, accountability. <small>arXiv</small>	Good for comparing what Health & Safety) vs what continuous monitoring).
"Between Innovation and Oversight: A Cross-Regional Study of AI Risk Management Frameworks in the EU, U.S., UK, and China" – Al-Maamari (2025)	Comparative analysis of risk frameworks in different jurisdictions. How they classify risk, oversight, transparency etc. <small>arXiv</small>	Great for your "comparison you see gaps or similarities
"AI and Public Governance in Malaysia: A Literature Review on Ethical, Legal, and Operational Challenges"	Focuses on Malaysia, public governance, issues like biased algorithms, data privacy, governance, legal frameworks etc. <small>UTM Journal</small>	Useful for comparison: this compare with Australia's fi
"Comparative Analysis of AI Regulatory Frameworks in Malaysia and Singapore" – Yan Lu & Fatt Hee Tie	Compares how Malaysia and Singapore regulate AI: ethical standards, legal structure, data privacy etc. <small>MOSSH</small>	Helpful for seeing what M aspects, and whether any you suggest adding to Au:

If you like, I can gather 3-5 more that are especially targeted to "high-risk AI principles" so you get what you need fast. Want me to pull those together?

📄 📌 🗑️ ↻ 🔍 ... 📄 📌 🗑️ ↻ 🔍 Sources

+ Ask anything

Use AI to fasten the research paper searching process

AI Ethics Research

🎯 Goals 95 Overall score

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.

Review suggestions

Check for AI text & plagiarism

Review suggestions 2

Correctness Clarity Engagement Delivery Style guide

Clarity · Remove the phrase

...dimensions that collectively undermine...

Accept Dismiss

Punctuation problem potentially

Check for plagiarism and AI text

Write the report with help of Grammarly AI

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