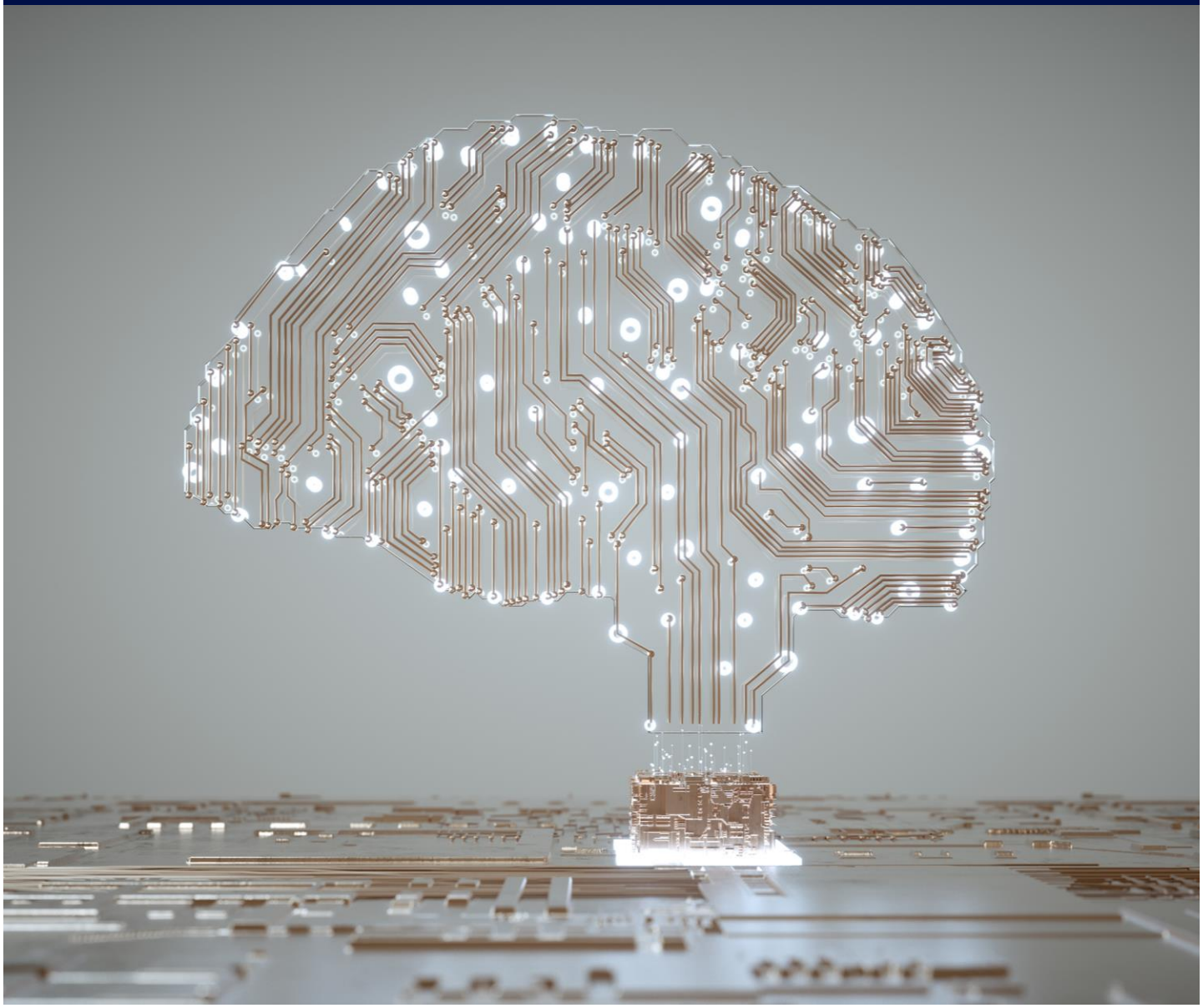
Introducing mandatory guardrails for AI in high-risk settings: proposals paper

Submission to the Department of Industry, Science and Resources

October 2024

THE UNIVERSITY OF MELBOURNE



Executive Summary

The University of Melbourne welcomes the opportunity to respond to the Department of Industry, Science and Resources on the *Introducing mandatory guardrails for AI in high-risk settings proposals paper.*

The University of Melbourne has been deeply engaged with responding to the challenges and risks posed by AI tools and systems, particularly since the emergence of widely available generative AI tools such as ChatGPT in late 2022. We have established a University-wide approach to generative AI, promoted innovation in the use of new tools, and built understanding and awareness of the risks and potential benefits associated with these tools.

The risks that GenAI tools represent for teaching and research institutions are now widely acknowledged, at least in general terms. An important subset of these challenges concerns the possibility of deception linked to GenAI's ability to produce outputs that plausibly resemble that produced by humans and the potential threat this poses for academic and research integrity. Other risks include the tendency of the AI systems used in research to replicate and amplify the biases already embodied in input data, and relatedly the danger of researchers and non-experts overestimating the reliability of AI outputs.

These risks are real and require a response. However, we need to ensure that attempts to address them do not prevent us from realising the opportunities made possible through AI tools and systems. The proposed mandatory guardrails attempt to balance these risks and opportunities but they do not neatly apply across all sectors, particularly in higher education and research. This means that further consultation with individual sectors will be required to develop tailored, sector-specific guidance on the guardrails. This will be necessary regardless of the regulatory approach taken by the Government.

This submission is divided into two parts. The first outlines the University's current initiatives and approaches to generative AI, including the University's ten AI principles, some of which align with the Government's proposed guardrails. The second provides a high-level response to the proposals paper. The University makes recommendations to ensure that principles and guardrails are fit-for-purpose and acknowledge nuances between the development and deployment of AI in different sectors.

The University recommends that the Government:

1. Refine the proposed principles for assessing high-risk AI to improve their clarity and coverage, including impacts on future generations and the environment.
2. Conduct sector-specific consultation and provide tailored guidance on the guardrails to ensure the framework supports innovation across the economy while safeguarding the public. This will be necessary regardless of the regulatory approach taken.
3. Establish clear, distinct principles that differentiate between the development and deployment phases of AI, including in research contexts.
4. Ensure that any regulatory body overseeing AI has appropriate expertise and training in AI.
5. Ensure that any ISO standards referenced in the guardrails are publicly available.
6. Reframe the guardrails requiring organisations to make information publicly available, clarifying that the primary audience for this information will be consumer advocates.
7. Offer clearer guidance and certainty regarding the application of copyright laws to AI.
8. Include an obligation for organisations to ensure AI literacy among staff.
9. Align Australia's regulatory framework with international practices to prevent excessive burdens on businesses, particularly small and medium-sized enterprises.

For further information or to discuss the submission, Professor Gregor Kennedy, Deputy Vice-Chancellor (Academic) can be contacted at gek@unimelb.edu.au.

Current AI initiatives at the University

Generative AI Taskforce (GAIT)

The University's Generative AI Taskforce (GAIT) was established in 2023 to oversee the response to the opportunities and challenges associated with the emergence of GenAI tools. The Taskforce is chaired by the Deputy Vice-Chancellor (Academic), with members drawn from the leadership of key divisions within the University, ensuring a coordinated strategy across three domains: Teaching and learning, Research and research training, and Planning and operations. The University is building a suite of resources relevant to each of these three domains, available to staff and students.

University of Melbourne AI Principles

In early 2024, the University endorsed a set of ten AI principles to articulate its position regarding the implications of AI for university activities, and to help guide actions around the adoption and use of AI tools and systems. These principles include fairness, responsibility and human oversight, and collaboration and ongoing review. The principles were drafted by GAIT in consultation with a range of internal stakeholders, but draw from existing principles frameworks, including *Australia's AI Ethics Principles*, published by DISR in 2019 as well as the Group of Eight's principles on the use of generative artificial intelligence. Significantly, the principles are closely aligned with the guardrails articulated in the proposals paper.

Whole-of-University Risk Review

The University's Risk and Assurance team conducted a risk and opportunity assessment on the use of GenAI across the University, finalised in August 2024. The point-in-time risk assessment was informed by a desktop review of documentation and consultation with stakeholders and subject-matter experts within the University. The report identifies potential risks and opportunities and provides an outline of the current and planned mitigation strategies (in part captured by the University's Teaching and Learning Action Plan, submitted to TEQSA) as the basis of the University's approach to GenAI impacts.

Supporting staff AI literacy and preparedness

The University has developed a suite of professional development programs for staff, aimed at building AI literacy and capability, and at raising awareness and understanding of the risks and opportunities generated by AI tools and systems. These professional development programs include the "Navigating GenAI" workshop, delivered online, which provides staff with a foundational knowledge of using GenAI tools in a university setting, and a hands-on workshop designed to provide staff with the knowledge and tools to leverage generative AI in their teaching practices. The University's Centre for the Study of Higher Education (CSHE) has also run a range of AI-relevant sessions. More than 800 staff members have attended at least one professional development session this year.

The University's GenAI in Teaching Community of Practice was established in 2023, to create a space for academics to share ideas and explore the technology as it applies to university teaching and learning. The group meets monthly and currently has more than 500 members. Additionally, CSHE's Assessment, AI and Academic Integrity website provides practical advice to teaching staff on the use of GenAI tools for assessment and to identify academic misconduct.

Assessment reform

Across 2024, the University has been conducting a review of all assessment types used in our course programs, with a view to reforms that provide assurance of learning outcomes in the context of readily available GenAI tools. The review's focus is on addressing types of assessment that have a higher vulnerability to integrity breaches and will establish an evidence base to support potential reforms. This may

include policy changes supported by the Academic Board's review of the University's Assessment and Results Policy in 2025. It is anticipated that an intensive program of reform will likely start from January 2025.

Research and research training

The University of Melbourne is home to leading research into AI and the implications of new technologies and is committed to supporting innovation through engagement with industry and other partners. The University's Centre for Artificial Intelligence and Digital Ethics facilitates cross-disciplinary research, teaching and leadership on the ethical, technical, regulatory and legal issues relating to AI and digital technologies. Our Faculty of Engineering and Information Technology supports basic research into AI as well as partnering with industry to enhance performance via AI-based interventions.

The University has recently appointed two academic convenors to oversee a collection of activities relevant to AI use in research and research training, including the establishment of a University-wide community of practice for the use of AI in research and the development of a framework for training to support researchers and graduate researchers on responsible use of AI.

In 2024, advice to research staff and to graduate researchers on the use of digital assistance tools has been updated, explaining the risks associated with the use of these tools and the requirements around acknowledging their use.

Spark AI

Spark AI is a platform created by the University of Melbourne to provide a safe and secure environment where staff can experiment with generative AI using information that may be sensitive to the University. The platform helps to safeguard privacy, data security and intellectual property rights within the University. It is also helping to build AI literacy and capability among our staff by improving access to AI tools, with more than 1,000 registered users of Spark AI among University of Melbourne staff.

The University has also developed an AI learning assistant, a chatbot embedded in the Learning Management System (LMS), that responds to student queries about course content, based on information provided and guidance set by subject coordinators. Currently in beta, the AI learning assistant is available only to staff during Semester 2, 2024. Depending on the results of the trial, the tool may be made available to students in 2025.

Response to the proposals paper

Principles for defining high-risk AI

Some of the proposed principles could be adjusted to provide greater clarity and coverage. For example, principle (c) could be rewritten to give a better understanding of what it is trying to achieve (additions in bold):

> *The risk of adverse legal effects~~, defamation or similarly significant effects on an individual~~* ***on individuals and groups, due to the use of AI technology.***

The principles could also be expanded to consider the legacy that we might leave for future generations to ensure a sustainable development approach. While AI holds great potential, we must meet present needs without compromising future generations' ability to do the same. Assessing AI's impact on future generations should be integral, and using methods like back-casting can help align present actions with a desired future. Wales's Future Generations Act requires an impact assessment on the effects of innovation on the interests of future generations. This provides a useful model for drafting a principle focused on this issue.

Additionally, while it is referenced in principle (e), sufficient attention is not placed on the potential environmental harm that AI technologies may cause. As with other historical developments in technology, AI

poses a major but largely unknown risk to the environment. The United Nations Environment Programme recommends that governments should ensure environmental risk assessments and mitigation strategies are incorporated into AI management policies. Environmental harms could be called out with the inclusion of:

> *The risk of adverse impacts to the environment.*

Recommendation: *Refine the proposed principles for assessing high-risk AI to improve their clarity and coverage, including impacts on future generations and the environment.*

Application of guardrails to specific sectors

The proposed AI guardrails, while critical for ensuring safety in high-risk settings, do not neatly apply across all sectors, particularly in higher education and research. These sectors often engage in experimental, exploratory uses of AI that may not align with typical commercial applications, and which could be unintentionally constrained by regulations designed for more immediate or high-risk deployment contexts. Universities and research institutions frequently develop AI models for academic purposes that may not directly impact public safety or individual rights in the same way as commercial or industrial AI systems.

For example, when researching AI models, requirements to "manage data quality," "publish an accountability process," or undertake "conformity assessments to demonstrate and certify compliance" need to strike a balance between maintaining safety and allowing the exploration that underpins research. Some of these necessary constraints are already implicit in the *Australian Code for the Responsible Conduct of Research* to which all Australian universities must adhere. As a result, these additional requirements may stifle research rather than enhance safety.

Additionally, research is inherently transnational and work is often published outside the jurisdictions where the researchers live or work, or is conducted collaboratively across borders. This creates challenges for national regulations.

As such, regardless of the regulatory approach taken, further sector-specific consultation and guidance on the guardrails will be required to ensure that the framework can support innovation across the economy while safeguarding the public.

Recommendation: *Conduct sector-specific consultation and provide tailored guidance on the guardrails to ensure the framework supports innovation across the economy while safeguarding the public. This will be necessary regardless of the regulatory approach taken.*

Development and deployment of AI

There is a lack of clarity in the distinction between the development and deployment stages of AI. These stages involve different actors, risks, and responsibilities. This distinction is even more complex in a university or research context – what is deployment in a research context?

This distinction is particularly important when considering guardrail 4 (test AI models and systems to evaluate model performance and monitor the system once deployed). Would this guardrail apply to prototypes in research contexts? If so, it could be read as specifying a single appropriate research methodology or even dictating how research should be done.

Clear principles should delineate the obligations and risks associated with each phase. Defining responsibilities during development and deployment will help avoid gaps in accountability and improve safety measures.

Recommendation: *Establish clear, distinct principles that differentiate between the development and deployment phases of AI, including in research contexts.*

Expertise in AI regulation

Regardless of the regulatory approach taken by the Government (domain-specific, framework, or whole-of-economy approach), it is crucial that any regulatory body overseeing AI has appropriate expertise and training in AI. Given the complexity and rapidly evolving nature of the technology, regulators without AI knowledge may struggle to effectively address the associated risks and nuances. Expertise in AI will ensure that the regulatory framework is both practical and effective.

Recommendation: *Ensure that any regulatory body overseeing AI has appropriate expertise and training in AI.*

Public access to information and standards

The paper references several ISO standards, many of which are closed, that is, accessible only behind a paywall. To ensure transparency and wide adoption of safe AI practices, it is essential that any ISO standards used in Australia's AI guardrails are made publicly available. Restricting access to critical standards can hinder compliance, particularly for smaller businesses and the general public.

Separately, some of the guardrails emphasise making key information publicly available. Transparency is critical and this information should be made public. However, the primary audience for this information will be consumer advocates who wish to hold AI developers/deployers accountable, rather than the general public.

Recommendations:

- *Ensure that any ISO standards referenced in the guardrails are publicly available.*
- *Reframe the guardrails requiring organisations to make information publicly available, clarifying that the primary audience for this information will be consumer advocates.*

Interaction with copyright laws

The paper mentions the work of the Attorney-General's Department and its Copyright and AI Reference Group in considering how Australia's copyright laws deal with AI. This work is important, as the ethical and legal boundaries surrounding the training of models on copyrighted materials without proper licensing are highly contested. Additionally, there is debate over the extent to which a model's outputs must avoid infringement.

In Australia, copyright law does not have a broad "fair use" doctrine, in contrast to the United States. Instead, it relies on "fair dealing," which is more limited and applies only to specific purposes. For example, researchers can use a "reasonable portion" of copyrighted works for their research or study. Generally, this means up to 10% of a book or one chapter, whichever is greater. Use outside of these narrowly defined categories may require permission from the copyright holder. Greater clarity is needed around other uses in digital and research contexts.

Recommendation: *Offer clearer guidance and certainty regarding the application of copyright laws to AI.*

AI literacy training

The EU AI Act specifies obligations for providers and deployers of AI to ensure AI literacy among staff:

> *Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used.*[1]

Drawing from the EU AI Act, there should be an obligation on organisations to train staff on AI systems' operations and impacts. This will ensure that employees understand both the benefits and risks of the AI they use, promoting safer deployment and reducing the risk of misuse or harm.

Recommendation: *Include an obligation for organisations to ensure AI literacy among staff.*

Alignment with practices overseas

Australia's regulatory framework should align with international practices to prevent excessive burdens on businesses, especially small and medium-sized enterprises. Similarly, if the regulatory burden is too high, some overseas companies may decide not to sell their products in Australia, stifling innovation and economic growth. We saw this in the EU, where Apple has announced that it will not roll out Apple Intelligence there due to the "regulatory uncertainties." By aligning with global practices, Australia can foster innovation while ensuring safety, avoiding a regulatory environment that disadvantages local businesses.

Recommendation: *Align Australia's regulatory framework with international practices to prevent excessive burdens on businesses, particularly small and medium-sized enterprises.*

[1] European Union. (2024). *European Union Artificial Intelligence Act*. Chapter 1, Article 4: AI literacy. https://artificialintelligenceact.eu/article/4/

The University of Melbourne

Grattan Street, Parkville, Victoria 3010 Australia

t 13 MELB (13 6352)

+61 3 9035 5511 (International)

unimelb.edu.au

THE UNIVERSITY OF MELBOURNE