UNSW Submission – Introducing mandatory guardrails for AI in high-risk settings

UNSW welcomes the opportunity to comment on the Department of Industry, Science and Resources' proposals paper for *Introducing mandatory guardrails for artificial intelligence (AI) in high-risk settings*.

This submission is primarily focused on teaching and learning as the UNSW Artificial Intelligence Institute is preparing a separate submission focused on research.

Executive Summary

The Department of Industry, Science and Resources should consider the following as part of its consultation on introducing mandatory guardrails for AI in high-risk settings:

Developments in the higher education sector

1. Maintaining academic excellence and integrity in research, teaching, learning, and assessment will continue to be a priority as universities adapt to and lead in the responsible use of AI, bearing in mind expectations from employers in an evolving AI world. We can support government efforts to build trust in AI by ensuring students and the public are educated in its ethical, responsible, and safe use.
2. The Government should continue to work closely with and consult leading experts on the risks of AI, including university academics and researchers. The Government could leverage this expertise to increase awareness of the types of risks that are specific to the higher education sector.

The proposed guardrails

3. UNSW supports a principles-based approach to risk management, which is more adaptive and technology-neutral in its application. The proposed principles for guiding consistency in assessment by organisations of whether the use of an AI system is high risk are appropriate, particularly if they protect existing rights, impacts and protections.
4. We support reinforcing existing regulatory mechanisms. As currently proposed, the guardrails amount to an entirely new regulatory regime, which we do not support. For organisations such as universities, which could participate in any level of the AI supply chain, there are existing regulatory regimes and legislative requirements that do or can be adapted to cover risks and harms posed by this emerging technology.
5. We suggest focusing on the core issues that need to be protected (fairness, privacy, etc) by extending and/or modifying those protections to ensure they capture new ways of operating, rather than through a new regulatory or legislative regime which risks ageing out of relevance as these new technologies continue to develop at pace.

Defining high-risk AI

6. We recommend implementing a technology-neutral approach which allows organisations to identify risks and frame a policy response that is not dependent upon the conception of the technology associated with those risks.
7. List-based use cases, as identified later in the proposals paper, have useful illustrative purposes for organisations seeking to assess their activities, and can be helpful for governments and industry to identify and share best practice and new applications (along with their appropriate management) of AI and other technologies. However, no list of use cases can provide full coverage for the proposed regulatory options being considered by government to mandate guardrails for such a rapidly evolving technology.
8. In the case that list-based use cases are utilised by government, we strongly recommend that education and training, including higher education, are not considered as *de facto* high-risk across all of their activities. The use of AI in education environments, such as UNSW, encourages responsible, safe, and ethical experimentation from both a teaching and learning, and a research and development perspective.
9. Government should apply a balanced approach to use cases and defining high-risk settings, so as not to unintentionally capture some low risk uses through generalisations, and better enable potential future use cases of AI during suitable periods of review. It should be noted that within some of the specified high-risk domains, there may be individual use cases which would pose limited or minimal risk.

Regulatory options for mandating the guardrails

10. Any proposed options for regulating AI should be flexible and adaptable to keep up with technological advancements. The best way of achieving this is through analysis of existing regulatory and legislative functions that are designed to protect, for example, privacy, identity, intellectual property rights, use of data, freedom from discrimination, consumer protection etc, and consider amendments that enable these protections to be robustly maintained, independently of the emerging field that is AI, and future technologies
11. The Government should focus primarily on Option 1. This includes adapting existing legal and regulatory frameworks to ensure they are effective in achieving their goals. The primary benefit of this approach is that it aligns the law with regulatory objectives and values, including protections, rather than with a point-in-time conception of an evolving technology.

About UNSW

UNSW is ranked in the world's top 20 universities. UNSW is a world-leading research and teaching-intensive university, known for innovative, pioneering research and high-quality education with a longstanding global impact. Since our foundation in 1949 and through celebrating our 75th anniversary year, our aim is to transform all lives through excellence in research, outstanding learning and teaching experiences, and a commitment to advancing Australia's economic growth and prosperity.

UNSW proactively equips both staff and students to navigate the realm of AI-assisted learning and professional practice, aligning with the UNSW mission of developing students and researchers who are globally-focused, are rigorous scholars, capable of leadership and professional practice in an international community.

Developments in the higher education sector

AI has ushered in substantial transformations within the higher education sector – disrupting conventional assessment methodologies, altering student learning paradigms, and prompting evaluation of research conduct. The impacts of AI are complex and wide-ranging for universities:

- Learning and teaching: Through new and emerging technologies, academics can reimagine learning and teaching practices. In similar ways to how the COVID-19 pandemic encouraged an accelerated adoption of digital technologies, the emergence of AI has encouraged staff to adapt their assessment methodologies (e.g. using oral assessments; in-class quizzes; and assessments where students reflect on what they have learnt). With privacy safeguards, AI tools can assist with indicative marks and generating feedback. The workload efficiencies AI promises may well mean that staff have more time to create personal connections with students, enhancing the overall student experience.
- Research: For researchers, the availability of large language models can have both positive and negative implications for their work. On one hand, these models can assist with data analysis, text summarisation, and other tasks that can speed up the research process. On the other hand, there is a risk of researchers relying too heavily on these models, which can lead to unoriginal work and potential plagiarism. Additionally, the ease of generating realistic-looking text and data with language models raises concerns about the validity and reliability of research results.
- Integrity: AI tools can often produce reasonable responses, blurring the distinction between genuine understanding and algorithmic output. One of the key concerns expressed about the use of AI in higher education relates to the potential for students to misuse the technology (e.g. relying on AI rather than generating their own ideas, and not referencing appropriately). These concerns are being addressed by universities through a range of strategies to provide our students, regulators and the public with confidence in the world-class education Australian universities provide (e.g. through student education campaigns to promote academic integrity and avoiding cheating and academic misconduct[1]; using a variety of assessment methods[2]; and encouraging critique of generative AI output). On the other hand, AI is an opportunity for the fundamental purposes of university education to develop critical and reflective thinking. Used responsibly, AI provides the opportunity to develop these uniquely human skills. Given the rapid pace of change, researchers, journals, and organisations are also considering their position on the appropriate use of these tools, especially in the context of research integrity.
- Biases: A study by Gartner found that "85 per cent of AI projects will deliver erroneous outcomes due to bias in data, algorithms, or the teams responsible for managing them". Engaging students in critical reflection on AI generated outputs encourages them to scrutinise the underlying assumptions, biases and limitations of AI technologies. This process not only sharpens their judgement but also cultivates a healthy scepticism and curiosity, essential for using generative AI responsibly. For researchers, the text generated by language models can contain biased, false, or harmful information, and it can be difficult to determine who is responsible for this content. It is becoming increasingly important for researchers to use language models in an ethical and responsible manner and to be transparent about their use in their methods and reporting.

[1] For example, UNSW's Key principles in AI usage and assessment, Six Categories of Permissible Use of AI in Assessment, Supporting Academic Integrity

[2] For example, UNSW's guide on designing alternative assessments in the world of AI

- Authorship: Academic authorship is generally reserved for individuals who have made significant contributions to the design, conduct, or analysis of a study or other research project. While language models can assist researchers in generating text, they do not have the capability to independently design, conduct, or analyse research, and therefore should not be considered as authors. However, it may be appropriate to acknowledge the use of language models in the methodology or materials and methods section of a study or research paper, to ensure transparency and to allow for accurate interpretation and replication of results.
- Students and equity: While there may be students seeking an advantage by using AI tools, for the most part we have found that students want to do the right thing and want to make sure that they are using emerging tools appropriately. Some students have reported feeling anxious and uncertain around the use of AI in their studies and assessment, without breaching university guidelines. A major issue raised in discussions regarding the use of AI in education is the impact on equity and diversity in education. Assumptions are made that students coming into university are familiar with the use of AI. A critical gap for students and staff is training and increasing comfort levels with AI.

Maintaining academic excellence and integrity in research, teaching, learning, and assessment will continue to be a priority for UNSW as it adapts to and leads in the responsible and safe use of AI, bearing in mind expectations from employers in an evolving AI world. We can support government efforts to build trust in AI by ensuring there is sound education for students and the public in the ethical, responsible and safe use of AI, including policy professionals with responsibility around governance and integrity.

The Government should also continue to work closely with and consult leading experts on the risks of AI, including at universities such as UNSW. The Government could leverage this expertise to increase awareness of the types of risks that are specific to the higher education sector.

The proposed guardrails

We have a strong commitment to progress for all, supported by world leading AI researchers and ethicists, exceptional teachers and a thriving community that is deeply engaged in responsible experimentation. In this context and more broadly from a societal perspective, we see our role is to help contribute to digital literacy and developing the critical thinking skills of our students and staff to understand what AI is and ensure AI is used ethically, responsibly, and safely.

As an early adopter, UNSW supports the ethical and responsible use of AI in research, learning, teaching, administration, and thought leadership. For example, UNSW's *Ethical and Responsible Use of Artificial Intelligence* guide assists the University in the development and deployment of AI. The principles are aspirational, outcomes-focused, and seek to effectively balance regulation with innovation. We recognise that we will need to continue to update our policies and guidelines to ensure they remain appropriate as new AI products and technologies emerge.

Additionally, within the higher education sector, universities such as UNSW already have systems and processes in place that reflect the type of guardrails proposed. This includes robust mechanisms to identify and manage risk, and strategies for regulatory compliance, such as guidance for the safeguarding of data and organisational avenues to raise concerns or complaints.

UNSW supports a principles-based approach to risk management, which is more adaptive and technology-neutral in its application. The proposed principles for guiding consistency in assessment by

organisations of whether the use of an AI system is high risk are appropriate, particularly if they protect existing rights, impacts and protections.

We therefore support reinforcing existing regulatory mechanisms. As currently proposed, the guardrails amount to an entirely new regulatory regime, which we do not support. For organisations such as universities, which could participate in any level of the AI supply chain, there are existing regulatory regimes and legislative requirements that do or can be adapted to cover risks and harms posed by this emerging technology.

This includes focusing on the core issues that need to be protected (fairness, privacy, etc) by extending and/or modifying those protections to ensure they capture new ways of operating, rather than through a new regulatory or legislative regime which risks ageing out of relevance as these new technologies continue to develop at pace.

Defining high-risk AI

Firstly, defining AI as a *'regulatory object'* is an almost impossible task. All agencies and governments that have attempted this task have struggled. For example, the European Union's definition of AI has been amended multiple times throughout the drafting process. The OECD's definition of AI on which the Australian definition is based has also already been amended. The struggle here is not that any given definition is 'good' or 'bad', it is that we do not yet understand the object that we are trying to define.

While defining AI is a difficult task at this stage of the technology's development, there are risks with which government is rightly concerned. Identifying high-risk scenarios is a useful exercise preliminary to law reform, particularly to better understand broader clusters of problems. These include, for example, the bias that can result from machine learning approaches which rely on inferences drawn from statistics, issues relating to non-transparency and accountability, and ownership and consumption of data, including public and individual data. For the purposes of regulation, however, an adaptive, technology-neutral approach focusing on mitigation of risks is likely to be more effective, as it is not dependent upon the conception of the technology associated with those risks.

Secondly, we note that the paper has specifically identified '*education/training'* as a potential high-risk use case based on international developments, where AI systems intend to be used to:

- Determine access or admission to educational institutions
- Evaluate learning outcomes
- Assess the appropriate level of education that individuals will receive or can access
- Monitor and detect prohibited behaviour of students during tests

List-based use cases, as identified in the proposals paper, have useful illustrative purposes for organisations seeking to assess their activities, and can be helpful for governments and industry to identify and share best practice and new applications (along with their appropriate management) of AI and other technologies. However, no list of use cases can provide full coverage for the proposed regulatory options being considered by government to mandate guardrails for such a rapidly evolving technology.

In the case that list-based use cases are utilised by government, we strongly recommend that education and training, including higher education, are not considered as *de facto* high-risk across all of their activities. The use of AI in education environments, such as UNSW, encourages responsible, safe, and ethical experimentation from both a teaching and learning, and a research and development perspective.

Further, government should apply a balanced approach to use cases and defining high-risk settings, so as not to unintentionally capture some low risk uses through generalisations, and better enable potential future use cases of AI during suitable periods of review. It should be noted that within some of the specified high-risk domains, there may be individual use cases which would pose limited or minimal risk.

We therefore recommend that the most appropriate approach is to provide protections against potential harms, rather than implementing a new regulatory regime. This has the additional benefit of minimising operational disruptions for organisations and businesses, so that they can integrate new regulatory compliance requirements into existing processes without developing and resourcing entirely new processes and procedures.

Regulatory options for mandating the guardrails

Any proposed options for regulating AI should be flexible and adaptable to keep up with technological advancements. The best way to achieve this is by analysing existing regulatory and legislative functions that are designed to protect, for example, privacy, identity, intellectual property rights, use of data, freedom from discrimination, consumer protection etc, and considering amendments that enable these protections to be robustly maintained, independently of the emerging field that is AI, as well as future technologies.

The Government should therefore focus primarily on Option 1. This includes adapting existing legal and regulatory frameworks to ensure they are effective in achieving their goals. The primary benefit of this approach is that it aligns the law with regulatory objectives and values rather than with a point-in-time conception of an evolving technology.

Conclusion

Thank you for the opportunity to provide a submission to this consultation.

Should you wish to discuss any issue raised in this submission, please do not hesitate to contact our Senior Government Relations Manager, Ms Cassandra Switaj, on 02 9348 2246 or c.switaj@unsw.edu.au.