Centre for Data Science

Response to Proposals Paper on introducing mandatory guardrails for AI in high-risk settings

Prepared for the Australian Government, Department of Industry, Science and Resources

Submitted 4 October 2024

By the Queensland University of Technology (QUT) Centre for Data Science

Lead Authors: Bernadette Hyland-Wood [*1,3,4,5] and Aaron Snoswell[1,2,3,4,5]

Contributing Authors: Sandeep Reddy[6], Chun Ouyang[1,8], Dimitri Perrin[1,7], Erwin Fielt[1,8], Aiden Price[1], Kerrie Mengersen[1]

[*] Corresponding Author: **b.hylandwood@qut.edu.au**

[1] Queensland University of Technology Centre for Data Science
[2] Queensland University of Technology GenAI Lab
[3] Queensland University of Technology Digital Media Research Centre
[4] Queensland University of Technology School of Communication
[5] Australian Research Council Centre of Excellence for Automated Decision-Making and Society
[6] Queensland University of Technology School of Public Health and Social Work
[7] Queensland University of Technology School of Computer Science
[8] Queensland University of Technology School of Information Systems

Centre for Data Science

Table of contents

About the QUT Centre for Data Science

Data science is a powerful force for addressing our challenges across all sectors — health, environment, business, society and industry. Its ability to solve global challenges is at the forefront of discussions and strategic activity in most commercial and government organisations, research entities and universities.

The Queensland University of Technology Centre for Data Science (CDS) brings together QUT's strengths and track record in data analysis, computation and research, as well as cross-disciplinary projects and connections to centres of excellence, cooperative research centres, and national and international networks.

The CDS also draws together capabilities in data science from across Australia, providing a centralised hub for world-class data science research, unique training opportunities, and active external engagement. At the CDS, we leverage our diverse capabilities in data science to deliver world-class research, unique training opportunities and active external engagement through:

1. Building Australia's data science network and research profile
2. Bridging the gap between deep research and applications
3. Creating a trusted source for advice on data science issues
4. Bringing together a critical mass of data science researchers to develop start-of-the-art solutions
5. Creating partnerships through an Australian data science network
6. Providing research and professional training opportunities for undergraduate students through to mid-career researchers and senior professionals
7. Leveraging investment in data science (e.g. strategic appointments, Australian Research Council Centres of Excellence and eResearch)
8. Prioritising approaches and methods that embed Indigenous Cultural and Intellectual Property (ICIP), Caring for Country and Indigenous Data Sovereignty (IDS) concerns in culturally appropriate ways.

For general inquiries, please get in touch with us at datascience.submissions@qut.edu.au.

For inquiries relating to this submission, please get in touch with the corresponding author, Dr. Bernadette Hyland-Wood, b.hylandwood@qut.edu.au

Our interest in the Proposals paper for introducing mandatory guardrails for AI in high-risk settings

This submission is a collaborative output with contributions from leading scholars and practitioners in data science, public health, and social work affiliated with the **Queensland University of Technology Centre for Data Science.**

The QUT Centre for Data Science draws together nationally and internationally renowned experts in ethical and explainable AI (XAI), data governance, data sovereignty, data federation, digital media research, public health and social work, and technology law. We collaborate to develop novel data science methods for domain-specific research to solve real-world challenges for our clients. Our postgraduate students and research experts are advancing the medical, physical, technological, and social sciences by developing and applying new data science methods to achieve rigorous scientific discovery and generate evidence-based insights.

Our work supports critical technologies impacting Australia's national interest, including economic prosperity, national security, and social cohesion. Our Centre's domain expertise in advanced information and communication technologies, autonomous systems, robotics, and health, biological, environment and natural systems allowed us to prepare this response to the Department of Industry, Science, and Resources regarding the discussion paper on introducing mandatory guardrails for AI in high-risk settings.

We enthusiastically support the Australian Government's focused attention to AI in high-risk settings including critical infrastructure, healthcare, financial services, and national security. These areas require heightened scrutiny and ongoing monitoring to anticipate and mitigate adverse impacts on the Australian nation. We support the principle of helping the industry balance innovation with the government's role of regulating, ensuring that AI technologies benefit all Australians while protecting individuals and the economy from unintended harm.

We have focused our responses on the questions where Centre members possess the most expertise. Therefore, some of the questions in the proposal paper are not addressed below.

In preparing our response, we considered existing reports, including the discussion paper on Supporting responsible AI, Rapid Response Information Report on Generative AI: Language models and multimodal foundational models (Australian Government, 24 Mar 2023), Critical Technology Statement (Australian Government, 22 May 2023), the Department of Industry, Science and Resources publication *8 AI Ethics Principles aimed at safe, secure and reliable AI*, and *Australia's AI Ethics Framework* (7 Nov 2019), informed by the Human Rights Commissioner's Human Rights and Technology Project, the Attorney-General's review of the Australian Privacy Act. Several members of the QUT Centre for Data Science's authorship team have reviewed the Defence Science and Technology Group's Technical Report: A Method for Ethical AI in Defence and the EU AI Act.

Introduction

On behalf of the QUT Centre for Data Science, we would like to thank you for this opportunity to provide evidence-informed feedback and recommendations.

This submission supports a risk-based approach to ensure AI is developed and used safely and responsibly in Australia, particularly in high-risk applications such as healthcare (including medical devices), transportation, law enforcement, and financial services.

Our recommendations aim to align with fairness, transparency, and human rights principles, which must be articulated and regularly reinforced within the government as an exemplar to industry. Our recommendations are designed to be incorporated within various mechanisms, including principles, frameworks, regulations, amendments, and future AI legislation.

- **Fairness** - Policies should be designed to treat all citizens equitably.
- **Transparency** - Government actions and decisions should be open to public scrutiny. One way this is achieved is by investing in supporting robust open government data programs (Berman & Carter, 2018; CSIRO, 2024; Eaves et al., 2019; Hyland-Wood, 2021).
- **Human rights** - public decisions must respect and protect individual rights.

We appreciate the Department of Industry, Science and Resources' leadership on supporting safe and responsible AI, and the proposed mandatory guidelines for high-risk AI. We believe that without public trust, the benefits of AI may not be fully realised. Public concerns related to transparency and accountability in AI-powered decision-making, particularly in areas such as employment, healthcare, and delivery of government services, must be addressed proactively.

We recognise that *principles* are intended to express foundational values and ethical or normative guidance for contemplated decisions and actions. *Frameworks* move closer to defining a structured approach that organises policies, actions, and strategies around objectives. Frameworks provide a more defined, step-by-step approach than more broadly defined principles.

To date, the Department's publications offer helpful guidance on how policies should be pursued and used to shape sectoral initiatives, AI principles and frameworks; however, they are *not enforceable as laws*. We further acknowledge that regulatory options to mandate AI guardrails for high-risk AI will be a time-consuming and potentially unrealisable goal.

We are recommending a framework approach to existing legislation with amendments in the *near term*, with mandatory guardrails for AI in high-risk settings. We recognise that several countries leading AI regulation reforms are adopting a risk-based approach focusing on *ex-ante* (preventative) measures, which is laudable. However, given the complexity and pace of AI technology development and uptake, a regulatory regime that anticipates *preventative intervention* or *mandating standards* will be **insufficient**.

Preventative intervention **and** ***mandating standards*** **cannot fully satisfy the objective and substantiated concerns of** ***algorithmic bias and discrimination***, ***lack of transparency and accountability***, ***erosion of cultural and data sovereignty, digital exclusion and access gaps, and surveillance and privacy concerns*****.**

In the following responses, we provide feedback and justify our recommendations.

Consultation Topic 1: Defining high-risk AI

Question 1: Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove?

Please identify any:

- low-risk use cases that are unintentionally captured
- categories of uses that should be treated separately, such as uses for defence or national security purposes.

Our response:

At a high level, the broad set of proposed principles for identifying high-risk AI seem well aligned with risks including bias due to inaccurate, insufficient, unrepresentative or outdated data; bias in design or configuration; and harms to people, groups of people, organisations, and collective harms to society. Emerging norms, such as those described by members of the research community interrogating automated decision-making, AI fairness, accountability, and transparency, are considered, as are key principles from the EU's AI Act and Canada's proposed AI Data Act (AIDA).

Modern AI systems risk amplifying often incomplete, lossy or inaccurate data and information. For example, modern AI systems risk amplifying social disadvantage for First Nations peoples. GPT-based systems are trained on vast, often historical datasets that may promote a deficit narrative or a framework that portrays Indigenous peoples and communities primarily based on their shortcomings, problems and disadvantages (Walter et al., 2020). Also see our responses to Questions 2 and 9, below.

Question 2 text: Do you have any suggestions for how the principles could better capture harms to First Nations people, communities and Country?

Our response:

Australia can lead by incorporating Indigenous rights supported by observing Indigenous Data Governance and Indigenous Data Sovereignty into AI regulations.

Notably, Canada's current AI and Data Act lacks explicit protections for First Nations communities.

GPT-based systems, trained on vast public datasets, can unintentionally misuse Indigenous cultural knowledge, especially when sensitive or sacred information is involved. To mitigate potential harms, the Australian Government must co-design AI policies with Aboriginal and Torres Strait Islander peoples, ensuring their voices guide regulation concerning the use of Indigenous data. To achieve this, the views and expertise of Aboriginal and Torres Strait Islander people, including Elders, Traditional Owners and Native Title holders, communities, and organisations,

must guide any discussions on policy and regulation due to the reliance on *Indigenous data.* We strongly recommend that mandatory AI guardrails are co-designed with existing well-defined principles of Indigenous Data Governance and Indigenous Data Sovereignty (Carroll et al., 2019, 2021; Walter et al., 2020 & 2021; Indigenous Data Network, 2024).

It is well understood that GPT-based and other model-based systems are trained on vast publicly accessible text from websites, academic articles, and government sources. This can be problematic for First Nations peoples. Cultural knowledge that is accessible does not infer that it is appropriate for use in training GPT-based systems. This is especially true if the data and information involve sensitive or sacred information.

Australia has an opportunity to demonstrate leadership globally on this topic. Notably, Canada's AI and Data Act (AIDA), a subcomponent of Bill C-27, currently has no explicit provisions for the rights or concerns of First Nations, Inuit, and Métis communities about AI regulation. This is despite the overarching goal of AIDA to ensure the responsible and safe development of AI systems, specifically high-impact AI systems, by incorporating standards for privacy protection, human rights safeguards, and anti-discrimination (source: ISED-ISDE Canada https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document).

Given the speed with which GPT-based systems are being used within government settings, this may disproportionately impact Australia's Aboriginal and Torres Strait Islander people. The Australian Government has a vital role to play in co-designing a risk-based approach that mitigates potential harms to First Nations peoples, communities and Country. To achieve this, the views and expertise of Aboriginal and Torres Strait Islander people, including Elders, Traditional Owners and Native Title holders, communities and organisations, must guide any and all discussions on policy and regulation due to the reliance on *Indigenous data.*

> *Indigenous data* refers to information, in any format or medium, collected, analysed, stored, and interpreted within the context of Indigenous individuals, collectives, populations, entities, lifeways, cultures, knowledge systems, lands, biodiversity, water and other resources. It includes data collected, used, or stored by any agency, department, laboratory, organisation, corporation, statutory body, university, or research institute, conducted by, with, and about Indigenous people or peoples, and data that Indigenous communities have generated and maintained themselves (Indigenous Data Network, 2024).

The use of First Nations peoples' data in training large language models (LLMs) also raises significant concerns related to *Indigenous data sovereignty (IDS)*, or the right of Indigenous peoples and tribes to govern the collection, ownership, and application of their data (Raine et al., 2019). IDS affirms the rights of Indigenous Peoples to control the creation, collection, access, analysis, interpretation, management, dissemination, and reuse of Indigenous data (Kukutai & Taylor, 2016; Snipp, 2016; Maiam nayri Wingara Indigenous Data Collective 2018). We recommend that mandatory AI guardrails be kept up with well-defined principles of Indigenous data government and Indigenous data sovereignty.

The Australian Research Data Commons (ARDC) is currently supporting the development of a national Humanities, Arts and Social Sciences Research Data Commons. The ARDC collaboration with the Indigenous Data Network will support the safe and responsible use of AI by the government and industry in Australia. The HASS Research Data Commons (RDC) is funded to help institutions share data more freely, ethically and cooperatively, following the FAIR data principles and Indigenous data governance protocols maintained by the Indigenous Data Network.

The HASS RDC brings together existing and ongoing investments in text analysis, cultural collections, linguistics, social sciences and indigenous data. It will leverage existing NCRIS capabilities, including the Population Health Research Network (PHRN) and the Australian Urban Research Infrastructure Network (AURIN).

For example, the HASS RDC is developing a framework intended to ensure the proactive identification and mitigation of risks to prevent harm and discriminatory outcomes.

Question 3: Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more defined approach, with a list of illustrative uses, needed?

- If you prefer a list-based approach (similar to the EU and Canada), what use cases should we include? How can this list capture emerging uses of AI?
- If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?

Our response:

Due to AI systems' complexity and algorithms' opacity, relying on a principles-based approach is neither practical nor feasible. The burden would be conferred on high-risk AI developers and deployers, incentivising them to voluntarily 'opt-out' through a narrow definition of its relevance.

Please see our response to Question 10 on the distribution of responsibility across the AI supply chain and throughout the AI lifecycle.

Question 4: Are there high-risk use cases that government should consider banning in its regulatory response (for example, where there is an unacceptable level of risk)? If so, how should we define these?

Our response:

We propose that high-risk AI systems, especially those that could result in harm or involve critical decisions like the use of force, *require a human-in-the-loop to approve, oversee, or directly intervene* in the AI's decision-making processes.

The role of human operators would be to verify and approve actions recommended by the AI, ensuring that any decisions involving force or harm are still ultimately judged and then approved or rejected by humans.

We further note that other respondents working in high-risk fields such as autonomous weapons systems and autonomous drones, the Australia Defence Force, national data collection and intelligence organisations, and other defence-adjacent organisations are better placed to respond in detail to this question.

Question 5: Are the proposed principles flexible enough to capture new and emerging forms of high-risk AI, such as general-purpose AI (GPAI)?

Our response is described in related questions herein.

Question 6: Should mandatory guardrails apply to all GPAI models?

Our response:

There are pros and cons to having mandatory guardrails apply to all GPAI models. Classifying all general-purpose AI as high-risk may be painting with a too broad stroke, mainly if using foundation models becomes a common way of developing and deploying new AI systems, as seems likely from current developments in industry practice (e.g., Schneider, Meske, & Kuss, 2024). QUT Centre for Data Science members had varying opinions on this point. However, there was consensus that what is considered 'AI' is infamously tricky to pin down (let alone 'General Purpose') - the meaning of these terms has and will continue to change radically over the coming years.

The boundaries between general and specific purpose models, as we presently understand them, are very permeable, and clear delineation may be very difficult in some cases. For these and other reasons, this seems to be an essential point that should be considered and discussed.

Question 7: What are suitable indicators for defining GPAI models as high-risk? For example, is it enough to define GPAI as high-risk against the principles, or should it be based on technical capability such as FLOPS (e.g. 10^25 or 10^26 threshold), advice from a scientific panel, government or other indicators?

Full response:

General-purpose systems, by definition, could be applied in both low-risk and high-risk use cases, suggesting that the notion of risk stemming from the application of a model and risk from the model itself are distinct concepts.

In our view, there are also specific problems with any attempt to use technical metrics as a definitional tool for high-risk models of any sort.

Firstly, the number of operations used to train a model is a poor proxy for its capabilities. The quality of the data is at least as important as the volume of data used for training, and smaller models trained on small, high-quality datasets can outperform more complex ones (Gunasekar et

al., 2023). An absolute threshold on the number of operations would not be able to capture these nuances.

And secondly, the proposal paper uses the incorrect definition on p.54; FLOPS, as in Floating-Point Operations per Second, is a measure of training throughput, whereas the more correct metric would rather be FLOPs, Floating-Point Operations - a measure of total training operations.

Consultation Topic 2: Guardrails ensuring testing, transparency and accountability of AI

Question 8: Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings? Are there any guardrails that we should add or remove?

Our response:

The ten proposed mandatory guardrails strongly focus on testing, transparency, and accountability, which are critical for ensuring the deployment of safe and responsible AI in high-risk settings. Below, we note some specific areas where the guardrails might be elaborated on or expanded.

a) Increased emphasis on the importance of data-centric risks in the guardrails

Firstly, the proposed guardrails *should* make data, specifically the privacy consequences of training data, more explicit.

Vast public and privately available data are used to train contemporary AI models, including Personally Identifiable Information (PII) (i.e., personal emails, social media posts, and confidential information), and this has led to unintended consequences, particularly around privacy, security, and ethical concerns. Some AI models also pose heightened risks of re-identification of users (even from de-identified data) where AI models cross-reference or triangulate between data, posing risks of re-identification and data breaches (Fell, 2023).

The existing proposed guardrails also did not seem to adequately address the issues of *informed* and *ongoing* consent about data.

b) Consider including a dedicated Guardrail on robust cybersecurity measures for AI systems

To further and comprehensively address the broader range of potential risks posed by AI, additional guardrail-strengthening cybersecurity measures to protect AI systems could be beneficial. For instance, a new guardrail could be included, such as:

Guardrail 11 (proposed): Implement and deploy robust cybersecurity measures to protect AI systems from hacking, data breaches, and adversarial attacks in high-risk settings.

While data governance (Guardrail 3) and system monitoring (Guardrail 4) are addressed, there is no explicit consideration of cybersecurity measures to protect AI systems from malicious activities.

Cybersecurity is especially critical in high-stakes or high-risk domains such as healthcare, national security, autonomous vehicles, critical infrastructure, and finance, where compromised AI systems could lead to catastrophic outcomes for Australian individuals and society.

AI systems in these and similar domains are prime targets for data hacking, data breaches, and adversarial attacks, where malicious inputs are crafted to deceive the AI model into making incorrect predictions. For instance, adversarial attacks can subtly modify input data to make an AI misclassify images – one such example being altering an autonomous driving system to ignore 'stop' road signs.

Thus, a dedicated guardrail would ensure organisations take specific steps to fortify their AI systems against evolving cybersecurity threats. In drafting such a guardrail, reference could be given to existing guidelines such as the European Union Agency for Cybersecurity (ENISA) Cybersecurity of AI and Standardisation Report (ENISA, 2021), or the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (NIST, 2023).

c) Consider introducing redress mechanisms beyond challenging AI outcomes in Guardrail 7

We propose a refinement to the existing guardrails that would introduce a formalised redress and compensation mechanism for people adversely impacted by AI in high-risk settings in Guardrail 7.

Guardrail 7 focuses on allowing individuals to challenge AI outcomes, but there could be more specific guidance on what happens *after* a challenge. Is there a fast-track process for addressing grievances? How is harm quantified and compensated when AI systems malfunction? Redress and compensation mechanisms could provide clearer paths to restitution when AI-driven decisions cause harm (OECD, 2022).

d) Clarification on transparency across the supply chain in Guardrail 8

While transparency across the AI supply chain is essential, *full* transparency about data, models, and systems may only sometimes be feasible or desirable due to intellectual property concerns, privacy concerns, and cybersecurity risks.

For example, requiring *proportional* transparency or ensuring that transparency applies more specifically to critical aspects of AI systems could make this guardrail more practical.

Overly stringent transparency requirements could limit innovation or expose AI systems to IP, privacy concerns, and security vulnerabilities. It is more important to balance transparency with business, privacy and security needs, and for organisations to mitigate risks without compromising proprietary technology or security (Publications Office of the European Union, 2024).

Question 9: How can the guardrails incorporate First Nations knowledge and cultural protocols to ensure AI systems are culturally appropriate and preserve ICIP?

Our response:

The Australian Government should engage with the locus of work, including well-defined principles and application of Indigenous data government and Indigenous data sovereignty, including automated decision-making and AI, which is facilitated by the Indigenous Data Network (IDN).

Established in 2018, the IDN supports and coordinates governance of Indigenous data for Aboriginal and Torres Strait Islander peoples and communities. The IDN is a national network of Aboriginal community-controlled organisations, university research partners, Indigenous businesses, government agencies, and departments.

Additionally, the **Australian Research Data Commons** (ARDC) is leading the development of a national Humanities, Arts and Social Sciences Research Data Commons. The ARDC collaboration with the Indigenous Data Network will support the safe and responsible use of AI by the government and industry in Australia. The HASS Research Data Commons brings together existing and ongoing investments in text analysis, cultural collections, linguistics, social sciences and indigenous data. It will leverage existing NCRIS capabilities, including the Population Health Research Network (PHRN) and the Australian Urban Research Infrastructure Network (AURIN).

The HASS Research Data Commons will help institutions share data more freely, ethically, and cooperatively, following the FAIR data principles and Indigenous data governance protocols maintained by the IDN.

Question 10: Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately? For example, are the requirements assigned to developers and deployers appropriate?

Our response:

No single entity can fully control risks across the AI supply chain for high-risk AI applications. Modern computer systems, especially GPT-based technologies, are highly complex and, by design, have minimal transparency.

While developers, platforms, and regulatory frameworks play a role in mitigating risks, the unpredictable nature of these systems, combined with their continuous evolution, makes *complete* risk control aspirational but unlikely.

Notably, the EU's AI Act also recognises that organisations alone cannot control risks from AI models and applications.

A detailed and prescriptive risk-based approach focusing on *ex-ante* (preventative)

measures to identify, target, and mitigate known AI risks is recommended for high-risk AI applications.

That said, while developers, researchers, legal teams, and the public play a role in identifying, mitigating, and managing risks, **no individual, group, or company can fully control the risks.**

No regulatory regime emphasising preventative intervention or mandating standards fully address the complexity of AI systems. Thus, there is limited opportunity for *shared responsibility* for meaningful risk control.

Modern software engineering, especially with the integration of GPT-based technologies, has introduced powerful capabilities and complex challenges in managing risks associated with AI-driven products and services. This is due to deep learning and massive datasets that are intricate and behave in ways that are not predictable. The complexity of systems is such that only very few highly trained engineers generally understand**.** While a small group of expert developers may understand how a given model is trained and tested, the lack of transparency and explainability makes it nearly impossible to interpret or predict the model's decisions in real-world scenarios. Unlike the example in the Proposals Paper on the therapeutic goods regime where provenance (chain of custody) is carefully and transparently tracked, developers, digital platforms, and AI supply chains are opaque and completely decoupled. One of the largest challenges in GPT-based systems is the 'black box' nature of AI models.

How developers, digital platforms and AI function is opaque and largely hidden from the public, government regulators, and even external researchers, and this has led to a black box problem (Pasquale, 2015; Von Eschenbach, 2021; Zednik, 2021). Modern digital platforms are owned by highly capitalised private companies, primarily based in the USA. Digital platforms and AI vendors' market power and influence has consolidated, leading to concerns around transparency, accountability, and control over public digital spaces.

Modern GPT-based systems leverage deep learning and massive training datasets measured in the hundreds of billions of tokens derived from various sources. The scale of these datasets is critical to the models' performance and ability to generate coherent, contextually relevant responses.

GPT-based systems are often deployed as learning models, meaning they *evolve* based on new data post-deployment. This makes risk control a moving target as the model's outputs behave in not only potentially unintended, but unforeseen ways as the model adapts.

Quality control efforts, such as filtering out low-quality or harmful content, are an active area of applied research. While removing spam, detecting, and addressing misinformation, disinformation, and biased content is likely to remain a priority for most platform vendors, current research suggests that GPT-based models will continue to struggle to reflect accurate information consistently.

While modern software engineering tools and best practices such as continuous integration, continuous deployment (CI/CD), automated testing, and formal verification can *help mitigate risks* introduced by GPT-based systems, **the spread of harmful content, irrelevant information, or**

misinterpretation (due to lack of context) is very likely to persist.

Thus, planning for potential and unintended consequences of AI decision-making systems is recommended. It is recognised that controlling risks arising during development and deployment is neither realistic nor feasible for developers and deployers across the AI supply chain and throughout the AI lifecycle.

Question 11: Are the proposed mandatory guardrails sufficient to address the risks of GPAI? How could we adapt the guardrails for different GPAI models, for example low-risk and high-risk GPAI models?

Our response:

Ensuring compliance with such regulations across various sectors with unique challenges would require oversight and adaptability that exceeds traditional governance structures. As a result, the capacity of mandatory guardrails to effectively contain the risks associated with GPAI is inherently limited by the system's unpredictable deployment and use patterns.

Regulatory frameworks designed for specific use cases are *unlikely* to be sufficiently agile or comprehensive to account for the vast array of potential applications that GPAI enables.

Mandatory guardrails, while essential for *risk mitigation*, *will always be insufficient to fully address the risks associated with GPAI* due to its unpredictable uses and impacts. Our rationale is as follows.

One of the core challenges with GPAI is that its generality allows it to be applied across a wide range of contexts and purposes, many of which may have yet to be envisioned during development. As a result, potential risks—particularly those introduced by unforeseen or novel applications—are difficult, if not impossible, to predict and mitigate in advance.

Though guardrails such as transparency, accountability, and fairness are typically use-case agnostic, their application becomes more challenging when faced with the inherent unpredictability of GPAI. For example, developers may design a system for a specific set of tasks. Still, the flexibility of GPAI means it could be repurposed or extended in ways that introduce new, unforeseen risks. The range of uses increases the possibility of unintended consequences, such as bias amplification, security vulnerabilities, or ethical concerns that were not identifiable in the original development stage.

Additionally, as discussed in the response to Question 10 (above), the evolving nature of GPAI models compounds these predictive difficulties. Unlike narrowly focused AI systems, GPAI adapts and evolves based on new data and environments post-deployment. This continuous learning process means that risks may emerge dynamically over time (perhaps with limited human developer input or oversight), not just at the point of development or initial deployment. As GPAI adapts, it could produce outcomes or impacts that were not originally anticipated, challenging the ability of static guardrails to keep pace with its changing behaviour.

For example, an AI system initially designed for benign purposes could be co-opted for harmful or

ethically questionable uses, such as spreading mis- and disinformation or enabling harmful automated decision-making in unforeseen domains. In such cases, it is difficult for any pre-set guardrail to fully account for the breadth of applications and their potential risks, particularly as GPAI systems can be deployed across industries or regions with different regulatory frameworks and cultural contexts.

While guardrails serve as critical tools in managing AI risks, their general applicability and reliance on pre-defined parameters limit their effectiveness in addressing the unpredictable risks associated with GPAI. The evolving, flexible, and dynamic nature of GPAI systems introduces layers of complexity that make it impossible to foresee all risks or regulate every potential impact, necessitating a more adaptive, ongoing approach to governance and risk management.

Question 12: Do you have suggestions for reducing the regulatory burden on small-to-medium-sized businesses applying guardrails?

Our response:

Given their resource constraints, the feasibility and ability to implement AI guardrails should be considered, specifically for SMEs and start-ups.

To reduce the regulatory burden on SMEs deploying high-risk AI, a tiered compliance framework should be implemented, with lighter documentation, streamlined audits, and scaled-down reporting obligations based on size and risk exposure. Additionally, providing SMEs access to shared compliance resources, including toolkits, AI governance frameworks, and third-party validation services, will help simplify and lower the costs of implementing mandatory guardrails for high-risk AI systems.

One way to reduce the regulatory burden is to have a tiered compliance framework, as with other forms of regulation, based on the size, resources and risk exposure of a given business. As with other forms of regulation, industry or government bodies provide a toolkit that includes checklists, automated compliance software, and guidelines tailored for smaller organisations. SMEs should have access to accredited AI auditing services to meet compliance requirements.

SME's deploying high-risk AI systems could be required by regulators to supply lighter documentation subject to streamlined audits, and scaled-down reporting obligations. Regulatory entities can provide access to shared compliance resources, support and services (leveraging standards and legal templates) to help SMEs and start-ups implement the guardrails, e.g., an AI governance framework, an AI risk management process, third-party validation services, an information sheets, documentation guidelines, to reduce the cost and complexity of implementing the guardrails.

Consultation Topic 3: Regulatory options to mandate guardrails

Question 13: Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach?

Our response:

Establishing a Commonwealth-level Office of the National AI Commissioner with expert leadership would ensure consistent enforcement and early risk mitigation while promoting collaboration with industry and alignment with international regulations. A multifaceted approach, including updating existing laws and regular reviews, would balance innovation with robust oversight.

The Australian government has various regulatory options to address AI use in high-risk environments, including mandatory standards, licensing systems, safety certification, and codes of conduct. Each approach offers distinct advantages and presents unique challenges, with crucial considerations including enforcement complexities, the delicate balance between innovation and regulation, and resource allocation.

To guarantee adaptability, the government should establish regular reviews, foster collaborative relationships with industry stakeholders, and uphold transparency throughout the regulatory process. A multifaceted approach is likely most effective when recognising existing gaps in Australian legislation, specifically concerning data privacy and liability for AI-related damages.

We believe that **Option 3: introducing a new AI-specific Act** — would best address the use of AI in high-risk settings. This approach provides the most consistency across the economy regarding definitions, coverage, and enforcement. It allows for comprehensive pre-market obligations on developers, which is critical for addressing risks early. Some key opportunities the government should consider:

- Designing the Act to complement existing regulatory frameworks, with carve-outs where domain-specific laws already provide sufficient protections (similar to Canada's approach)
- Enabling interoperability with international approaches like the EU AI Act and the US President's Executive Order on the safe use of AI.

Centre members supported moves to establish a dedicated AI regulator or commissioner (e.g., Commonwealth-level Office of the National AI Commissioner) with specialised expertise and jurisdiction over enforcing the high-risk AI guardrails. A commissioner (and associated office) is a better option than a regulatory body, especially given the speed with which the field of AI continues to develop.

This could follow a similar model to the national data commissioner, whose office was appointed following the Productivity Commission's 600+ page report on the government’s data sharing issues. Similarly, a National AI Commission could be appointed, supported by an Office for the

National AI Commissioner.

Concerning the criteria for who is appointed, the AI commissioner should be an expert in the field. The Office of the National AI Commissioner should include representatives from research, industry, First Nations, youth, and other sections of Australian society. Like the e-Safety Commissioner, the commissioner should have some powers rather than simply an advice-giving capacity. It was also noted that there is a need to establish such a body rapidly (i.e. moving past Terms of Reference quickly) so that this team can keep pace with the diffusion of AI technology across Australian society.

Question 14: Are there any additional limitations of options outlined in this section which the Australian Government should consider?

Our response:

Additional limitations to consider:

- For Option 1 (adapting existing frameworks): This could exacerbate regulatory fragmentation and inconsistency across sectors.
- Option 2 (framework legislation): This may still need to include more coverage and enforcement.
- Option 3 (new AI Act): It will require significant resources to establish a new regulator and enforcement mechanisms. See our response to Question 13, above.

Question 15 text: Which regulatory option/s will best ensure that guardrails for high-risk AI can adapt and respond to step-changes in technology?

Our response:

Option 3 (a new AI Act) is likely best positioned to adapt to technological changes. A dedicated AI regulator could more easily update guidance and rules as needed, compared to relying on multiple existing regulators to coordinate changes across different frameworks. The Act could also be designed with flexibility to accommodate future developments.

Question 16: Where do you see the greatest risks of gaps or inconsistencies with Australia's existing laws for the development and deployment of AI? Which regulatory option best addresses this, and why?

Our response:

The most significant risks of gaps/inconsistencies in existing laws appear to be:

- Lack of pre-market obligations on AI developers in many sectors
- Inconsistent definitions and thresholds for high-risk AI across different regulatory regimes

- Limited ability to address cross-sectoral and economy-wide AI risks
- Gaps in accountability for AI supply chain actors beyond just end-users/deployers

We recommend Option 3 (new AI Act) best addresses these by providing comprehensive, consistent coverage and extending obligations across the AI supply chain. It allows for a coordinated approach to defining high-risk AI and implementing guardrails economy-wide.

Additional recommendations on supporting and promoting best-practice governance

Our response:

Australia's active participation in global AI governance initiatives is crucial to shaping responsible AI development and addressing the ethical and societal risks posed by high-risk AI systems. *Genuine engagement* in international collaborations will ensure Australia contributes to and benefits from co-created standards, positioning the nation as a leader in ethical AI innovation and risk mitigation.

The field of AI is evolving very rapidly, and the AI supply chain carries significant ethical and societal considerations due to its scale and speed of uptake. The Australian Government has a once-in-a-generation opportunity to prioritise and support diverse representation in AI governance initiatives through participation in working groups and communities of practice. Representation cannot be tokenistic or a 'tick and flick' once or twice per year exercise for the purposes of *appearing* to be a contributor.

The Australian Government must support its experts and representatives in international standards forums, working groups and communities of practice (national and international) on an ongoing basis.

External oversight, audits, and standards co-created by international initiatives, working groups and communities of practice play an essential role in ensuring responsible AI development and deployment.

Just as the internet has benefited by multilateral cooperation, and longstanding initiatives, governments adopting a risk-based approach must prioritise engagement with international initiatives, working groups and communities of practice. Technical, policy and social scientists, as well as youth and community representatives are vital to the process of defining standards and procedural requirements for risks related to privacy, information and physical security, bias and discrimination as part of a risk mitigation strategy.

For example, forums convene policy advisors and AI innovators from Canada, the United Kingdom, and the European Union. Informed by the Hiroshima AI Process Code of Conduct and the Frontier AI Safety Commitments, representatives from these jurisdictions have been collaborating online and in-person to co-create governance frameworks.

Australian-based applied researchers and advocates (e.g., youth representatives, Indigenous leaders) should be joining these countries, and contributing to these forums. Australian-based researchers and advocates must be empowered and supported in doing so–It cannot remain an altruistic 'side hustle' or supplementary effort supported by a benevolent manager.

Those of us involved in applied research in AI and data governance recognise that the Canadian Government is a leading voice in AI innovation and the first OECD country to create a national strategy for AI (2017). Canada has a *pan-Canadian* AI Strategy comprised of independent organisations dedicated toAI applied research and commercialisation across Canada. The pan-Canadian Strategy is led by three organisations: Amii in Edmunton, Mila in Montreal, and the Vector Institute in Toronto.

As discussed in the Mandatory AI Guardrails Proposals paper, the United Kingdom too has made significant progress in AI frameworks and emerging legislation including establishing the Office for AI (2018), announcing a National AI Strategy (2021), publishing the *Pro-Innovation AI Regulation White Paper* (2023) emphasing a decentralised regulatory approach (versus an overarching AI regulator), and non-statutory guidance, and exploring AI assurance frameworks through the Centre for Data Ethics and Innovation. And the European Council has applied a Data Governance Act and mechanisms to enable the 'safe reuse of certain categories of public sector data that are subject to the rights of others' (EU 2022 Council approves Data Governance Act). See https://www.consilium.europa.eu/en/press/press-releases/2022/05/16/le-conseil-approuve-l-acte-sur-la-gouvernance-des-donnees/

Australia's active participation in global AI governance initiatives is crucial to shaping responsible AI development and addressing the ethical and societal risks of high-risk AI systems. Engagement in national and international collaborations will ensure Australia contributes to and benefits from co-created standards, positioning the nation as a leader in ethical AI innovation and risk mitigation.

We thank the Department of Industry, Science and Resources for this opportunity to provide expert feedback and evidence-informed recommendations. We are available should you wish to discuss any part or the entirety of our response.

The authorship team from the QUT Centre for Data Science

Brisbane, Australia 4 Oct 2024

Additional recommendation: Reorder the 'work plan' per the proposal paper and put the readily achievable tasks first, and put time-consuming objectives last

The Safe and Responsible AI in Australia proposal paper (September 2024) discussed the Australian Government's Safe and Responsible AI work plan (p. 7) and included a Figure 1: *Actions the government is taking to support safe and responsible AI in Australia* (p. 6). While the image did not have arrows indicating order, a material change in how that is represented is required.

We recommend the government is well positioned to *'support and promote best practices'* and '*serve as exemplar*' in the near term (immediate to 12 mos). We believe that government leaders should prioritise active *international engagement* through genuine and substantive participation in international working groups and communities of practice a necessary function of government (immediate and ongoing).

Practically, *delivering regulatory clarity and certainty* is something that will take 5-10 years. Clear, understandable, and stable regulations that define expectations, responsibilities, and legal boundaries for individuals and organisations operating within a specified context are substantial undertakings at the best of times.

References

Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs). Australian Council of Learned Academies.

Berman, E., & Carter, J. (2018). Scientific Integrity in Federal Policymaking Under Past and Present Administrations. Journal of Science Policy & Governance, 13(1). www.sciencepolicyjournal.org

Carroll, S.R., Rodriguez-Lonebear, D., Martinez, A. 'Indigenous Data Governance: Strategies from the United States Native Nations'. Data Science Journal 18 (2019).

Carroll, S.R., Herczog, E., Hudson, M., Russell, K., Stall, S. 'Operationalizing the CARE and FAIR Principles for Indigenous Data Futures'. Nature, Scientific Data 8 (2021). https://doi.org/Https://doi.org/10.1038/s41597-021-008920-0

Carroll, S.R., Ibrahim, G., Figueroa-Rodriquez, O.L., Holbroo, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K. Rodriquez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., Hudson, M., 'The CARE Principles of Indigenous Data Governance'. Data Science Journal 19, no. 43 (2020): 1–12. https://doi.org/10.5334/dsj-2020-043

CSIRO Data Program https://www.csiro.au/en/research/technology-space/data?start=0&count=12 (Retrieved 2 Oct 2024)

Eaves, D., McGuire, B., & Carson, A. (2019). Open data around the world: North America, Australia, and New Zealand. In The state of open data: Histories and horizons (pp. 517–534). African Minds and International Development Research Centre.

European Union Agency for Cybersecurity (ENISA) (2021). Malatras, A., Agrafiotis, I. and Adamczyk, M. (eds). Securing machine learning algorithms. Accessible via https://data.europa.eu/doi/10.2824/874249

European Union (2024). Striking a balance: Open data and privacy in the digital era. Accessible via: https://data.europa.eu/en/news-events/news/striking-balance-open-data-and-privacy-digital-era

Fell, J., (2023). See your identity pieced together from stolen data. ABC News. https://www.abc.net.au/news/2023-05-18/data-breaches-your-identity-interactive/102175688 (Retrieved 1 Oct 2024)

Government of Canada Artificial Intelligence and Data Act, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act (Retrieved 1 Oct 2024)

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., ... & Li, Y. (2023). Textbooks are all you need. arXiv preprint arXiv:2306.11644.

Hyland-Wood, B. (2021). Bridging the Open Data and Public Policy Gap: Barriers and facilitators for effective government data sharing. PhD Thesis, The University of Queensland, School of Political Science and International Studies

Indigenous Data Network, 2024. https://mspgh.unimelb.edu.au/centres-institutes/centre-for-health-equity/research-group/indigenous-studies-unit/indigenous-studies/indigenous-data-network (Retrieved

1 Oct 2024)

Janke, T. (2021). True tracks: Respecting Indigenous knowledge and culture. UNSW Press

Kukutai, T., Taylor, J. Indigenous Data Sovereignty Towards An Agenda. Vol. 38. Canberra: Australian National University, n.d.

Lovett, R., Lee, V., Kukutai, T., Cormack, D., Rainie, S. C., & Walker, J. (2019). Good data practices for Indigenous data sovereignty and governance. Good data, 26-36

Martin, K., & Mirraboopa, B. (2003). Ways of knowing, being and doing: A theoretical framework and methods for indigenous and indigenist re-search. Journal of Australian Studies, 27(76), 203–214 https://doi.org/10.1080/14443050309387838

National Institute of Standards and Technology (NIST) (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Accessible via https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Organisation for Economic Co-operation and Development (OECD) (2022). OECD Social, Employment and Migration Working Papers. Accessible via https://one.oecd.org/document/DELSA/ELSA/WD/SEM(2022)7/en/pdf

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information. Harvard University Press.*

Rainie, S. C., Kukutai, T., Walter, M., Figueroa-Rodríguez, O. L., Walker, J., & Axelsson, P. (2019). Indigenous data sovereignty.

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. Business & Information Systems Engineering, 66(2), 221-231. doi:10.1007/s12599-024-00851-0

The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. [online] The White House. Available at: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), 1607-1622

Walter, M., Kukutai, T., Russo Carroll, S., Rodriguez-Lonebear, D., 2021. Indigenous Data Sovereignty and Policy. Routledge: Oxon

Walter, M., Lovett, R., Maher, B., Williamson, B., Prehn, J., Bodkin-Andrews, G., & Lee, V. (2021). Indigenous data sovereignty in the era of big data and open data. Australian Journal of Social Issues, 56(2), 143-156. DOI: 10.1002/ajs4.141

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34(2),* 265-288


QUT
Centre
for Data
Science