ADM +S

ARC Centre of Excellence for **Automated Decision-Making and Society**

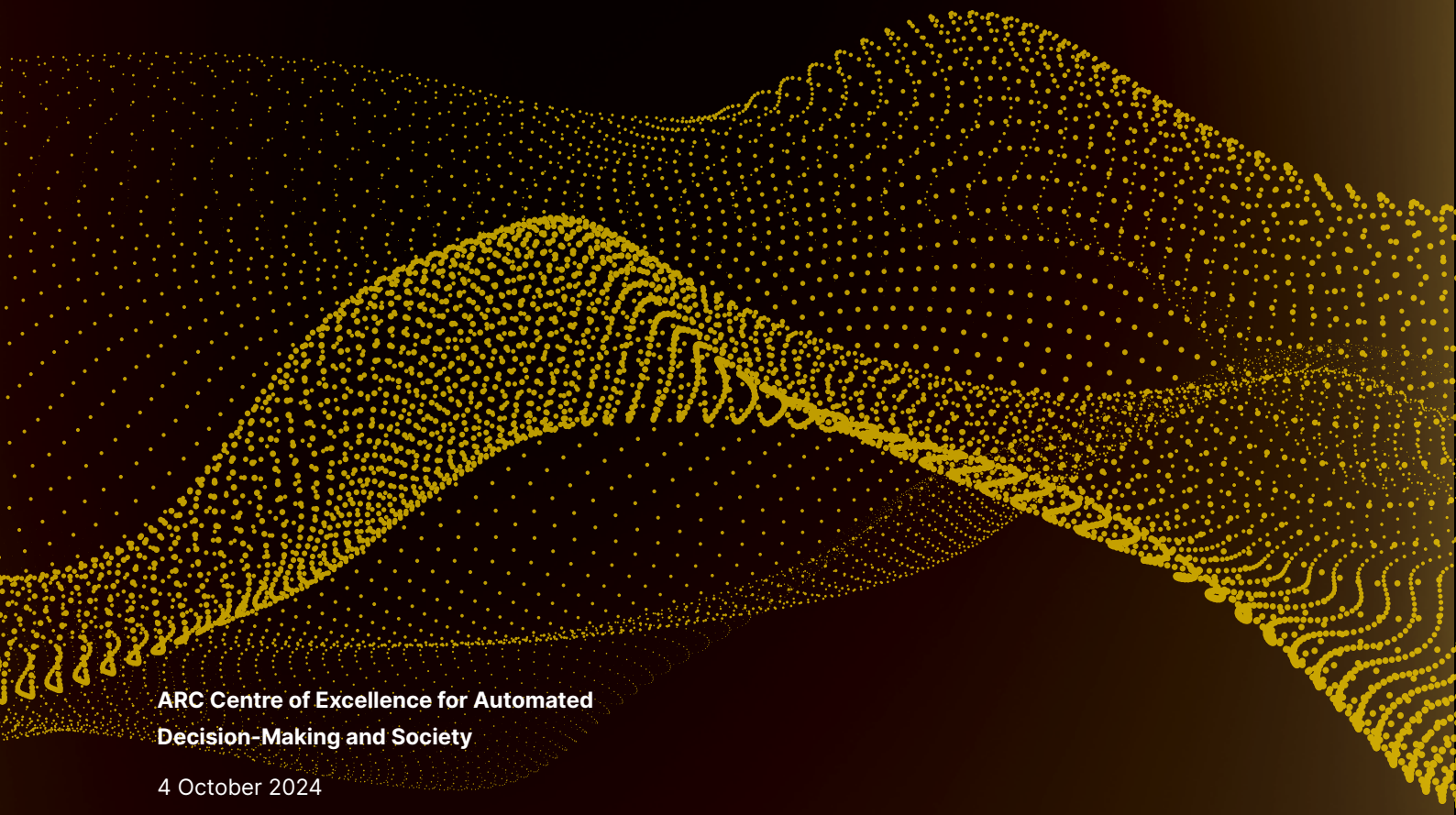
Safe and Responsible AI in Australia: Proposals Paper for introducing mandatory guardrails for AI in high-risk settings

Lead authors

Kimberlee Weatherall, Henry Fraser and Aaron Snoswell

Contributing authors

Daniel Angus, Francesco Bailo, José-Miguel Bello y Villarino, Jake Goldenfein, Paul Henman, Fabio Mattioli, Rita Matulionyte, Anthony McCosker, Luke Munn, Christine Parker, Lyndal Sleep, Milica Stilinovic, Johanne Trippas



ARC Centre of Excellence for Automated Decision-Making and Society

4 October 2024

Acknowledgement of Country
In the spirit of reconciliation, we acknowledge the Traditional Custodians of Country throughout Australia and their connections to land, sea and community. We pay our respect to their Elders past and present and extend that respect to all Aboriginal and Torres Strait Islander peoples today.

Suggested citation
Weatherall, K., Fraser, H., Snoswell, A. (2024). Safe and Responsible AI in Australia: Proposals Paper for introducing mandatory guardrails for AI in high-risk settings. ARC Centre of Excellence for Automated Decision-Making and Society.
DOI: 10.60836/mrgn-9e28

ARC Acknowledgement
This research was funded by the Australian Government through the Australian Research Council's Centre of Excellence for Automated Decision Making and Society (ADM+S) [CE200100005].

Australian Government
Australian Research Council

About ADM+S

The ADM+S is pleased to have this opportunity to engage with the Proposals Paper for introducing mandatory guardrails for AI in high-risk settings. The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) is a cross-disciplinary, national research centre established and supported by the Australian Research Council to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making (ADM). More information about ADM+S and its research may be found on our website, www.admscentre.org.au

This submission

This submission is the product of a collaborative process involving direct contributions from the above researchers from ADM+S, as led and consolidated by Kimberlee Weatherall, with significant contributions from Henry Fraser and Aaron Snoswell. ADM+S researchers come from many different institutions, disciplines and perspectives. It should not be assumed that every contributing author, or every member of the Centre subscribes to every comment or recommendation in this submission. The submission represents our best effort to consolidate research and thinking in a way that can be useful to the government's deliberations on safe and responsible AI and mandatory guardrails. We are happy to put members of the government's team in touch with experts in the Centre for further discussion.

The authors acknowledge the invaluable assistance of Annie Luo and Kathy Nickels in the preparation of this submission.

Key messages

The Australian government has made clear their view: Australia's current regulatory framework — our laws, and our systems for enforcing them — are **not fit for purpose** when it comes to the distinct risks and governance challenges posed by AI. We agree.

One problem is that (we have the risk that) AI gets deployed — rushed to market — without proper testing, creating risks of harm to human rights, democracy, the rule of law; and societies, and the environment that could and should have been mitigated. Most of our laws remedy harms caused by breach of the law after the event, but don't explicitly or specifically require organisations to take steps in advance to reduce harms. To address this problem, the Proposals Paper sets out mandatory, whole-of-economy (public and private sector) processes and requirements ("Guardrails"), to apply to developers and deployers (as appropriate) of high-risk AI systems. "Guardrails" are systems and processes: such as risk management, testing, transparency, accountability, and data governance. The Guardrails are not a complete answer to how, as a society, we will manage AI risks, but they are an **important step, and one that can be taken now**.

The Proposals Paper addresses a range of questions: which systems are high risk and require the application of Mandatory Guardrails; should any AI systems be prohibited; what should be the content of the Guardrails; who in the AI supply chain needs to do what.

The answers to all of these questions need to be driven by **three clear understandings**:

1. Technology in this area is shifting, will continue to shift, and not necessarily in predictable ways. The same is true of business models, investment patterns, societal responses, public concerns, and environmental impacts.
2. Good processes (guardrails) for safe and responsible AI are, to a significant degree, generalisable (although questions such as what substantive uses are appropriate, and what remedies should be provided for harms, may not be). Certainly detailed guidance for different domains would help in implementation.
3. What we need to design is a framework that:
    a. Addresses immediate concerns and risks;
    b. Provides some coherence and coordination of government action;
    c. Is able to be refined, developed, and updated in an ongoing way.

The core imperatives, then, are the need for **action**, and the need for that action to be able to be — and to be — refined, updated and adjusted over time. We need to design processes, more than rules. Achieving these aspects of institutional design must drive the answers to **all** the questions in the Proposals Paper.

[1] The Proposals Paper Attachment E (p 65) sets out a proposed 'division of responsibility' for the Mandatory Guardrails.

This submission therefore presents arguments in favour of:

1. **Foundational, central, legislated Mandatory Guardrails, applied horizontally through the introduction of an AI Act (that is, Option 3).** The Guardrails may not need to be identical across all sectors. In fact, as we note in discussing Topic 5, there are arguments for applying more stringent standards to public sector uses, which could be immediately applied without any kind of staged or delayed implementation, consistent with government's commitment to be an **exemplar** for safe and responsible AI use. Some heavily regulated sectors might be carved out, as is true in Europe.
2. **Updateable detail** of every aspect of the framework: the content of what counts as a high risk system; the content of the Mandatory Guardrails; the determination of which GPAI systems are high risk: must be able to be updated, not through legislation but through more flexible (but binding) means — such as via disallowable instruments. More detailed guidance — for example, on what risks are likely, and how best to mitigate them — can also be provided, including from domain-specific agencies and regulators.
3. **A Central AI Body** — a Regulator, or Commission, with **expertise** (and access to wider networks of expertise and research); **mandate** to impose the Guardrails across public sector uses; and a chief role of keeping the whole system up to date and responsive. Such a body may not be permanent.

We have provided input (below) on various questions of detail that arise in relation to aspects like the definition of high risk systems, and the Guardrails. It is important however not to get lost in these details. Some of the details we discuss below would need to be settled in the process of legislative drafting; others would likely be dealt with in guidance.

But importantly, we are **not arguing for certainty in every detail, in legislation, in advance**. Requiring certainty as to every detail of the framework is misguided, in light of shifting technology. We cannot write highly certain legislation to regulate uncertain and unpredictable technology.

Finally, we note the **limits of what the Mandatory Guardrails can address**. While (particularly over time) there will need to be **consequences for breach** of the Mandatory Guardrails, which could include enforcement action from a regulator, the question of **liability to individuals for AI-related harms** is a **separate (critically important) question**. Liability for AI harms needs to be addressed at two levels:

1. Updates to our substantive laws, including long-overdue updates to privacy law, and consumer protection.
2. Detailed consideration at the level of principle: regarding the appropriate apportioning of responsibility and liability, to ensure that people harmed by AI can recover, should be considered in detail, preferably the Australian Law Reform Commission. There is much international work in this area on which to build, but an Australian-focused discussion is warranted.

What follows is detail in response to each of the questions in the Proposals Paper, in the order set out in that Paper. We have separated out two areas for consolidated discussion: questions relating to **Australia's First Nations people** and **GPAI**. There is a great deal of detail below, contributed by researchers and research from across ADM+S. We have aimed to structure the submission so that it does not have to be read cover to cover to make sense. We are also happy to connect the Department to individual researchers within the Centre on any of the specific points below that are of interest.

Contents

Topic 1: Defining high-risk (and unacceptable risk) AI

Question 1: Principles to identify High Risk AI

Question 1: Do the principles adequately capture high-risk AI? Identify any:

- Low risk cases that might be unintentionally captured, and
- Categories of uses that should be treated separately (eg defence, national security).

Regarding the drafting of these principles, we make four comments:

First, we agree that the listed considerations are relevant to considering a system to be high-risk. Risks to human rights; risks to health and safety; risks of adverse differential treatment of groups; and risks to the economy, society, environment and the rule of law are all relevant. We would advocate for some refinement of the considerations, in consultation with experts in both emerging AI regulation and domestic and international human rights law to avoid redundancy[2] and to align with international instruments such as the *Council of Europe Framework Convention on Artificial Intelligence and Human Rights* ('COE Framework Convention)[3] and equivalent systems overseas:

- The COE Framework Convention considers 'adverse impacts on human rights, democracy and the rule of law'. Although we understand impacts on democracy would fall within 'impacts to the broader society', the specific mention of democracy may serve a useful signalling purpose;
- The Canadian Directive on Automated Decision-Making[4] requires consideration of 'the equality, dignity, privacy and autonomy of individuals'; and 'the economic interests of individuals, entities or communities': potentially providing greater clarity as to the kinds of 'similarly significant effects' on an individual, or 'adverse effects' on a community that ought to be considered in designating a system as high-risk.

Additional or expanded considerations would not by themselves produce new expansive rights to dignity, autonomy, or protection of economic interests. Rather they would simply increase the likelihood of a system's being designated as high-risk, with commensurate guardrails.

[2]For example, international human rights protect health and safety; from an international human rights law perspective the redundancy between (a) and (b) might cause interpretive difficulties.
[3]*Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, opened for signature 5 September 2024, CETS No. 225 (not yet in force) ('COE Framework Convention'). Note that despite the name, the Convention is open to non-European signatories, and has already been signed by Andorra, Georgia, Iceland, Norway (with declarations), Republic of Moldova, San Marino, United Kingdom, Israel, United States of America, and the European Union.
[4]Directive on Automated Decision-Making' (Web Page, 25 April 2023) <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592> ('Canadian ADM Directive'). This Directive applies to Federal Government use of ADM (not only AI); it does not apply to use by the private sector or by provincial governments.

Second, the principles don't set any **threshold** for what is **high** risk. Severity and extent are identified (generally) as considerations. This contrasts with the **Canadian ADM Directive**,[5] which sets out risk levels (I - IV), or the **NSW AI Assurance Framework**[6] that identifies certain scenarios as giving rise to 'elevated risk' (slide 11), and also sets out different risk levels (slides 19-20).

Third, the Principles as drafted do not take explicit account of **monoculture** effects, or the **cumulative** effects of AI adoption in essential services. An AI system that controls individual commercially-supplied thermostats may not be high-risk, but it *becomes* high-risk if the majority of the population is using those thermostats, such that system failure leaves a significant proportion of the population without cooling in a heat wave. The world recently had a taste of monoculture effects in the Crowdstrike cybersecurity incident that locked users out of computers globally. The Crowdstrike incident is also an example of **cascade effects**: where a technology (in that case related to cybersecurity) is embedded into infrastructures that impact many other parts of society, such that failure has widespread and varied impacts across different parts of society (from hospitals to airlines, in that case). The Principles may be broad enough to cover such effects; we highlight this issue as relevant to further refinement of the drafting, or for when guidance is provided. In short, significant or dominant market share, lack of diversity, or the absence of redundancy/alternatives to an AI system or AI-controlled system can tip that use of AI into being high-risk.

Fourth, there may be some uncertainty around which 'adverse effects' are intended to be covered, and whether 'adverse effects' or adverse impacts *during training* and *development* are intended to be covered — or only impacts from use. When AI is trained on personal data, or copyright-protected content without authorisation or exception, is the AI module/system 'high-risk' owing to its 'adverse legal effects' or adverse impacts on human rights?[7]

The overall lifecycle approach adopted in the Proposals Paper implies that impacts during training are relevant. If so, there are questions regarding whether the Mandatory Guardrails are suited to identifying all possible risks.

For example, it is arguable, especially under current Australian law, that training and fine-tuning using almost any database of any size of material generated since the second half of the 20th Century risks infringement of copyright, even if licences are obtained from some copyright owners (as it will be, in many cases, impossible to ensure all copyright owners have been identified). We may not, however, wish to conclude that all AI systems trained without permission on copyright-protected content are high-risk, on the basis that questions relating to copyright protection and AI development ought to be dealt with separately from the Mandatory Guardrails.

[5]Canadian ADM Directive (n 4).
[6]NSW Artificial Intelligence Assessment Framework | Digital NSW' (Web Page) <https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assessment-framework#anchor-nsw-artificial-intelligence-assurance-approach>.
[7]IP rights are recognized as human rights under international human rights law: *Universal Declaration of Human Rights*, GA Res 217A (III), UN GAOR, UN Doc A/810 (10 December 1948) art 27 para 2; International Covenant on Economic, Social and Cultural Rights, opened for signature 16 December 1966, 993 UNTS 3 (entered into force 3 January 1976) art 15 para 1(c).

On the other hand, when it comes to **environmental risks**, every stage of the AI lifecycle creates environmental impacts.[8] The main stages include the extraction of resources to build new infrastructure such as data centres, servers, and undersea cables; the energy used to power data centres and water used to cool them; the deployment and application of AI in various facilities and on individual devices for more or less environmentally beneficial purposes; and the disposal and waste of infrastructure and devices.[9] While computation is generally becoming more efficient, the rapid overall expansion and growth in the training and use of AI means that AI's overall environmental impact is growing, creating potentially unacceptable climate and other environmental risks. For example, Sasha Luccioni of Hugging Face and colleagues have suggested that 'the amount of energy needed to power AI now outpaces what renewable energy sources can provide' in the United States.[10] In addition, larger models generally use more power and water to train and use,[11] and the impact depends on where computation is sourced.[12] Both AI developers and deployers need to be able to track and understand likely environmental impacts through the whole lifecycle in order to make an assessment as to whether the environmental benefits outweigh the risks, and how to minimise the overall impact.

Automated Decision-Making Systems

While the Proposals Paper is focused on AI, there are a wide range of 'high-risk' uses of automation that do not involve AI. Robodebt is a clear example, as is the UK's Post Office Horizon scandal. The factors defining high-risk systems were clearly present in the systems giving rise to both Robodebt, and the UK Post Office Scandal, given their adverse impacts on rights and economic interests.[13]

Applying Mandatory Guardrails only to AI could allow, and even incentivise organisations to avoid the Guardrails by developing and deploying non-AI automated systems. We would encourage the government to consider applying the Guardrails to high-risk AI and automated decision-making systems. This is particularly important in relation to *government* use of ADM systems. Recent ADM+S research has shown that use of ADM is widespread and increasing across government — and more common than the use of AI.[14] The Canadian ADM Directive provides a precedent for the application of similar risk-based guardrails to ADM systems in government.[15]

[8]Lynn H. Kaack *et al*. Aligning artificial intelligence with climate change mitigation (2022) 12 *Nature Climate Change* 518–527; Anne-Laure Ligozat *et al*, Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions (2022) 14(9) *Sustainability* 5172.
[9]Sasha Luccioni *et al*, The Environmental Impacts of AI--Primer (2024) Hugging Face; OECD, "Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint (November 2022).
[10]Luccioni *et al* (2022) (n 9).
[11]Gaël Varoquaux *et al*, Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI, (2024).
[12]Anne Pasek *et al*, The World Wide Web of Carbon: Toward a Relational Footprinting of Information and Communications Technology's Climate Impacts (2023) 10 *Big Data & Society*: 20539517231158994.
[13]It might be argued that the UK Post Office Horizon case, involving an accounting system, was not 'high risk' in itself, but the use of the system as the basis for prosecution for fraud and theft caused adverse impacts, which should be sufficient on any interpretation of what constitutes a 'high risk AI system'. Every AI system is not purely technical, but sociotechnical — it is the system *in context and in use* that creates risk. Nick Wallis, *The great post office scandal: The fight to expose a multimillion pound scandal which put innocent people in jail* (Bath Publishing Limited, 2021).
[14]Kimberlee Weatherall *et al*, 'Automated Decision-Making in New South Wales: Mapping and Analysis of the Use of ADM Systems by State and Local Governments' (Research Report, ADM+S, March 2024) ('ADM in Governments').
[15]Canadian ADM Directive (n 4).

Exemptions for Low Risk systems

An **exemption** to exclude unintentionally-captured low risk uses or systems is only required where (a) there is a list of high-risk systems or use; *and* (b) the regulation applies to all systems falling in the list categories: that is, there is no explicit 'second step' in the analysis, determining the level of the risk.[16] Article 3 of the ***Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law***[17] has no exemption, and does not need one, because the scope of covered systems in that Convention is defined at the level of principle, rather than with a list.[18]

If Australia adopts an indicative or prescribed list of high-risk systems, then an exemption may be appropriate. We note the Paper's reference to art 6 of the *EU AI Act*.[19] Art 6 of the EU AI Act states:

> Art 6(3) an AI system referred to in Annex III shall not be considered to be high-risk where it does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.
>
> The first subparagraph shall apply where any of the following conditions is fulfilled:
>
> (a) the AI system is intended to perform a narrow procedural task;
>
> (b) the AI system is intended to improve the result of a previously completed human activity;
>
> (c) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or
>
> (d) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

We have some concerns about this drafting. What constitutes a 'narrow procedural task' or 'preparatory task' is open to interpretation (and there will be incentives to interpret it broadly). Automated systems performing preparatory tasks are common in government, and can shape the information available to a decision-maker.[20] In Europe, further definition of these concepts will be undertaken by the Commission.[21] Similarly, if Australia adopts a list, with exemptions, then there will need to be a mechanism for continuous development of both the list, and the exemption. As we argue generally across this submission, the Australian legislation must allow an AI Body to issue disallowable instruments to add to, or refine drafting of any list; domain regulators or departments might also be tasked with providing guidance.[22]

---

[16] No exemption is needed as such in the Canadian ADM Directive or the NSW AI Assurance Framework, both of which require categorisation of an AI system as (some variation of) low, medium or high risk: see references above.
[17] COE Framework Convention (n 3).
[18] The Framework Convention applies to 'systems that have the potential to interfere with human rights, democracy and the rule of law' (art 3).
[19] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 art 6 ('EU AI Act')*.
[20] Weatherall *et al*, ADM in Governments (n 14).
[21] *See EU AI Act* (n 19) art 6(5), providing that 'The Commission shall, after consulting the European Artificial Intelligence Board (the 'Board'), and no later than 2 February 2026, provide guidelines specifying the practical implementation of this Article in line with Article 96 together with a comprehensive list of practical examples of use cases of AI systems that are high-risk and not high-risk.'
[22] Such questions cannot be left to the courts, because (a) it is unlikely there will be much litigation if individual remedies are not provided for in the legislation; (b) litigation takes too long in any event; *and* (c) the courts lack expertise in the relevant issues, and the means to access generalised expertise of a technical and socio-technical nature (court processes can access *specific* expertise on the *specific dispute* but not *general* on risks and impacts *generally*).

Question 3: List or Principles?

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

We see four potential approaches to defining high risk AI systems:

1. The 'principles' approach;
2. The 'principles' approach supplemented with examples (which could be in legislation, or (more updateable) guidance from an AI Body);
3. A **list** of high-risk uses in legislation or regulations with a mechanism for updating the list;
4. A list, supplemented by **principles** that identify high-risk uses, to be applied by the developer or deployer for cases for scenarios not already listed.

It will not be possible, in our view, to rely exclusively on principles: it would place all the onus — and confer all the power — to define high-risk uses on developers and deployers with incentives to apply them narrowly to reduce regulatory burden.[23]

If that is correct, then the Principles must be made concrete somehow, and the real question is how: a prescribed list, or examples/guidance.

In our view, there is sufficient evidence and international consensus, reflected in the EU list and the Canadian draft legislation, that a subset of uses are high-risk. A preliminary list at the time that legislation is introduced will help organisations. It will also increase public trust, as there will be an understanding that development and deployment of known high-risk systems are regulated in Australia.

The list should be *inclusive*, rather than exclusive, because we cannot identify all high risk uses in advance. Developers and deployers should also be obliged to consider the risks posed by systems not on the list, and provide for accountability, transparency and testing for other systems where non-listed systems pose similar levels of risk.

The question *then* becomes how the list should be mandated and updated. In our view, the list should be readily adapted by inclusion in disallowable instruments such as regulations rather than legislation, with updates managed by an AI Body (see further below in Topic 5). A primary list should be in disallowable instruments such as regulations, rather than 'guidance', so that the

[23]See Luke Munn, 'The Uselessness of AI Ethics' (2023) 3 *AI and Ethics* 869, 871-872. The incentive to apply the definition narrowly could in theory be reduced by applying penalties where the definition is not appropriately applied, but there is reason to doubt this would work. Significant penalties for getting it wrong are likely to increase pressure from industry and organisation to make the definition more specific; or to apply penalties only in the case of egregious or bad faith failures to apply the Guardrails, which would be difficult for regulators to enforce in court.

In addition, it should also be possible for specialist regulators or responsible government departments to publish their own additional guidelines or views on the kinds of systems that could be high-risk within their area of competence, in consultation with an AI Body.[24] For example, the Department of Health and Aged Care, the Therapeutic Goods Administration, and Fair Work Australia could have their own guidance on what AI systems they do and do not consider high risk.

Question 4: What AI systems, if any, should we ban?

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?

There are three questions to resolve here:

1. Are there systems which we can categorically say pose unacceptably high risk to Australians, Australian society, the economy and/or environment, and how can we articulate a definition of those systems?
2. Are such systems **already** prohibited by other Australian laws, and if so, does that mean a prohibition in AI regulation is unnecessary?
3. Is there a better alternative to prohibition that would still provide protection to Australians, Australian society, the economy and/or the environment?

We also point out the **critical interaction** between the question of **prohibitions** and Guardrail 2 (risk management). Especially if Australia chooses **not** to prohibit any AI systems, it will be critical to make clear that Guardrail 2 may require that an AI system not be used, where risks cannot be eliminated or reduced to a level that is acceptable. This is clear in the current drafting

Are there systems, or uses, that pose unacceptably high risk, and how could we articulate such prohibitions?

There are reasons to prohibit certain uses of AI systems. Article 5 of Europe's AI Act prohibits several applications of AI including:

- systems that use purposefully manipulative or deceptive techniques;
- systems that exploit vulnerabilities of individuals or groups to materially distort their behaviour in a manner that causes or is likely to cause significant harm;
- expansive 'social scoring' systems that lead to detrimental treatment in contexts unrelated to the context where scoring data was collected;
- systems used to predict risk of a person committing a crime based solely on profiling of personality traits and characteristics;

[24]We note, for example, that regulators in the US such as the Federal Trade Commission have a practice of issuing statements on the kinds of business models or activities which they consider breach various consumer protection laws: see, most recently for example, Federal Trade Commission, 'FTC Announces Crackdown on Deceptive AI Claims and Schemes' (Press Release, 25 September 2024). For a list of FTC activity to protect consumers from harmful AI use, see <https://www.ftc.gov/news-events/news/press-releases/2024/07/ftc-submits-comment-fcc-work-protect-consumers-potential-harmful-effects-ai>.

- systems that conduct untargeted facial image scraping from the internet;
- systems that infer emotions in the workplace or educational institutions (except when used for medical or safety reasons);
- biometric categorisation systems that infer sensitive personal information from biometric data; and
- real time biometric ID systems except in highly limited and regulated circumstances.

It is unclear why any of these systems should be allowed in Australia. Without a ban, legitimate questions may be asked by Australians, as to why the government considers that Australians deserve less protection from obviously egregious systems than their counterparts in the EU.

There may not be a consensus among all ADM+S researchers regarding a set of unacceptably high risk systems. The following, however, have been suggested in our internal consultations:

1. Any systems which exploits, for profit or other advantage, human rights-related vulnerabilities (that is, vulnerabilities or protected categories recognised under Australian anti-discrimination laws, such as age, disability, gender etc);
2. Any systems purposefully used to create 'deepfake' images, video or other content that can be used for sexual exploitation or shaming, including 'fake' child exploitation material.
3. Systems that produce psycho-social, physical or economic harm for workers.

A prohibition like the second of these, based on the *purpose* of the system, may be avoided by organisations noting such uses are not intended. In such a case, application of the Guardrails across the lifecycle would still require the developer to take steps to mitigate or prevent ongoing use for the prohibited purpose and its adverse impacts.

We draw particular attention to the need for a prohibition on the indiscriminate use in public places of facial or other biometric recognition and surveillance technologies, by policing, law enforcement or other public authorities. As in Europe, we would suggest that only very limited, targeted and regulated uses with judicial authorisation processes be allowed.[25]

Any list of prohibitions should be enacted in binding (disallowable) form with mechanisms for update and refinement (as discussed above in relation to Question 3).

We further suggest that whatever remedial options are considered for other parts of the AI regulations proposed, breach of a prohibition on unacceptably high-risk systems should be subject to criminal and civil prosecution (the latter with civil penalties available), and regulator powers to order immediate withdrawal of technology from the Australian market.

[25]See, for example, *EU AI Act* (n 19) art 5. See discussion in R Matulionyte, M Zalnieriute (eds), The Cambridge Handbook of Facial Recognition in the Modern State (Cambridge University Press, 2024).

Where should prohibitions be found (in AI legislation or general law)?

Some uses of AI systems that we or others may suggest for prohibition are arguably already prohibited by existing Australian law, or may be soon. Discrimination based on protected characteristics is already prohibited; the first proposed prohibition above, for example, would also arguably fall within a general prohibition on unfair trading practices which has been proposed to be included in the *Australian Consumer Law*.[26]

We nevertheless recommend a prohibition list even where other prohibitions arguably apply, for the following reasons:

- **Clarity**: we do not want to have to wait for case law to be clear that prohibitions in general law, not specifically drafted to mention AI, nevertheless apply;
- **Responsibility allocation**: we want developers and deployers to take proactive responsibility for ensuring systems are not designed or deployed in ways that pose unacceptably high risk, and do not want to have to wait for case law in which multiple parties in the AI supply chain fight for years over who is responsible for what;
- **Signalling**: to signal government disapproval and protection for Australians;
- **Findability**: to ensure prohibitions are readily identified by developers and deployers — prohibitions dotted throughout Australian legislation and case law, and not drafted by reference to technology but behaviour or actions are not easy to identify.
- **Remedies/enforcement**: to ensure that remedies or enforcement options (such as immediate withdrawal from the Australian market, at the order of an AI Body) follow from breach.

Clearance or pre-authorisation as an alternative to prohibition

We note that if the government chooses not to *prohibit* certain technologies or uses, an alternative would be to create a clearance or pre-authorisation system for certain systems. A list of systems required to be cleared in advance by an appropriate authority (see Topic 5 below) could potentially be longer than a list of prohibitions.[27]

[26] The Treasury, *Protecting Consumers from Unfair Trading Practices* (Consultation Regulation Impact Statement, August 2023).

[27] For example, consistent with proposals from the Human Technology Institute, indiscriminate use of facial recognition systems by public or private sector actors could be subject to pre-authorisation (or preferably, biometric identification systems). As noted in research by Choice, the use of facial recognition for example by private retail is increasingly pervasive in Australia — and highly intrusive. See Nick Davis, Lauren Perry and Ed Santow, *Facial Recognition Technology: Toward a Model Law* (HTI, 2022).

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

Question 8: The Mandatory Guardrails (drafting)

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?

General comments

On the whole, the guardrails look logical and comprehensive.

We note that the Mandatory Guardrails are not a complete answer for addressing potentially harmful uses of AI. They set out *how* AI should be developed and deployed: *not what it should be used for*.[28] They establish requirements such as risk management, accountability and human oversight; testing and monitoring; data governance; transparency (of various kinds); contestability; record-keeping; and conformity assessment.

The Guardrails do not impose substantive standards for safe and responsible AI. Mandatory Guardrails do not define what counts as illegal discrimination using a system: we would hope that they help reduce bias and discrimination by requiring developers and deployers to identify risks of discrimination, mitigate the risk, and test and monitor for it.

Consequences of breach

As we understand it, the Guardrails do not create liability if harm arises where the Guardrails are breached. A question that has come up repeatedly in our internal consultations is what **consequences** are intended to flow from a failure to comply with the Mandatory Guardrails:

- Is it purely regulatory (for example, a fine)? Or,
- Could breach lead to an obligation to compensate for harm caused?

Our preliminary view is that liability for harm should be considered separately: connecting the Guardrails to compensation for harm would have the implication that they would need to be more narrowly and clearly defined.

It is possible that the Guardrails could have indirect effects on liability, such as influencing standards of reasonable care (in negligence);[29] or providing records (via Guardrail 10) that could be relied on in evidence;[30] failure to implement the Guardrails might be relevant in considering whether company directors are fulfilling their duties to the company.

[28]Jacqueline McIlroy, Sara Luck and Henry Fraser, 'Decoding Canada's Directive on Automated Decision-Making', *Medium* 24 May 2024) <https://medium.com/automated-decision-making-and-society/decoding-canadas-directive-on-automated-decision-making-08124bcdf250>.
[29]Henry Fraser and Nicolas P. Suzor, 'Locating liability for AI harms: A systems theory of reasonable foreseeability, control and fault in the AI value chain' (*Law Innovation and Technology*, Forthcoming Autumn 2025) <https://eprints.qut.edu.au/251116/>.
[30]Henry Fraser, Rhyle Simcock and Aaron J Snoswell, 'AI Opacity and Explainability in Tort Litigation', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) <https://dl.acm.org/doi/10.1145/3531146.3533084>, p 190.

We do not think all the liability questions need to be resolved before enacting Mandatory Guardrails. But our view is **that the government should also take active steps to address the question of liability for AI-generated harms**. There are ongoing developments in this area overseas which could be considered, such as in Europe where there have been amendments to the Product Liability Directive and a proposed AI Liability Act.[31] Detailed consideration by a body such as the Australian Law Reform Commission could be helpful in guiding organisations, lawyers and the courts.

Drafting and mechanisms for updating and providing guidance

Much will depend on how much of the *text describing* each Guardrail becomes part of the law or *standards applying the law*. For example, Guardrail 3 states that organisations should protect AI systems, and implement data governance measures to manage data quality and provenance. The text *describing* the Guardrail states that:

- organisations must ensure they have appropriate data governance, privacy and cybersecurity measures in place
- data used to train, fine-tune or test a model must be fit for purpose and representative
- data must be legally obtained, and must not contain illegal and harmful material
- data sources must be disclosed

To operate horizontally (see Topic 5 below) the Guardrails will need to have flexibility: in that deployers and developers would be expected to implement testing and risk mitigation proportionate to the nature and severity of any risks of harm arising from development or deployment of an AI system; different tests and mitigations will be appropriate for different kinds of identified risks of harm.

Flexibility will also be important, because some aspects of the Guardrails would need to be nuanced in practice, as our discussion of some individual Guardrails below illustrates. See, for example, the discussion of Guardrail 6, and our argument that watermarking content as AI-generated should not always be required.

[31]See: European Commission, *Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and Robotics*, Com/2020/64 final; *Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence*, COM/2022/496 final; *Proposal for a directive on liability for defective products* (text adopted March 2024).

We expect there will be an ongoing need to update the Guardrails, through a combination of:

- Industry standards, or guidance from industry bodies where guidance specific to particular domains is warranted
- Disallowable instruments where general updates are required in response to new technologies or better understanding.

An AI Body (see Topic 5 below) could provide ongoing guidance, based on scanning of industry practice, research, and engagement with experts on harms caused by AI systems that need to be mitigated by organisations, and how to mitigate them. Any such taxonomy of harms and mitigations will have to be constantly updated and so should not be part of legislation. A key challenge for any such body will be to ensure that it has both expertise in the technical aspects of AI and machine learning, and their legal and social impacts; and to ensure that stakeholder groups, including marginalised groups, are appropriately represented.[32] We provide further comment below highlighting areas where we can already foresee the need for flexibility and/or updating. We do not comment on every Guardrail.

Mandatory Guardrail 2: Risk management and mitigation

A critical question for Guardrail 2 is whether a risk acceptability criterion is built in. Article 9(5) of Europe's AI Act requires risks to be judged 'acceptable'. In other words, under the European AI Act, risk management obligations are not met until the risk is reduced to a certain level, according to a certain set of criteria. In Europe, the risk acceptability criterion for high-risk systems is 'as far as technically feasible' — a very exacting standard. Other risk management standards (the Proposals Paper contemplates implementing Guardrail 2 through standards) leave it to their users to pick their own risk acceptability criterion, based on their regulatory, reputational and commercial situation. In the Canadian Directive, by contrast, there is no such requirement.

The risk acceptability criterion chosen, or indeed the choice not to specify a risk acceptability criterion, may significantly impact on the lengths to which developers and deployers must go to manage risks,[33] and it is worth giving consideration to whether risk acceptability should be explicitly addressed in the Guardrails or related guidance.

Risk acceptability may also tie into Guardrail 10, conformity assessment — meeting a particular acceptability criterion is a relatively straightforward condition against which conformity could be judged.

Another question the government may wish to consider is based on the **NSW AI Assessment Framework**. That Framework asks government entities to consider whether the proposed use of AI is an improvement on the existing system/activity.

[32]Henry Fraser, Christine Parker, Fiona Haines, José-Miguel Bello y Villarino and Kimberlee Weatherall, 'Should Australia follow Europe's approach to AI standards and regulation?' (Under review with *ANU Journal of Law and Technology*), working paper available at <https://eprints.qut.edu.au/251557/>.
[33]Henry Fraser and José-Miguel Bello y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough' [2023] *European Journal of Risk Regulation 1*.

It strikes us that organisations ought to be asking themselves whether the benefits from using AI outweigh the potential risks of doing so. For example, organisations should consider the benefits against environmental costs. A concept of this kind may need to be referenced in legislation or rules, because we cannot be sure that the market will ensure that the environment impacts will be costed in, given the extraordinary funding that is going into developing and deploying AI, and the 'race to domination' occurring between larger entities. The Government of France has adopted as one of its principles for AI innovation the concept of 'frugal' or 'sufficient' AI by which they refer to a range of ways to promote 'moderation in the production and consumption of energy and material resources'.[34]

The relationship between Mandatory Guardrails 2, 4 and 8: Risk management and mitigation, Testing and Monitoring and Transparency across the AI supply chain

As the Proposals Paper indicates, contextual factors will play a significant role in determining how responsibility will be apportioned across the AI supply chain and throughout the AI lifecycle. Many AI risks, including risks from general purpose models and systems only become clear once the context of use and the conditions and capabilities of users are known or at least anticipated.[35] It follows that responsibilities should be distributed according to which actors are best equipped to address risks associated with a particular stage of development. But there are also other implications.

Firstly, supply chain participants with an understanding of the context of use of a system will generally be best placed to foresee and manage risks associated with that context, test and monitor for those risks on an ongoing basis, and communicate risks to other parties in the supply chain. The distinction between deployer and developer may not be as important, in this respect, as the distinction between 'upstream' actors with little knowledge of the context of deployment, and 'downstream' actors with richer knowledge.

Secondly, risk assessments may need to specifically state, and be tailored to, a contemplated context of use, or set of contexts. Deployers and developers will also likely seek procedures and clear processes for updating risk assessment as new contexts of use are discovered or developed. For example, there will be a need for clear guidance about how testing and monitoring at one point in the supply chain should feed into documentation and transparency requirements, and how these in turn should feed further risk management processes will likely be very much in demand.

Thirdly, the Guardrails cannot work separately from each other. Since some risks only become apparent in specific contexts of use, developers and deployers will likely need guidance on the relationship between Guardrails 2, 4 and 8.

[34]AFNOR SPEC 2314, "General Framework for frugal AI: An AFNOR SPEC to measure and reduce the environmental impact of AI", (June 2024), Ministere de la Transition Ecologique et de law cohesion des territoires.
[35]Henry Fraser and Nicolas P. Suzor, 'Locating liability for AI harms: A systems theory of reasonable foreseeability, control and fault in the AI value chain' (*Law Innovation and Technology*, Forthcoming Autumn 2025) <https://eprints.qut.edu.au/251116/>.

Fourthly and relatedly, effective monitoring and lifecycle risk management is only possible through the cooperation and coordination of multiple actors in the value chain. Transparency (under Guardrail 8) is a necessary but not a sufficient condition for this kind of coordination. Developers and deployers all across the supply chain not only need effective (and cost-effective) tools and processes for finding out important matters about what is happening upstream and downstream of them, but more importantly, processes for working together to fix problems that none of them can fix in isolation.

Supply chain participants are likely to demand more clarity about *how to respond* to information they receive through monitoring, transparency and documentation. General purpose AI providers will, for example, need clarity about *when they will be deemed to be on notice of a risk or harm* in a downstream application of their model, and when notice of a harm tips over into responsibility to address a harm or assist in addressing it. Will duties to monitor — especially duties for upstream providers — be active duties, or passive? Generally, online platforms have been very resistant to any kind of responsibility to actively monitor for risk, although the European Digital Services Act[36] and Australia's eSafety regulations have begun to impose more active duties. For the most part, in internet regulation, responsibility sharing for matters such as copyright infringement or harmful content has settled into 'safe harbour' or 'notice and takedown' regimes, where upstream actors are only required to take action in cases where they have detailed and specific notice of clearly scoped problems. Another possible model of coordination in supply chains is the 'bug bounty' — where upstream actors provide incentives for downstream actors and third parties to discover and report software vulnerabilities. In any event, uniform and standardised reporting mechanisms, and clear thresholds for when responsibility passes between supply chain participants, or when supply chain participants are required to work together on a problem, will likely be necessary. These can be provided via guidance, rather than in any legislative framework.

These issues have particular importance in the context of government use and procurement of AI. A number of recent tribunal decisions have shown how software supply chains and software-as-a-service procurements of automated systems frustrate transparency obligations under Freedom of Information.[37] Many existing transparency obligations as well as proposed audit obligations apply to the entity involved in decision-making, service delivery, or other forms of conduct that affect individuals. However, software-as-a-service means that the organisational location of service delivery is not always the organisational location where the computational (and automated) contributions to decisions and service provisions primarily occur. The cases noted above demonstrate the need to ensure that government entities providing services or making decisions are able to access and have control over the documentation necessary to fulfil transparency and audit requirements, despite those materials being controlled by third parties and potentially including protected commercial information and trade secrets. This current problem also speaks to what accountability mechanisms under Guardrail 1 might require — distributions of transparency obligations through procurement and contracting processes.

[36] *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC* ('Digital Services Act').
[37] *'EC3' and Department of Jobs, Precincts and Regions (Freedom of Information) [2022] VICmr 47 (27 June 2022); O'Brien v Secretary, Department Communities and Justice [2022] NSWCATAD 100.*

Mandatory Guardrail 3: Data quality and provenance

We agree that **data quality** is critical: data must be high quality and must measure what is purported to be measured. The 2016 Alleghany County Department of Human Service's Family Screening Tool, in Pennsylvania provides a useful illustration.[38] AI was used to calculate a predictive risk score for families, and guide social workers deciding whether to investigate a family for child abuse or neglect. However, the algorithm used data on the number of times a family had been *reported*, not the number of times there had been a *finding* of child abuse or neglect; data on the outcomes of reports was not included. As a result, economically marginalised non-white families, who were disproportionately reported to child protection services, were given a higher risk score regardless of the outcome of previous reports. In other words, the algorithm did not calculate the risk of child abuse or neglect, but the risk of being reported for child abuse or neglect. More broadly, the quality and accuracy of administrative data is recognised as a longstanding challenge.[39]

A requirement that data be **legally obtained** is complex, and may be impossible to comply with, given how many different laws may apply to the collection and use of data: privacy law, copyright law, other IP; confidentiality obligations; defamation; national security laws; and more. Australian copyright law could make some AI training impossible (or impossible for all but the largest players, including social media and other entities which have already collected extensive data from Australian users as allowed by Australia's inadequate privacy protections). Australia's repeated failure to update copyright exceptions means that, unlike many similar jurisdictions, it does not have any exceptions to allow for text and data mining, or fair use, meaning it is at least arguable that a licence for **all content** is required where copyright content is copied in Australia, by developer or deployer, for the purposes of training, fine-tuning or retrieval-augmented generation. It is doubtful that AI developers can determine and claim whether the dataset is legal in most cases due, to variety of data scraped from various sources: even were a licence obtained, would the developer (or even the data provider) be able to guarantee there were no infringements: of copyright law; or defamation; or privacy?[40] It is also unclear what information developers or deployers would or could provide to support their claims regarding the legality of data.

[38]Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin Press, 2017) contains further examples.

[39]Philip Gillingham, "The development of algorithmically based decision-making systems in children's protective services: Is administrative data good enough?." The British Journal of Social Work 50.2 (2020): 565-580; Aileen Rothbard, "Quality issues in the use of administrative data records" in Actionable intelligence: using integrated data systems to achieve a more effective, efficient, and ethical government. New York: Palgrave Macmillan US, (2015). 77-103.

[40]For more detail on why relying on licensing alone is insufficient, see Martin Senftleben, Win-win: How to Remove Copyright Obstacles to AI Training While Ensuring Author Remuneration (and Why the European AI Act Fails to Do the Magic) (July 04, 2024) Chicago-Kent Law Review, Volume 98 (2024) forthcoming.

A further question is whether 'legally obtained' should be assessed with reference only to Australian law. If a training dataset is compiled overseas, would the Guardrails require that an AI developer or deployer ensure the creation of the dataset complied with every relevant law of the country where the individual or entity was residing at the time of compilation, or a place of servers where the data was stored?[41]

In short: a requirement that data be legally obtained, while superficially attractive, may be impossible to comply with in practice, especially for Australian companies who, unlike companies overseas, cannot afford to ignore Australian law. It could also lead to the Mandatory Guardrails being used as a tool for achieving legal goals unrelated to responsible and safe use of AI. This is not to say that we should entirely ignore legal breach involved in the creation of AI, but rather, to say that the Mandatory Guardrails may not be the place to solve those problems, which may require global, and coordinated, rather than purely Australian resolution.[42]

One question is whether **data sources** should also be **disclosed**. Disclosure of data sources may be important as a step towards accountability, enabling the public to monitor potential sources of harm, in particular in relation to systems used by government.[43] Much work has been done in this space to develop mechanisms for disclosure that can be referred to.[44]

We note on the other hand however that there may be reasons for non-disclosure. In practice, it is rare for datasets to simply be downloaded and then used for training. Dataset pre-processing before training can be complex and time-consuming, and can legitimately involve engineering knowhow that might transform a public dataset into something of legitimate proprietary value.[45] The specific ways different data sources are combined, weighted, filtered, cleaned, arranged, and used in model training are examples of legitimate sources of proprietary 'know how' that companies should have a right to protect.[46]

[41]As a general rule, copyright law is territorial, the law of the place where the activity took place (the copying) is the one which applies. Copyright law, however, might envisage an exception and require that, wherever the data compilation took place, it should comply with Australian laws. Such requirements are not unknown in Australian copyright law, but such an approach could preclude Australian companies from activities that are allowed in many equivalent jurisdictions to our own, where copyright law has been amended to create text and data mining exceptions or fair use, including Japan, Singapore, Germany, and the US.

[42]As a further point in support of the argument that laws relating to Mandatory Guardrails are not well-suited to address copyright concerns, we note that problems of copyright infringement are not confined to high-risk AI modules/systems (unless the concept of 'high risk' is broadly read to include any risk to any legal right; see discussion earlier in this submission). For example, there are already AI systems that were trained on musical works and only produce music; they are not high risk but information about their training datasets is certainly relevant to right holders intending to exercise their rights.

[43]See further a report by ADM+S researchers: S Maitra, L. Sleep, S. Fey, and P. Henman (2024) *Building a Trauma-Informed Algorithmic Assessment Toolkit*. ADM+S Working Paper Series. DOI:10.60836/f01c-4a18 ('ADM+S Toolkit')

[44]See Margaret Mitchell et al, 'Model Cards for Model Reporting' [2019] *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220; Timnit Gebru et al, 'Datasheets for Datasets' <https://arxiv.org/abs/1803.09010v8>; and the collected sources available via HuggingFace.

[45]For example, the pre-processing of an existing dataset can sometimes be so involved that the pre-processed data are actually published as a new academic output, as has occurred several times with the ImageNet dataset (see, e.g. Tal Ridnik et al, 'Imagenet-21k pretraining for the masses' arXiv preprint arXiv:2104.10972 (2021)).

[46]For a recent illustration of the depth of data pre-processing that goes into several contemporary foundation models, see e.g. <https://magazine.sebastianraschka.com/p/new-llm-pre-training-and-post-training>.

A requirement to publish all metadata, such as titles of works and their authors/right holders, would not be possible in many cases since much of the data will not have this information. As an alternative to full disclosure, perhaps AI developers might be required, upon request, to disclose whether specific items were contained in the datasets where reasonably possible.[47]

In summary: the nature of any disclosure required will likely require elaboration, which will need to be informed by both the *purpose* of disclosure, and the extent of disclosure required to meet that purpose, and the potential broader social and legal *implications* of different kinds of disclosure. Again, this would suggest that legislated guardrails should be carefully drafted and allow for flexibility; further disallowable rules and guidance, which can be updated, will be needed to address the kinds of issues identified here.

Mandatory Guardrail 4: Testing and monitoring

We have commented above on the relationship between Guardrails 2, 4 and 8.

Mandatory Guardrail 5: Human oversight and intervention

Risk-based regulatory systems typically include requirements for human oversight in cases of high risk systems. These new oversight requirements are designed differently than prior requirements for a final-human decision maker, rights to contestation or obligations for 'meaningful' human control of a system.

We note that:

1. Human oversight requirements are different for system developers and system users. Developers must integrate human oversight into their designs, paying attention to organisational and cognitive realities like the risk of automation bias, and create information and instructions on how human oversight must be carried out by system users. System users and deployers are obliged to follow those instructions, as well as assign oversight to natural persons with sufficient competence, authority, and support to perform that role.
2. Human oversight does not need to be performed by the system user themselves. There is no requirement that system deployers actually perform oversight, only that they assign the oversight task to a natural person and create conditions necessary for them to fulfil their role.
3. How oversight should be performed is only vaguely specified, with forthcoming standards likely to be important for offering guidance on compliance with regulations.
4. Human oversight obligations interact strongly with questions of liability. In the EU, where human oversight is one of the 'guardrails' for high-risk systems, the related revised Product Liability Directive and AI Liability Directive contemplate how human oversight requirements might participate in the determination of both strict product liability and fault-based liability.

[47]For example, havelbeentrained.com app allows right holders to discover whether their images were in the LIAON training dataset that was used to train Stable Diffusion model; automated tools are already provided by online platforms such as YouTube to enable IP owners to find infringing content on their platforms.

When including human oversight as a guardrail, policy makers must pay attention to the type of human oversight requirements desired. Is this guardrail intended to ensure a human decision-maker, or — like the EU AI Act — is it more about establishing the parameters for a safe AI product that has a human somewhere in the workflow? Are policy makers content to allow human oversight to be outsourced and performed under gig-work like conditions coordinated by software development platforms? Are policy makers considering how human oversight obligations interact with questions of liability? There is a need to avoid overly responsibilising human overseers if the structural conditions of oversight mitigate against their having actual control or influence over outcomes.

Again, this would suggest that legislated guardrails should be carefully drafted and allow for flexibility; further disallowable rules and guidance, which can be updated, will be needed to address the kinds of issues identified here.

Mandatory Guardrail 6: Transparency to end users/impacted individuals

The design of transparency/disclosure systems is important and difficult, for a number of reasons.

First, **notification fatigue**, and the limits of notification need to be considered. Informing people again and again, in an undifferentiated way, that 'AI is being used' is likely to lead to fatigue and an inability on the part of end users to differentiate between notifications that matter — and those that don't.[48]

Second, any AI-generated content labelling, disclosure or watermarking regime needs to make sense with respect to systems that may emerge in the near future, including AI agents. There are differences between labelling a deepfake video of a politician's speech as AI-generated, and seeking to label a series of digital interactions as AI-generated ('my AI receptionist was actually the one emailing back-and-forth with you to you to organise that purchase order'). Rules regarding disclosure will need to be drafted in sufficiently broad terms, and be able to be updated to address agentic applications of Generative AI — with detail perhaps provided in further guidance.

Third, in some contexts, simply *informing* end users regarding the use of AI may be insufficient to mitigate the risk of harm. One example is in the area of social services delivery. Social services users are increasingly being required to engage with AI to receive essential support,

[48]See, e.g. the analogous and well-documented case of alarm fatigue in healthcare settings; Cvach, Maria. "Monitor alarm fatigue: an integrative review." Biomedical instrumentation & technology 46.4 (2012): 268-277. Or, the analogous case of notification fatigue from GDPR 'cookie consent' banners on the internet; P. A. J. Graßl *et al*, 'Dark and bright patterns in cookie consent requests' (2021).

chatbots being a notable example.[49] A choice not to engage with AI may in some cases mean forgoing essential services. This can be problematic in a context where Australians may fear automated and AI services after experiences with Robodebt or lack affordable access to digital

Fourth, flexibility around disclosure may be important, because sometimes disclosure ought not be required. Guardrail 6(3) states that: 'Organisations must apply best efforts to ensure AI-generated outputs, including synthetic text, image, audio or video content, can be detected as artificially generated or manipulated'. But in our view, this requirement should not apply unconditionally or in all cases. For example, when AI is only assisting in the creation of work or editing it, or when only one part of a work (film) is generated by AI (eg special effects), organisations should not be obliged to make best efforts to ensure that this AI-generated content is detectable as AI generated or manipulated. This conclusion is also supported by survey results which show that users expect no/limited disclosure when AI contribution is minimum.[51] We note that if AI is used only for minor editing, this may fall outside the category of high-risk AI, depending on the approach to defining high risk use cases and any exemptions.

Alternatively, the expectation could be that, for generated content, any disclosure, labelling or watermarking regime should be applied in situations where an ordinary person might reasonably assume that AI had not been used in the creation of the content. This could be difficult to apply generally: in many cases people cannot identify which content has been generated by AI,[52] and it may not be clear whether they are likely to assume AI is or is not used (and indeed that assumption or expectation could shift, potentially rapidly). An alternative might be that there are certain contexts (say, news media) where a higher expectation of disclosure applies (such as a publicly available policy that explains how AI editing is used, together with some watermarking or labelling). A distinction might also be drawn (with difficulty) between AI-*edited* and AI-*generated* content. No Guardrail is likely to be drafted at anything like this level of detail: we make these points to illustrate the importance of not drafting Guardrails in absolute or unqualified terms.

Fifth, a notification or disclosure regime needs to clarify the purpose(s) of notification in different contexts, and be mindful of different ways to achieve the stated goals. For instance, if the objective is to make sure AI-generated data can be automatically detected and filtered

[49]Previous research by ADM+S researchers demonstrates that a common government use of AI and automated systems is to enable self-service by social services users. For some end-users, this has positive impacts, but this will not necessarily be true for everyone. See Kimberlee Weatherall et al, ADM in Governments (n 14); Lyndal Sleep, Brooke Ann Coco and Paul Henman, 'Mapping ADM in Australian Social Services' (Working Paper, ADM+S, October 2022) 92 ('ADM in Social Services'). The latter study found that of the 28 systems investigated, 67% involved self-service. The report also found that accessing services was the most common use of AI in Commonwealth, State/territory and non-state agencies that delivered social services: at 92. The former study (Weatherall et al) found a wide range of uses of automated decision-making systems, but also supports the conclusion that governments are making broad use of automated systems for the purpose of enabling self service by citizens and residents. Many examples may be found in the Compendium of AI systems published by the NSW Ombudsman and building on the research in Weatherall et al: see New South Wales Ombudsman, 'Compendium of ADM Systems' in 'A Map of Automated Decision-Making in the NSW Public Sector' (Special Report under s 31 Ombudsman Act 1974, March 2024).

[50]For information about access, see ADM+S research, the Australian Digital Inclusion Index.

[51]Paul Formosa, Sarah Bankins, Rita Matulionyte, Omid Ghasemi, 'Can ChatGPT be an author? A Mixed Methods Study of Perceptions of Generative AI creative writing assistance on authorship, creatorship, responsibility, and disclosure', *AI & Society: Knowledge, Culture and Communication* (forthcoming 2024), unpublished version available at: <https://link.springer.com/article/10.1007/s00146-024-02081-0?utm_source=rct_congratemailt&utm_medium=email&utm_campaign=oa_20240927&utm_content=10.1007%2Fs00146-024-02081-0>.

[52]Frank, Joel, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. "A representative study on human detection of artificially generated media across countries." In 2024 IEEE Symposium on Security and Privacy (SP), pp. 55-73. IEEE, 2024.

out from other types of online data in the future to prevent model collapse,[53] watermarks that are invisible to human users are appropriate. If on the other hand, a purpose is to allow human users to exercise discretion regarding information sources when interacting with content online, then any label or notification scheme would need to be perceptually detectible.

It is necessary to be explicit in guidance about both the level of the expectation and the reasons why disclosure is needed, so that organisations applying the Guardrails can apply them appropriately.[54]

Sixth, increasingly AI is being used in classification and triage systems, sometimes based on risk assessment and profiling. Such systems may be viewed as mundane and commonsense and can occur upstream from human made decisions, such as in risk profiling of child protection, parole, or long-term unemployment.[55] Such systems are often quite invisible to people subjected to these assessments, and yet the decisions arising directly or indirectly can be quite profound, such as loss of autonomy or rights. Notification and disclosure regimes need to take account of these often indirect, yet high-stakes uses.

Mandatory Guardrail 7: Contestability (challenging outcomes)

Contestability is important. Regulation of AI can take essentially two forms. First, a risk based framework, as the one suggested in the Proposals paper. This would normally involve an ex-ante clearance for higher-risk systems, and separate institutional and operational mechanisms to assess that risk. The second alternative is a legal framework that relies on the affected parties to complain about the effects of those systems once deployed, using those processes to create the right incentives for developers and deployers to perform adequate assessments before using those systems. Contestability is essential for the latter, but, in theory, not for the former. A proper ex-ante assessment, accompanied by adequate monitoring should have considered cases that require contestation and evaluated their consequences in the process of assessing the acceptability of the system.

However, in reality, there are many instances when AI systems will produce outcomes that still cause harm. There must be mechanisms to contest those outcomes and explore if they have been adequately addressed in the assessment process and whether they are part of their normal operation. Contestability will also allow organisations to refine their decisions regarding whether the mitigations of risk are sufficient to allow for deployment of a system. At least for public sector systems, contestability could be facilitated as part of a monitoring process. As discussed below (as a potential additional guardrail), civil society may play an important role in monitoring AI systems; in considering contestability it would be helpful to ensure that civil society has some access to mechanisms to contest systems or their results.

[53]Model collapse refers to the scenario where future AI systems get less intelligent due to a large fraction of training data being generated by other AI systems, not by humans. See, e.g. <https://theconversation.com/what-is-model-collapse-an-expert-explains-the-rumours-about-an-impending-ai-doom-23641>.

[54]This may be another situation where, as discussed elsewhere in this submission, it will be appropriate for domain-specific regulators or government agencies to provide additional guidance: ACMA for example could have a role in relation to news media expectations.

[55]Sam Desiere, Kristine Langenbucher, and Ludo Struyven, 'Statistical profiling in public employment services: An international comparison' (2019) OECD <https://www.emerald.com/insight/content/doi/10.1108/17538351111143330/full/html>; Sarah Sacher, 'Risking children: The implications of predictive risk analytics across child protection and policing for vulnerable and marginalized children' Human Rights Law Review (2022) 22(1): ngab028. Gijs Van Dijck, 'Predicting recidivism risk meets AI act' European Journal on Criminal Policy and Research (2022) 28(3) 407–423.

We note that it is important that contestability is not inextricably linked to liability and compensation. For example, contestation by interested parties of outcomes generated by AI systems deployed in the public sector will be in the public interest to ensure an adequate functioning of the systems even if those negatives outcomes were already anticipated in the initial assessment and considered acceptable, in a similar manner that some medical devices cause harm to some people, it is important to contest it and understand what happened even if it may not involve compensation, if it as a known risk that was already anticipated and explained.

Mandatory Guardrail 8: Transparency across the supply chain

We have commented on Guardrail 8 above, in considering its relationship to Guardrails 2 and 4.

One additional area in which AI transparency across the supply chain will be of increasing importance is the environmental impact of the various components of the AI lifecycle. Market expectations and mandatory requirements that all commercial and public organisations fully report their carbon emissions, biodiversity impacts and other environmental impacts and act on plans to minimise these impacts will continue to grow. This will create increasing demand to be able to track and calculate the carbon and other environmental impacts of the AI supply chains embedded in the AI applications deployed by businesses and public agencies. It is therefore worth requiring larger higher impact model developers and deployers to build in ways to track and make transparent carbon emissions and other environmental impacts of their AI systems. This is broadly recognised in the EU AI Act, and standards to be developed to support the EU AI Act. The EU AI Act requires high-risk AI systems to conduct risk assessments and create technical documentation processes that address their impact on 'health, safety [and] fundamental rights', a term which is defined to include 'environmental protection' (see EU AI Act Articles 1, 9). To assist this, the Act requires the European Commission to request the development of standards for all requirements for high-risk and general purpose AI models and specifically 'on reporting and documentation processes to improve AI systems' resource performance, such as reducing the high-risk AI system's consumption of energy and of other resources during its lifecycle, and on the energy-efficient development of general-purpose AI models' (EU AI Act Article 40(2); see also article 112(6)). In addition the EU AI Act encourages the development and adoption of voluntary codes of conduct for other (non high-risk) AI systems for various purposes including 'assessing and minimising the impact of AI systems on environmental sustainability, including as regards energy-efficient programming and techniques for the efficient design, training and use of AI' (EU AI Act Articles 95(2)(b); see also article 112(7)).[56]

[56]See Philipp Hacker, 'Sustainable AI Regulation' (2024) Common Market Law Review 2(61) pp. 345-386.

Mandatory Guardrail 9: Record-keeping

Guardrail 9 is an essential element for enabling and locating accountability and responsibility. Currently, the wording is quite broad, and could benefit from additional guidance to illustrate the types of records that could be in scope.

Decisions about deployment, including authorising actors, provide the basis by which individuals within organisations can be held motivated to consider the requirements for consideration and assessment of Guardrail compliance.

The results of testing provide evidence of the extent to which problems may have been foreseen and assessed or not. When problems arise, evidence of testing can highlight whether the issues were not considered or were identified and ignored. Testing also relates to the results of AI audits and risk assessments undertaken. There is a rapidly growing body of AI audit tools and frameworks to support organisations in ethical, responsible and safe use of AI,[57] including for example an Trauma-Informed Algorithmic Assessment Toolkit developed by ADM+S researchers.[58]

In providing guidance on this Guardrail, we should also consider the minimal amount of publicly available information about record keeping for Guardrail compliance, such as a list of types of records. In the EU, for example, there is a growing practice of publicly reporting outcomes of AI bias testing to demonstrate that outputs of AI are not biased by gender or ethnicity.[59]

Mandatory Guardrail 10: Conformity assessment

Consideration could be given to extending this draft Guardrail to also include making allowance for independent third parties, such as researchers, to assess the AI model or tool.

The history of social media platforms provides an illustration. In early days, platforms such as Facebook, Twitter and YouTube built their services through open web standards which facilitated interoperability, user choice, and opportunities for third-party development. In parallel, REST became a dominant web framework, through which a range of Application Programming Interfaces (APIs) emerged that were usable for a range of purposes, including by independent researchers to understand and evaluate the nature of social media discourse and patterns in algorithmic recommendations and streaming of content. This initial work was crucial in providing insights into the potential negative social effects of such platforms as well as how to advance prosocial outcomes.[60] However, as platforms have grown they have increasingly tended towards the development of 'walled gardens' to limit interoperability, keep users within their own services, and maximise control over user data and monetisation. These platforms have subsequently shut down or significantly limited access to their APIs creating a greater opaqueness about how their algorithms operate, such as whether disinformation is privileged or not.[61]

[57]The organisation For Humanity (https://forhumanity.center/) provides training in mandated and recommended AI audit tools and is central to growing the nascent professionalisation of AI auditors.
[58]ADM+S Toolkit (n 43).
[59]Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act' (2024) *Computer Law & Security Review* 53: 105957.
[60]Robert Bodle, 'Regimes of sharing: Open APIs, interoperability, and Facebook.' (2011) *Information, Communication & Society* 14(3): 320-337.
[61]Axel Bruns, "After the 'APIcalypse': Social media platforms and their fight against critical scholarly research" *Information, Communication & Society* 22(11): 1544–1566.

As platforms have evolved, their use of algorithmic systems has also increased, woven into every aspect of the everyday user experience. Paradoxically, as these platforms have grown in complexity and influence, their transparency has decreased. This observation extends beyond social platforms to recommender and risk assessment systems in social services which have increasingly become more complex. Often developed by commercial organisations the ability to know whether they are compliant with anti-discrimination laws or fit-for-purpose in a different geographical or jurisdictional location is not assessable. As AI systems become increasingly integrated into various everyday platforms and systems there is an urgent need to ensure greater transparency about all facets of these systems and their operations. Such requirements include a need for mechanisms to provide access for independent researchers for evaluation, which can be achieved in ways that do not compromise commercially sensitive or private data.

ADM+S researchers have been working on investigating these black box systems, via ADM+S' *Ad Observatory Project*, and the development of the Australian Internet Observatory.[62] The Ad Observatory project[63] uses a population-driven approach to enrol ordinary Australian platform users as data donors, using their deidentified ad data to develop accurate collective accounts of the highly personalised, ephemeral and intransparent platform advertising ecosystem. This project has been highly successful in revealing potentials for harm and lack of explainability of platform-based algorithmic advertising,[64] and critiquing limitations of existing self-regulated platform transparency approaches.[65] The Australian Internet Observatory will expand and extend this and similar innovative methods over the next four years with co-investment from the Australian Research Data Commons (ARDC). The observatory will deploy a range of technical mechanisms including user data donation, web scraping, and authorised platform data access to build independent accounts of the role of automated decision making systems in a range of platform functions and services. We would be happy to provide further information and consultation on this point to strengthen the assessment and reassurance of conformity.

An additional Guardrail? What role for stakeholder engagement

There may be arguments for including some obligation related to stakeholder engagement in the Mandatory Guardrails, in line with the 10 guardrails for the Voluntary AI Safety Standards, Guardrail 10: '*Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness*.' This Guardrail is equally and perhaps more important for high-risk AI.

[62]https://internetobservatory.org.au/

[63]Angus, Daniel, Lauren Hayden, Abdul Karim Obeid, Xue Ying Tan, Nicholas Carah, Jean Burgess, Christine Parker, et al, 'Computational Methods for Improving the Observability of Platform-Based Advertising' (2024) *Journal of Advertising*, doi:10.1080/00913367.2024.2394156.

[64]J. Burgess, N. Carah, D. Angus, A. Obeid, M. Andrejevic, 'Why Am I Seeing This Ad? The affordances and limits of automated user-level explanation in Meta's advertising system' (2024) *New Media & Society*, 26(9), 5130-5149 <https://doi.org/10.1177/14614448241251796>.

[65]Nicholas Carah, Lauren Hayden, Maria-Gemma Brown, Daniel Angus, Aimee Brownbill, Kiah Hawker, Xue Ying Tan, Amy Dobson, Brady Robards, 'Observing "tuned" advertising on digital platforms' (2024) *Internet Policy Review* 13(2).

A *stakeholder first* approach to responsible AI development and deployment can support the scrutiny of risks and harms more effectively than internal or self-audits.[66] Stakeholder engagement presents an opportunity to boost the inclusiveness of AI systems through either their development phase or deployment, improving the likelihood of representation of those most likely to be affected or potentially harmed by the deployment of AI systems.[67] The absence of stakeholder engagement in the mandatory guardrails creates a substantial gap in the capacity to identify emergent risks and harms as they impact diverse people who use or engage with AI systems.

Results of a study from UTS would also support engaging with workers as stakeholders in the decisions about the development, deployment, and use of AI systems in the workplace.[68] In the sectors studied in the report, workers were generally left out of discussions about AI in their workplaces and consequently did not trust the system and felt the use of AI created serious risks.

Civil society and not-for-profit organisations can play a leading role in establishing best practice in identifying emergent AI risks for communities and target groups and mitigating harms through their capacity to support or advocate for vulnerable and marginalised people and groups. Key stakeholders across parts of civil society can support the implementation of the guardrails, to ensure individual, group and cultural harms can be avoided, including financial services, community health, disability, aged care, employment services, family services, housing and homelessness services, First Nations organisations and consumer rights organisations.

Civil society organisations and not-for-profit organisations are well placed as key stakeholders in mitigating the harms of high-risk AI through their experience and expertise in dealing with sensitive data, maintaining privacy rights, autonomy and dignity for marginalised and disadvantaged groups in society. ADM+S research recognises the urgent need for supporting data and AI governance practice across public services, the NFP sector and public health sector.[69]

[66]D. Domínguez Figaredo, J. Stoyanovich, 'Responsible AI literacy: A stakeholder-first approach' (2023) *Big Data & Society*, *10*(2), 20539517231219958; A. Bell, O. Nov, J. Stoyanovich, 'Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance' (2023) *Data & Policy*, 5, e12.
[67]A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, S. Mohamed, 'Power to the people? Opportunities and challenges for participatory AI' In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (2022, October) pp. 1-8; A. McCosker, X. Yao, K. Albury, A. Maddox, J. Farmer, J. Stoyanovich, 'Developing data capability with non-profit organisations using participatory methods' (2022) *Big Data & Society*, 9(1), 20539517221099882.
[68]UTS Human Technology Institute, 'Invisible Bystanders: How Australian workers experience the uptake of AI and automation', (2024) Sydney: UTS <https://www.uts.edu.au/sites/default/files/2024-05/EssentialResearch%2BUTS_Invisible_Bystanders_0524_D4.pdf>; see also S. Gosper, J. Trippas, H. Richards, F. Allison, C. Sear, S. Khorasani, F. Mattioli. 2021. Understanding the Utility of Digital Flight Assistants: A Preliminary Analysis. 3rd Conference on Conversational User Interfaces (CUI '21), ACM, 1-5.
[69]Y.B. Kang, A. McCosker, M. Savic, T. Graham, P.P. Jayaraman, 'AI Governance in the Smart City: A case study of garbage truck mounted machine vision for roadside maintenance.' (2023) Swinburne University of Technology, Melbourne, DOI: 10.25916/a2fn-yb49; A. McCosker, X. Yao, K. Albury, A. Maddox, J. Farmer, J. Stoyanovich, 'Developing data capability with non-profit organisations using participatory methods' (2022) *Big Data & Society*, 9(1), 20539517221099882; A. McCosker, F. Shaw, X. Yao, K. Albury, 'A Data Capability Framework for the Not-for-Profit Sector.' (2022) Swinburne University, Melbourne, DOI: <https://doi.org/10.25916/h2s6-r178>; K. Albury, S. Mannix, 'Digital and Data Capabilities for Sexual Health Policy and Practice: Stage One Report' (July 2023) Melbourne: Swinburne University of Technology. <https://doi.org/10.25916/782g-nk95>.

Civil society and not-for-profit organisations are also well-placed to work with end users who are experiencing intersectional disadvantages by bringing a strengths-based trauma-informed[70] approach to community engagement. These approaches do not focus on looking for harm but on engaging with stakeholders in a way that focuses on care and healing, acknowledging the widespread impact of trauma across the population. Often, general AI ethics principles are in tension with trauma-informed approaches and the professional principles of those working in social services delivery (for example, social work ethical principles and professional codes of conduct[71]). In *Building a Trauma Informed AI Assessment Tool*,[72] ADM+S researchers showed how trauma-informed principles can be combined with AI technological knowledge to guide AI development, deployment, and monitor impacts. This research can inform the application of the Guardrails and ensure stakeholder engagement is meaningful for end users in situations of intersectional disadvantage, and to do so in a way that empowers and heals rather than causes additional harm.

Question 10: AI supply chain and AI Lifecycle

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?

AI supply chains can be varied, complex, dynamic, and in many cases multi-jurisdictional.[73] In practice, organisations seeking to operationalise the Mandatory Guardrails framework will find that neither AI 'systems', nor the distinction between development and deployment are in practice, as clear cut as the Proposals Paper might suggest. We can therefore expect that feedback on the Proposals Paper will raise questions about how the definitions are intended to apply in a range of situations known to or identified by various organisations.

[70]Harris, M., & Fallot, R. D. (2001). Envisioning a trauma-informed service system: A vital paradigm shift. *New directions for mental health services*, *2001*(89), 3-22 <https://doi.org/10.1002/yd.23320018903>.
[71]Australian Association of Social Workers (2020) Australian Association of Social Workers Code of Ethics 2020 <https://www.aasw.asn.au/about-aasw/ethics-standards/code-of-ethics/>; Australian Association of Social Workers (2023) AASW Practice Standards <https://www.aasw.asn.au/about-aasw/ethics-standards/practice-standards/>
[72]ADM+S Toolkit (n 43).
[73]For example, see the ADM+S study, ADM in Social Services (n 49) 93-4, highlighting how many systems use multiple AI applications, in various stages, in some cases originating in or being managed from multiple jurisdictions. In the 28 case studies they investigated, out of a total of 42 identified technologies there were at least 27 different AI technologies that were sourced from outside Australia, from 10 different jurisdictions, spanning five continents.

In terms of the definitions themselves, we note that it may be to recognise that AI supply chains include companies who harvest user data in order to sell to AI developers. In our internal consultations, one suggestion was that the definition of 'AI supply chain' could be amended to read as follows:

> **AI Supply Chain:** The complex network of actors and organisations that enable the use and supply of AI throughout the AI lifecycle from model design, testing and fine tuning to deployment and integration into the local IT system. These actors and organisations include any third party who harvests, supplies or onsells existing data to AI developers.

We suggest that a **degree of uncertainty or fuzziness of these concepts is both inevitable, and in practice, tolerable** in establishing, and operationalising, a framework for Mandatory Guardrails as proposed. The purpose of this regulatory framework is not to strictly divide and apportion responsibility for the purposes of liability, but to require organisations to adopt responsible internal and external processes for managing the risks of AI. It is likely in practice that many organisations will be both developer and deployer of different components of AI systems, as we illustrate below. And that is ok: internal systems and governance can and should be designed to allow organisations to fulfil the closely related responsibilities that apply when an organisation is developing an AI system, and when they are deploying.

Defining AI systems, developers and deployers

'Deployers' and 'developers' are identified by reference to their role in relation to an 'AI system'. When entities operationalise the Guardrails, they may thus need to work out what the system is, to work out their own obligations. As ADM+S researchers have noted elsewhere, this can be challenging, and views can differ:

> 'System' can … have a range of meanings and operate at a number of levels. A database used for multiple automated activities (such as issuing a range of licences, or licence renewals, and tracking demerits against those licences) could be considered a single 'system', or a series of ADM systems, on the basis that there are multiple different decisions being made, with potentially different degrees of impact on people.[74]

As part of a process of mapping automated decision-making (ADM) systems in NSW governments, ADM+S researchers asked government entities (via both survey, and in interviews) to identify ADM systems that they were using. Views within entities differed on what the boundaries of any given system were:

> We observed that local governments and small state government agencies often reported ADM systems that were small in scope or size, whereas large agencies with high levels of automation (such as the Transport portfolio) reported ADM systems at higher-level systems, with multiple automated functions. A good example is our case study of LifeLink and the Online Birth Registration systems. These systems make possible, among other functions, customer-driven online registration of births. At different points the views within the NSW Registry of Births Deaths and Marriages (BDM) diverged: for some LifeLink was the reportable ADM system, whereas for others, focusing on the function (the registration), the Online Birth Registration system (OBR) was the ADM system. Depending on the approach, this could be reported as one system, or multiple systems to match the number of functions it could perform (e.g., registration of births, marriages, deaths, etc.), increasing as these functions are progressively added.[75]

[74] ADM in Governments (n 14) 15.
[75] ADM in Governments (n 14) 15-16.

We can also think about the complexity of 'AI systems' by thinking about how larger 'systems' can be created from multiple components. Imagine, for example, the following:

> Company Y runs a Learning Management System (LMS) platform used by high schools. Company Y develops AI features available on the platform which allows students to interact with their course material, AI functions for self-testing and testing for assessment (including AI systems to suggest feedback on and assign marks to essays and presentations). The AI features is powered by Claude 3 (Anthropic) but also involves integrations from Company Z's text-to-speech AI model and Company W's speech recognition AI model which is available via API without a user interface. School B procures access to the system (which is offered as a cloud-based service) for use by its teachers and students.

Assume, for the moment, that the system is considered high risk. In this scenario, we have a broader AI system (the LMS), as well as a text-to-speech AI function/model, and a speech recognition function/model which could themselves be thought of as 'AI Systems'.[76] It may well make sense to treat the latter two integrations as themselves AI systems, meaning the responsibilities of developers should fall Y, Z, and W in relation to different aspects of the overall LMS. It is also arguable, in this scenario, that *both* Company Y and School B are deployers: both are 'supplying or using an AI system to provide a product or service': Company Y is offering the LMS as a service; School B is procuring and using the AI-enhanced LMS to offer educational services to students. Company Y is also a developer of the LMS system.

This can still work, as the Guardrails can sensibly be applied to both:

1. Each must have accountability processes, including details of who is responsible for ongoing monitoring (probably both; preferably with the division of responsibility based on opportunity and capacity, and made explicit in any procurement contract);
2. Each will need to follow instructions: Company Y will need to follow instructions from earlier developers in the supply chain (Z and W); School B will need to follow Y's instructions. Both will need to engage in ongoing oversight and observation.
3. Regarding data: each of Y and B are responsible for the data they provide/use/input;
4. Regarding testing: Y (as developer) is responsible for testing; Y and B will both need to engage in ongoing monitoring of operations (once again, this could be via a sensible division of responsibility, with Y monitoring the technical system in operation and B monitoring for student and teacher reactions, responses, questions or complaints);
5. Regarding human control, intervention and oversight: each will need to train their staff;
6. On informing end users: it is most obviously B who is responsible for this, but Y will need to inform B of how AI is being used; Both Y and B should have avenues for concerns or complaints (it would not be legitimate for either Y or B to simply refer complaints to the other!)
7. Both Y and B should have avenues for concerns or complaints (it would not be legitimate for either Y or B to simply refer complaints to the other!)

[76]See, for example, the definition of 'AI System' in art 3 of the EU AI Act of an 'AI System': a 'machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.

8. B should inform Y, and Y should inform Z and W of any incidents or risks
9. 9 and 10: Both will need to undertake record-keeping and conformity assessment.

Other matters

We draw to the attention of the Department some other matters of detail raised by ADM+S researchers.

First, as we have highlighted throughout this submission, as the Guardrails develop, further work will likely be required in articulating other key responsibilities and roles within AI supply chains. A challenge will be to find a taxonomy of participants in the supply chain that is sufficiently simple to operationalise, but which is sufficiently nuanced to capture differences in capacity and responsibility between different participants. The 'fuzziness' of the developer and deployer roles is, as noted above, inevitable, and need not for now be a cause for concern.

There are however other actors within AI supply chains who have the capacity to foresee or respond to risks and harms in ways that are different from, and in some cases better than, developers or deployers. In time it may make sense to single them out for additional responsibilities and roles. For instance, platforms or model marketplaces such as HuggingFace and Civittai provide a model 'hub' or repository (sometimes alongside other services). In this capacity, they supply models, but not necessarily with the goal of providing a product or service. They are not developers or deployers, but it may nonetheless be appropriate for them to bear some burden of responsibility for their role in model distribution. Veale and Gorwa have recently explored the capacity of such actors to operate as regulatory 'chokepoints'.[77] They have a unique capacity to enforce the licences under which certain AI models or datasets are made available, and to operate 'notice and takedown' processes for non-compliant or harmful models. Indeed, over the past 24 months HuggingFace has already been acting in this way to an extent – for instance by making platform-wide decisions to promote the use of certain model reporting standards for all new model uploads.[78] Perhaps the word 'model host' or 'model distributor' could be included somewhere to expand the typology of AI ecosystem actors (noting however the multiple meanings of the word 'host' in terms of internet infrastructure)?

Second, in a world where some AI systems are offered 'to the world' or direct to the public, what is an organisation's responsibility should employees use such a public system rather than one centrally procured by the organisation? For example, imagine an organisation has specifically procured and made available one AI system incorporating generative AI features, but some subset of employees within the organisation have sought to use another 'public' system which is better at a particular task (or which they perceive to be better for some particular task, such as coding, or obscure language translation, or something else). Is the organisation to be seen as a deployer, because their employee is using the system to provide the organisation's services? We note this may be a matter for further consideration if, as we have suggested, questions of

[77]Robert Gorwa, Michael Veale, 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries.' (2024) *Law, Innovation and Technology* 16(2) <https://doi.org/10.31235/osf.io/6dfk3>.
[78]See, e.g. https://huggingface.co/docs/hub/model-cards

liability are referred to a body such as the ALRC — highlighting the importance, as a subsequent step, of engaging with liability questions. For now, Guardrails covering accountability and governance, and risk mitigation should assist organisations to develop systems for 'line of sight' as well as training of staff.

Third, should local deployers using AI from developers overseas, or incorporating AI components sourced from overseas entities into their locally developed AI system, bear any additional responsibility to ensure that overseas developers are in compliance with the Mandatory Guardrails?

- On one view, the goals of the Guardrails may not be achieved if overseas developers do not comply with the Guardrails. For example, it could be argued that the Guardrails requirement that developers document and disclose data sources will be ineffective if developers are overseas, so deployers should have some responsibility to ensure the developer guardrails are met for any systems they propose to use or procure. On the other hand, this could limit options and opportunities for local developers and deployers.
- Or should it be sufficient for local developers or deployers to ensure that they use AI systems that are governed by an equivalent legal framework (for example, Europe's framework) or where overseas developers or deployers have voluntarily adopted a relevant international standard such as the NIST framework?

Questions of international interoperability are addressed in more detail in a separate submission from Bello y Villarino and Weatherall.

Question 12: Regulatory burden

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

We recognise that the question of regulatory burden is important; the application of all mandatory guardrails for all high risk systems has the potential, if the Guardrails are stringently drafted, to preclude participation in AI innovation from some SMEs. On the other hand, while in some cases the degree of risk posed by an AI system will be less for a SME (for example, because their reach, or the number of people affected is reduced) this is not necessarily true (for example, if only a small number of individuals are affected but the harmful consequences are very high for those individuals).

We also recognise, however, that the requirements of the Guardrails are intended to have some flexibility in their operationalisation: that steps for risk mitigation and oversight may be proportional to the particular risks involved, and their likelihood.[79] This form of proportionality through principles of 'reasonableness' is well established in privacy law, tort law and elsewhere in Australian law.[80]

[79]There are other possible options. 'AI guardrail compliance' 3rd party services may assist; another alternative is that entities offer insurance or coverage for SMEs that sign up to a charter of best practices.
[80]See Henry Fraser and José-Miguel Bello y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough' [2023] *European Journal of Risk Regulation* 1.

Should the government wish to formalise a system for tailoring regulatory burden to risk and to organisational capabilities, there may be lessons to learn from the concept of 'maturity' as it is understood in the field of cybersecurity. Cybersecurity maturity is the degree of readiness of an organisation to respond to cyberthreats. Guidance on Australia's Essential Eight Maturity Model for cybersecurity suggests that 'When implementing the Essential Eight, organisations should identify and plan for a target maturity level suitable for their environment. Organisations should then progressively implement each maturity level until that target is achieved.'[81] The thrust is that all organisations with cybersecurity responsibilities or concerns are generally expected to work towards higher levels of maturity. It is, however, understood that there will be different baseline expectations and different levels of maturity for organisations depending on their size, capabilities, goals, activities, and of course the sensitivity of the data they handle.

The CIS controls, a leading industry standard for cyber resilience, provided by the Centre for Internet Security, is another example of a highly granular and tailored model of maturity.[82] There are 18 CIS controls — types of cybersecurity and cyber resilience practices — but CIS contemplates a tiered approach, with three implementation Groups. Implementation Group 1 is comprised of smaller businesses with fewer resources and less pronounced security concerns, and are only required to take some of the measures called for by each of the controls. They are expected to follow only essential 'cyber hygiene': 56 out of a total of 153 safeguards making up the CIS controls; usually fewer than half of the safeguards for each of the Controls. Implementation Group 2 follows an additional 74 safeguards, and Implementation Group 3 follows all 153 safeguards.

There will likely be some relationship between maturity (level of regulatory burden) and risk. Where the CIS controls designate different requirements to different implementation groups based on those groups' resources, the Canadian *Directive on Automated Decision-Making* provides specific guidance on how to meet general obligations in a way that is tailored to risk. So, for example, requirements for transparency or peer review are more stringent and expansive, and require more steps, for higher risk systems than for systems with a lower risk. The government may find this kind of differentiation of responsibility helpful in order to avoid excessive regulatory burden for small to medium businesses. It may also be helpful for AI systems which are, for example, at the lower end of the 'high-risk' spectrum.

[81] ASD, 'Essential Eight Maturity Model' <https://www.cyber.gov.au/resources-business-and-government/essential-cyber-security/essential-eight/essential-eight-maturity-model>.
[82] 'CIS Critical Security Controls Implementation Groups', Center for Internet Security (Web Page) <https://www.cisecurity.org/controls/implementation-groups>.

Topic 3: General-purpose AI

Question 5 and 7: Defining GPAI, and high or very high risk GPAI

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

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?

The Proposals Paper suggests applying the Mandatory Guardrails to **all** General Purpose AI (GPAI) models, on the basis that they are high risk because they pose unforeseeable risks. The proposed approach differs from developments overseas. The EU has high level requirements specifically drafted for GPAI (**Art 53**, Ch 5, EU AI Act), with additional obligations for **GPAI models with systemic risk** (**Art 55**),[83] which are dealt with by the Commission, on a case-by-case basis. There is also ongoing work in the EU to develop a Code of Practice for GPAI. There is also proposed legislation in the US, targeting only the largest models, that would impose a separate regime on such models. The US government has already been engaging with the large model developers for pre-release testing, via a series of voluntary commitments.

Definition of GPAI

A **general-purpose AI** model is defined in the Proposals Paper as:

> An AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems.

During internal ADM+S consultations, some have expressed some doubts about the proposed definition. The phrase 'adapted for use' is very broad: any machine learning system, and especially any neural network-based system, can be adapted for use for a variety of purposes and in a variety of applications. One option might be to confine the definition to models intended for general purpose use, but this too has imitations (it would exclude models that show 'emergent' abilities and/or happen to be general purpose without that being the intention of the designer). The EU AI Act definition in art 3 is perhaps more specific:

[83]In the EU AI Act, a general purpose model is presumed as having 'high impact capabilities' and therefore systemic risk (giving rise to additional requirements under art 55) when the cumulative amount of computation used for its training measured in floating point operations is greater than 1025. Note that the floating point operations threshold for GPAI to be 'presumed to have high impact capabilities' which triggers an obligation for the provider to notify the Commission. The Commission can then decide it is not OR if below the threshold it can still designate a GPAI as of systemic risk (51 (1) (a) and (b)). As always with the EU AIA, the Act establishes, not so much a rule, as a procedure to address the issue.
[84]We note that while the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 88 Fed Reg 75191 (30 October 2023) ('US Executive Order 14110') only addresses federal use of AI, the White House has also published a list of voluntary commitments from companies developing foundation models (September 2023). California's Bill, Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, SB 1047 (9 September 2024) would have targeted advanced AI models, requiring developers to make certain safety determinations before training AI models, comply with specified safety requirements, and report AI safety incidents; it was drafted to apply to models training compute measured at 10^{26} FLOPs, and costing more than $100M to train. The Act was recently vetoed by the Governor.

> an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market"

However, we note that the qualifiers "[significant] generality" and "[competently] performing" are vague and are likely to be points of contention.

The current proposed Canadian definition[85] perhaps strikes a better balance:

> general-purpose system means an artificial intelligence system that is designed for use, or that is designed to be adapted for use, in many fields and for many purposes and activities, including fields, purposes and activities not contemplated during the system's development.

In particular, the specification that the use adaptation is 'with respect to many fields and for many purposes, including [those] not contemplated during the system's development' seems to resonate well with current general purpose systems (such as foundation models), as well as potential near-future systems that are closer to aspirations of Artificial General Intelligence (AGI) held by AI developers such as OpenAI. These qualifiers also limit the scope more appropriately compared to the proposed definition in the Proposals Paper.

High Risk and GPAI

The core question implicitly asked in the Proposals Paper is: what is the relation between 'GPAI models' and 'high risk AI':

- Are there GPAI systems that are not high risk?
- Are there high risk systems that are not GPAIs?

In our view, there *may* be GPAI systems that are not high risk, but there are definitely high risk AI systems that are *not* GPAI. This means that GPAI is either a subset of high-risk AI, or an adjacent set with some overlaps. In any event, as (by definition) we do not know all the possible use-cases of a GPAI in advance, we cannot definitively exclude the possibility that it has high risk uses. It therefore makes sense from a regulatory perspective to treat GPAI as a subset of high-risk AI: that is, we should treat all GPAI systems as high-risk AI systems.

Technical metrics as a measure or indicator

A further question raised in the Proposals Paper is whether to define high-risk GPAI (or, perhaps, 'very' high (systemic) risk GPAI to which additional obligations should attach), by reference to technical metrics, such as effort used in training (e.g. measured by computation FLOPs, electricity consumption, engineering FTE, or fiscal training budget), or the size of the training data (e.g. total number of tokens) or model capacity (e.g. total number of trainable parameters) or system complexity (e.g. architectural configuration or number of hyper-parameters).

[85]Proposed as part of amendments proposed in November 2023 to the Canadian Bill, the Artificial Intelligence and Data Act, part of Bill C-27, introduced in 2022. See the Proposed Amendments to the Artificial Intelligence and Data Bill 2022, 28 November 2023, <https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf>. It is unclear whether, or when this Bill might become law, or whether further amendments are likely to be made.

In the course of internal consultations on the Proposals Paper, some argued that technical metrics are not a suitable measure for 'general-purposeness' or for risk thereof. There are several arguments against using such technical measures for defining high (or very high) risk GPAI:

1. First, the Proposals Paper has a subtle but meaningful typo: Floating Point Operations *Per Second* (FLOPS — uppercase S) is a measure of training throughput (how fast you can do the calculations required to train a model), not total capacity of training (how much training is performed or needed).[86] FLOPs (note the lower-case 's') is a measure of total training required or performed, and is more correlated with 'general-purposeness', although still problematic as argued further below.
2. Technical measures such as training required or training data required are only, so far as we know, *correlated* with general purpose model abilities/capacities — they are not *causally linked* with general purpose abilities. Although it is true that empirical model scaling 'laws' that predict model ability from training metrics have until now held quite tightly,[87] there is no guarantee these empirical predictions will continue to hold into the future.
3. Technical measures gloss over too many details. For example, 'total training tokens' glosses completely over the kind of tokenizer and input data-extraction processes used (which is increasingly important for multimodal models). References to FLOPs gloss over details about how effective the training was, and questions of the training schedule (how many epochs of training were done, the data chunking approach etc.). These decisions and hyper-parameters have important impacts on training efficacy and hence the resultant 'general purposeness'. Any technical metric would offer no insight into these details, and would therefore be a poor measuring stick for general purpose ability.
4. Technical metrics like FLOPs, Tokens, MegaWatts, etc. are tightly intertwined with model architectures and training algorithms, and physical computational infrastructure. Advances in model architectures (such as inference-time scaling[88]) or training methods (such as the advent of the Adam optimizer[89]) or training hardware (such as the TPU[90]) represent unpredictable step-changes in the general purpose capability obtained given a fixed size of any particular technical metric. There will continue to be more advances in architecture, training algorithms, and infrastructure. which will at best require technical metrics to be regularly reviewed, or at worst, make them totally untenable as guideposts for 'general purpose' ability.

All of these arguments are consistent with effort/data/complexity/capacity being *some* indicators of very high risk GPAI, but incomplete ones, and ones that is likely to change in their significance. In other words, the determination of which GPAI models pose very high or systemic risk will be affected by technical developments, including innovations in model architectures and training algorithms.

[86]See for example Dzmitry Bahdanau, 'The FLOPs Calculus of Language Model Training', *Medium* (Blog Post, 10 January 2022) <https://medium.com/@dzmitrybahdanau/the-flops-calculus-of-language-model-training-3b19c1f025e4>.
[87]See for example Jordan Hoffmann et al, 'Training Compute-Optimal Large Language Models' (2022) *arXiv* 2203.15556.
[88]Charlie Snell et al, "Scaling LLM test-time compute optimally can be more effective than scaling model parameters." *arXiv preprint arXiv:2408.03314* (2024).
[89]D.P. Kingma, J. Ba, 'Adam: A Method for Stochastic Optimization' (2014) CoRR, abs/1412.6980.
[90]Norman P. Jouppi et al. 'In-datacenter performance analysis of a tensor processing unit.' Proceedings of the 44th annual international symposium on computer architecture. (2017).

What this means is that any definition of high-risk GPAI is likely to need to **change over time**. We have argued throughout this submission that an entity (an AI Body), informed by technical and socio-technical expertise, will need to play a role in dynamic adjustment of any system for AI governance: for example, in updating any indicative or prescribed list of high risk AI systems or uses. That is, we will need a regulator body or scientific body with the power to independently declare certain specific models as high risk or very high risk (or systemic risk, in the EU parlance) GPAI, and therefore subject to more stringent regulation. For more, see Topic 5 below.

Questions 6 and 11: The Mandatory Guardrails and GPAI

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

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?

Can the proposed Mandatory Guardrails apply to GPAI?

The important implication of GPAI's *general purpose* nature is that it is difficult to predict uses and impacts. Their general nature means we cannot know in advance what all the uses (variety of purposes) will be, with the result that there *might* be increased risk that is not foreseeable. While for the most part the Guardrails are use case agnostic, and so capable of application to GPAI, applying some guardrails will be challenging when combined with the predictive difficulties introduced by GPAI. Specifically, we highlight:

- Guardrail 2 (risk management and mitigation): Identification of risks stemming from an AI system will be more difficult for GPAI systems where the potential applications, let alone risks, are harder to identify a-priori.
- Guardrail 3 (data quality and provenance): Protecting GPAI systems from poor quality or malicious data sources may be more difficult, due to the unforeseeable ways these systems might operate on data. For example more advanced LLM models are known to have a largern 'attack surface' for prompt injection attacks — that is, the general purpose nature means malicious users have more ways of tricking the model into performing forbidden actions.[91]
- Guardrail 4 (testing and monitoring): Testing GPAI systems may be much more difficult, as you can't test specific use cases in advance, or you might need to test very many use cases, or the tests must be are very general and therefore lose predictive power.
- Our proposed Guardrail 11 (stakeholder engagement): Engaging effectively with the right stakeholders may be very difficult for GPAI systems where the question of which stakeholders are relevant is itself difficult to answer.

[91]Yupei Liu *et al,* 'Formalizing and benchmarking prompt injection attacks and defenses.' *33rd USENIX Security Symposium (USENIX Security 24)* (2024).

We have assumed throughout this submission that the Guardrails are intended to have some internal flexibility, and that risk management, risk mitigation, oversight, disclosure and the like are all intended to be applied in a way that is proportional and adapted to the AI systems, and risks involved. If this is correct, it may be mostly possible to apply the Guardrails flexibly to GPAI (and we note the adjustments specifically noted in Attachment E to the Paper). In any event, as we have noted above in our discussion of Question 8, there will be more work to do drafting the particular requirements that are to be mandated as part of each Guardrail, and more specific adjustments and flexibility can be included at that stage. Ongoing work via the Working Group on trustworthy GPAI in Europe will also be relevant in providing more detailed guidance on how the Guardrails apply to GPAI. One possible avenue to deal with this is to insert in any legislation an obligation for providers of GPAI to cooperate with government to identify risks, similar to what is envisaged in the EU AI Act in article 53(3).

Are additional protections needed, perhaps for 'very high' or 'systemic' risk GPAI'?

Another question is whether there are additional requirements that ought to apply to very high risk or very large models — beyond the 'standard' Mandatory Guardrails. Specific, additional obligations have been included in the EU AI Act (art 55) and were proposed in California's (recently vetoed) SB1047. We note that the position in the US is fluid: while SB1047 has been vetoed, the developers of the largest models have engaged in some voluntary cooperation with the federal government; being voluntary it is inevitably unclear how long such cooperation will last, or how effective it will be.

The crux of additional existing or proposed requirements is that for GPAI that has potential large-scale, systemic risks, there should be additional cybersecurity and testing obligations, as well as deeper levels of engagement with government (such as reporting incidents). The recently vetoed SB1047 in California would have also included a requirement that such systems have capacities for shutdown built in. In Europe, further development of standards is expected to be developed through Codes of Practice (see art 56).

We make the following brief comments on this question:

- Given that the Mandatory Guardrails will mostly involve organisations engaging in *internal* processes such as risk assessment and mitigation and testing, it makes sense that further *external* requirements including reporting to government should apply for some subset of very high-risk models. To the extent that systemic risks arising from large models or incidents relating to large models pose cyber security risks or risks to critical infrastructure, reporting and cooperation obligations already exist under other legal regimes. Cybersecurity or critical infrastructure-related risks however do not cover the full range of risks arising from AI (as discussed above in relation to the definition of high-risk systems).
- Any model which poses systemic risk within Australia is likely to pose similar risks beyond Australia. This is therefore an area where it is especially critical that Australia promote

international cooperation, and that any national regulation be developed in such a way as to ensure effective interoperation with global efforts, including with efforts following on from the Bletchley Declaration[92] and the recent final UN Report on *Governing AI for Humanity*.[93]

- It will be important that Australia develops effective mechanisms for genuine engagement with international AI Safety cooperation, including via the growing network of AI Safety Institutes.

Topic 4: First Nations Australians rights and interests

Question 2: How could the principles better capture harms to First Nations people, communities and Country?

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

It is critical that Australia take active steps to manage the particular risks, and any particular opportunities that arise from use of AI impacting on Australia's First Nations people.

Current models are trained on data and embedded with values that tend to be white, Western, English-speaking, Global North skewed, and so on. Far from being universal, AI is perhaps better understood as 'artificial Western ethno-intelligence'.[94] Such models have the capacity to reinforce historical injustice, to amplify labour precarity, and to cement forms of racial and gendered inequality. To avoid this bias and its concrete impacts on lives and livelihoods, models must be redesigned to include more expansive ways of being and knowing, including those of the Indigenous and First Nations peoples in Australia. An alternate set of values and visions could be used to both design and evaluate AI models, as nascent research has explored in the context of Aotearoa New Zealand.[95]

While such a project is certainly ambitious and long-term, there are existing frameworks in parallel disciplines (e.g. environmental impact) and parallel contexts (First Nations peoples in Canada, the United States) that could be used for integrating some of this Indigenous perspective. One immediate need relates to the breadth and depth of data around Indigenous knowledge, customs, and practices. This issue in some respects replicates that of 'low-resource' languages, where inference, accuracy, and veracity all suffer due to a lack of training data. Of course, any generation or collection of such data in Australia should not employ the 'scraping' or harvesting model prevalent in Silicon Valley, but rather work alongside Aboriginal and Torres Strait Islanders to collect, curate, and order data in ways that are both ethical and

[92]'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', GOV.UK (Web Page, 1 November 2023). <https://web.archive.org/web/20231101123904/https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.
[93]AI Advisory Body, *Governing AI for Humanity* (Final Report, September 2024).
[94]Deborah H. Williams and Gerhard P. Shipley, 'Enhancing Artificial Intelligence with Indigenous Wisdom' (2021) 11(1) *Open Journal of Philosophy* 43.
[95]Luke Munn, 'The Five Tests: Designing and Evaluating AI According to Indigenous Māori Principles' (2023) 39 *AI & Society* 1673.

culturally-sensitive. These questions may arise in a wide array of circumstances in Australia due to traditional owner claims over — and sovereignty in — land, water and living species. This means that AI developers and deployers will need be mindful that need to engage with First Nations peoples and protocols may arise in relation to a wide array of training and deployments of AI enabled systems for purposes such as conservation monitoring of endangered species, mineral and resource exploration, and analysis of novel uses of plants for nutritional or pharmaceutical purposes.[96]

Relatively little detail is provided in the Proposals Paper on this question. There are ways that the Mandatory Guardrails could be adjusted, for example to explicitly require that any risk assessment address specific risks relating to First Nations Australians. ADM+S has an existing research program on questions relating to the digital divide, or digital gap affecting Australia's First Nations People, in particular through our ***Mapping the Digital Gap project***.[97] ADM+S can therefore offer some assistance. However, we emphasise that while the ADM+S may be able to assist, the much-needed further development must be Indigenous-led and recognise indigenous data sovereignty.

Topic 5: Regulatory options to mandate guardrails

Question 13-16

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?

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

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

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?

Assuming that we accept that mandating guardrails is something Australia should do, a key question is how.

The Paper canvasses 3 options:

- **Option 1**: Adapt the Guardrails for each domain, and insert them into domain-specific legislation or regulations.

[96]Jennifer Mairi Macdonald, *et al*, Indigenous-led responsible innovation: lessons from co-developed protocols to guide the use of drones to monitor a biocultural landscape in Kakadu National Park, Australia (2021) 8(2) *Journal of Responsible Innovation* 300–319; Vishal Rana, Indigenous Data Sovereignty: A Catalyst for Ethical AI in Business. *Business & Society* (2024) 0(0).
[97]'Mapping the Digital Gap', *ADM+S Research Projects* (Web Page) <https://www.admscentre.org.au/mapping-the-digital-gap/>.

- **Option 2**: Pass 'framework legislation' that includes the Guardrails but does not mandate their application until 'activated': that is, when other laws (or regulations) are amended to refer to the framework legislation.
- **Option 3**: Pass horizontal legislation, making the Guardrails mandatory across the economy for high risk AI systems. Where Options 1 and 2 involve action by departments and regulators responsible for specific domains, option 3 would require designating an entity responsible, and determining a division of labour between that entity and existing agencies and regulators. The new entity could be an existing regulator, with new powers,

Option 1, 2, or 3?

In summary, our view is that a combined approach is inevitable, but that, in light of rapidly changing technology, the fact that regulatory frameworks are already in place in a number of important jurisdictions, and the need for Australia to take action sooner rather than later, some form of horizontal legislation (identified in the Proposals Paper as Option 3) will be required. Horizontal legislation would not necessarily cover the entire economy: the *EU AI Act* for example carves out many areas where domain specific rules continue to apply, especially in heavily regulated sectors (transport, finance, energy etc).[98]

Option 1 is unlikely to achieve the goal of safe and responsible AI, or building public trust, at least in the near term. Option 1 poses a number of problems:

- **Time and delays**: the time available on the agenda of both government, and Parliament, is limited. It is simply unrealistic to think we could, within the kind of timeframes necessary, develop specialised legislation across a wide range of areas, undertaking the necessary further stakeholder consultation each piece of legislation would require, followed by separate Parliamentary consideration and negotiation of passage.
- **Gaps**: while some regulators and departments are likely to be active and engaged with questions relating to AI, there are others where the issue is likely to fall much lower down the priority list. This is not a criticism: in the area of work and employment, other priorities such as ensuring people are fairly paid; addressing workplace health and safety; and tackling modern slavery are all high priority areas requiring significant attention. Where there are other priorities, we would anticipate gaps in addressing AI risks would be likely to persist for some time. In other areas there may not be an obvious regulator with responsibility and powers to address issues arising in relation to AI.
- **GPAI**: domain-specific rules, possibly with variations across domains (since that is the point that justifies Option 1), is a poor way to address cross-economy issues raised, for example, by GPAI.
- **Complexity and regulatory burden**: many Australian firms operate across or offer services to more than one industry. System developers, for example, would have to be able to certify to deployers that they met any Mandatory Guardrails across every sector, which will become more challenging, the more variation there is.

[98] *EU AI Act* (n 19) annex I.

- **Incoherence**: it would likely lead to incoherence. As Weatherall has noted elsewhere,[99] there is already a lack of coherence developing around AI regulation and governance within and across Australian governments.

The chief argument in favour of Option 1 is that domain expertise, and domain-specific rules are required, because AI is used differently in different domains. This argument, however, misses the point of the Mandatory Guardrails. What they mandate are not general substantive rules regarding what AI can and cannot be used for (which might require domain specificity), but *processes*, which lend themselves to more ready standardisation, even across industries or domains. And while, for example, the particular risks or mitigations will vary across (and notably, *within*) sectors, particular risks or mitigations cannot be addressed at the level of legislation, which cannot be easily changed. We have argued throughout this submission for high level, generalised guardrails with flexibility, supplemented with updates (via disallowable instruments) and guidance.

Option 2 is unclear, but so far as we understand it, the Guardrails would exist at a central level, but would not apply to any given industry, or in any given context, until some kind of action were taken by a responsible regulator or Minister.

This seems to have many of the problems of Option 1. In particular, we will likely still have problems of **delay** and **gaps**. Responsible government agencies would need to engage in consultation prior to activation, and would be urged not to 'activate' Guardrails until detailed guidance on the application of the guardrails in the domain context has been developed. This risks making the promise to the Australian people, implicit in the passage of legislated, Mandated Guardrails an empty one. This would be destructive of already low public trust.

As noted by Weatherall elsewhere, Option 2 would seem to require coordinated action on the part of the Commonwealth government that is not supported by evidence to date.[100]

If Option 2 is preferred, it would need to:

- establish common terminology, processes and rules for AI systems that can be applied across domains;
- include rules to deal with conflicts of competence for situations where more than one regulator has jurisdiction;
- include rules for GPAI, which is inherently domain-crossing.

If all these matters are addressed, we are close to doing the work necessary for Option 3, but without gaining the benefits Option 3 provides. This strengthens the case for Option 3.

We also note that an advantage of Option 2 is that it allows for a federal approach, where states could also potentially 'activate' the guardrails: ie adopt the same guardrails as the Commonwealth? Many areas of regulation are state-based (eg property and rental laws are the domain of the States and this could be an important area). This is certainly an advantage over Option 1, but it may also be constitutionally possible to build mechanisms for state adoption model into Option 3. Consumer protection law, for example, operates via a cooperative scheme

[99]Kimberlee Weatherall, 'The Mandatory Guardrails Proposal Paper: An Opinionated Explainer', 25 September 2024.
[100]Weatherall, Opinionated Explainer (n 99).

where the Commonwealth and States have agreed to legislate the same law, which is a schedule to legislation.

Option 3: This analysis would seem to suggest that some version of Option 3 — horizontal legislation that applies Mandatory Guardrails — is the most realistic means of *mandating* Guardrails for the purposes of ensuring safe and responsible AI development and deployment in Australia, and to protect Australians as well as equip Australian businesses to engage with developing global rules and frameworks. The Mandatory Guardrails applied this way are analogous to consumer protection law, or privacy law which applies across public and private sectors.

We understand that some may take the view that guardrails need to vary across different parts of the economy, or different issues. For example, it might be argued that risk assessment or mitigation should look differently for banks, than it might for a university or a company involved in cutting-edge medical research. The concern is that 'whole of economy' guardrails will lead to silly results (guardrails that are simply inapplicable), or won't capture the unique risks that arise in different areas, or will mean that guardrails must be expressed at such a high level of generality that they are largely meaningless: little better than the AI Ethics Principles we already have.

We think these concerns are more apparent than real. The Guardrails would require processes and governance (testing, transparency, accountability, processes for risk identification and mitigation), rather than specific directions *how* to test, or how AI should or shouldn't be used (beyond perhaps a small subset of prohibited uses). To the extent that detail is required, it should be provided by a mix of disallowable instruments, and guidance, including from domain-specific regulators.

Importantly, having horizontal legislation does not necessarily mean the Mandatory Guardrails will be identical for all sectors. First: there is inherent flexibility within each of the Guardrails.

Second, the Guardrails could be subject to staged introduction and/or variation between the private and public sectors. In particular, it may be advisable at an early stage to establish more constraining or detailed rules for **public sector uses**. Australian governments (Commonwealth, state and territory governments) have committed to being **exemplars** in the safe and responsible use of AI, and to adopting a lawful, ethical approach that places the rights, wellbeing and interests of people first.[101] **This is appropriate.**[102] Addressing public sector use immediately and at a higher standard is consistent with the only existing treaty on AI use to date, the Council of *Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*,[103] which imposes stronger obligations in relation to government use. And both in Canada and in the US there are already quite strong binding rules for government AI use:[104] rules that have come in advance of any regulation for the private sector.

[101]The commitment was made via the *National Framework for the assurance of artificial intelligence in government*, 21 June 2024. This is a joint approach agreed and released by the Data and Digital Ministers Meeting of that date.
[102]See generally J-M Bello y Villarino, 'A Tale of Two Automated States: Why a One-Size-Fits-All Approach to Administrative Law Reform to Accommodate AI Will Fail'. In Z Bednarz, M Zalnieriute, eds. *Money, Power, and AI: Automated Banks and Automated States*. (Cambridge University Press, 2023) 136-151.
[103]COE Framework Convention (n 3). The treaty was signed on 5 September 2024 at Vilnius, by the EU, US, Israel, UK and six other non-EU European countries. Australia participated in its negotiation but has not yet signed.
[104]See the Canadian Directive on Automated Decision-Making (n 4) and the US Executive Order 14110 (n 84).

Third, we could design option 3 Mandatory Guardrails as a **default minimum set**: a set which can be expanded upon (in terms of additional detail, or additional guardrails) through industry codes of practice (or even updated legislation). Applying the Guardrails by default until superseded by more tailored industry-specific detail provides incentives for stakeholders to cooperate in designing domain-specific detail, rather than seeking an unattainable perfect ruleset. It would be important to ensure that any negotiated alternative to the default does not provide less/reduced protection for Australians: this could mean that the responsible entity or agency should have both involvement in, and power to refuse approval to, negotiated domain-specific variations.

We note that Option 3 requires consideration of constitutional issues. In relation to State public sectors, all the States and Territories have committed to the *National AI Assurance Framework*; we would urge all Australian governments to follow up on that commitment, and in a coordinated way. AI is not a technology that can or will be constrained by state borders, and all Australian stakeholders — *especially* industry but extending to Australian people and civil society — will benefit from avoiding duplication, overlap and gaps which will arise if a national approach is not adopted.

Another key issue which will need to be determined in Option 3 is how **questions of interoperability with the domestic frameworks of other jurisdictions** will be addressed.[105]

The critical question of institutional structure and the need for ongoing, iterative design

A key message at every point in this submission has been that any regulatory framework imposing Mandatory Guardrails will need **ongoing, iterative development**. In the context of rapidly changing general purpose technology fuelled by significant investment; unpredictable future technologies; and shifting business models and societal attitudes, successful government intervention requires:

- Action as soon as possible, to join equivalent countries in offering protection to citizens and residents from unsafe, and irresponsible development and adoption of AI.
- An understanding that Guardrails adopted now will need to be clarified, iteratively as circumstances change. Perfection now is not attainable and should not be sought. We will need to act, and iterate; designing a regulatory framework is as much if not more, in this context, about developing processes as it is about rules;
- Access to appropriately broad and independent technical and socio-technical expertise.

Addressing issues around safe and responsible AI is a whole-of-government endeavour: it would be a problem to isolate **all** expertise and responsibility in a single entity, and there is much good expertise in various parts of the Commonwealth government. The Commonwealth government has already moved to appoint accountable officials, who can take on a larger role than currently envisaged.

[105] Here we refer to the separate submission on this point by J-M. Bello y Villarino and K. Weatherall.

But equally, government action, especially at the Commonwealth level at present **lacks coherence**: as Weatherall has explained in some detail elsewhere.[106] One reason for this lack of coherence could be the absence of any entity within the Commonwealth Government with the resources to develop coherent, technically- and socio-technically well-informed and hence authoritative positions, and the power, on the basis of those positions to impose stronger obligations, and/or the willingness to use that power.[107]

The imperatives identified above — the need to act; the need to develop processes and powers to update and iterate — should drive institutional and regulatory design. This could require an AI Body (Commission, Regulator). Such an entity would coordinate with agencies and regulators across government, but to be most effective, would need to be the 'final word' on safe and responsible AI, with authority across government uses of AI.

A key role of such a Body would be the further updating, development, adjustment and refinement of aspects of the regulatory framework as set out elsewhere in this submission. Broadly we imagine such a Body would play a role in:

- Developing guidance on the systems that are considered high-risk (including in consultation with domain-specific regulators);
- Developing an understanding of which GPAI systems should be considered high risk (and, if necessary, very high or systemic risk, including updating or developing any technical metrics that might be used to identify such systems);
- Adding to or refining any list of prohibited AI systems/uses; or offering pre-licensing/ registration of especially high risk systems if prohibitions are not adopted;
- Developing and refining the content of the Mandatory Guidelines;
- Helping develop, and approving, alternative guidelines for specific domains, in cooperation with domain-specific agencies or regulators and industry (secondary role in development; final say in approval);
- Helping develop AI aspects of regulatory frameworks for regulated industries where the Mandatory Guardrails do not apply (secondary role);

To fulfil these roles, we would suggest that such a body have power to issue disallowable instruments: sufficient to keep the rules current but maintaining ultimate authority in the Parliament.[108]

Beyond managing the regulatory framework, an AI body would ideally also have broader roles, in connection with broader networks of researchers, industry and civil society:

- Advising and/or determining on questions related to maturity levels as an element of any adjustment adopted for small business;
- Identifying emerging risks arising from shifts in technology;
- Scanning (and possibly commissioning) research on the impact of AI on human rights, democracy and the rule of law, as well as broader societal and environmental impacts of AI, and considering the import of such research for policy; and

[106]Weatherall, Opinionated Explainer (n 99).
[107]Compare Canada, where the Automated Decision-Making Directive is based on specific powers held by Treasury; in the US the President's Executive Order is based on the President's Power over the US federal executive. At the Commonwealth level, the Digital Transformation Agency is an advisory body: it doesn't have the power to direct people how to use AI.
[108]See generally, Commonwealth Office of Parliamentary Counsel, *Instruments Handbook* (2022).

- Being a source of knowledge, and expertise on AI, its regulation and its impacts (often an underrated role, but across complex organisations like government, knowing there is someone to turn to for both knowledge, and networks of experts beyond government, could be of real assistance to the Public Service, and members of Parliament).

These roles should be separate from (but able to cooperate/work with) the existing National AI Centre, which plays an important promotional and facilitative role (enabling the adoption and responsible use of AI).

Given these roles, such a Body will need to:

- Have strong technical expertise, and the resources to access to further external technical, sociotechnical and legal/regulatory expertise;
- Be supported with appropriately skilled staff;
- Hold a mandate to monitor and require responsible technology design, procurement and use across all parts of the Commonwealth government

Such a body may not, in the end, be a permanent fixture within government: increasingly AI is likely to be 'everywhere'. But in the current period of (possible) transition, and for the near future, there is a need for a coordinating body.