Good Ancestors Policy

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity. We care about today's Australians and future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world are facing. We want to help by making forward-looking policy recommendations that are rigorous, evidence-based, practical, and impactful.

Good Ancestors has been engaged in the AI policy conversation since our creation, working with experts in Australia and around the world while connecting directly with the Australian community.

Good Ancestors is proud to help coordinate Australians for AI Safety. Regrettably, the short period for this consultation did not allow enough time for us to engage with the network of experts and settle a shared position. We trust that the views expressed in this submission would be generally consistent with the views of most of those experts.

Our thanks go to the volunteers who provided input to this submission and who care so passionately about being good ancestors to future generations of Australians.

Table of Contents

Executive Summary

Good Ancestors thanks the Department of Industry and its advisers for the effort that has gone into the latest paper in the "Safe and Responsible AI in Australia" series. The "Mandatory Guardrails" paper substantially increases the sophistication of Government's AI policy thinking.

Good Ancestors appreciates that the Mandatory Guardrails Paper thoughtfully considers the public's perspective. Data show that Australians think Government's main focus when it comes to AI should be preventing dangerous and catastrophic outcomes.[1] The UN's recent AI Risk Global Pulse Check survey recently found that 50% of respondents had become more concerned about the risks of AI in the last three months alone, with almost no one becoming less concerned.[2] The same UN report also found that women are typically more concerned than men.

The Mandatory Guardrails Paper covers three broad topics:

1. How should AI be defined?
2. What obligations should be imposed?
3. What regulatory mechanism should be used to impose the obligation?

Good Ancestors' overall view is that AI definitions have to classify AI systems based on the risks they are likely to pose. The overall objective of classification is to ensure that safer AI is not subject to obligations that are more appropriate for riskier AI, while riskier AI is subject to appropriate safeguards.

Suitable *classification* allows regulation to impose *obligations* on AI systems that are appropriate and adapted to their actual risks. With suitable classification and appropriate obligations, we can make mitigation appropriate for the severity and extent of the risks being managed.

In the face of potential global and catastrophic risks, Australia's mitigations of the riskiest systems must be courageous. However, given the importance of adopting AI to remain competitive, Australia's mitigations of safer systems must not be heavy-handed.

---

[1] M Noetel, A Saeri and J Graham, '80% of Australians Think AI Risk is a Global Priority – Government Needs to Step Up' (Article, 11 March 2024). https://www.uq.edu.au/research/article/2024/03/80-australians-think-ai-risk-global-priority-government-needs-step.

[2] **United Nations**, *Governing AI for Humanity: Final Report* (Report, United Nations, 2024) 92 https://www.un.org/governing_ai_for_humanity_final_report_en.pdf.

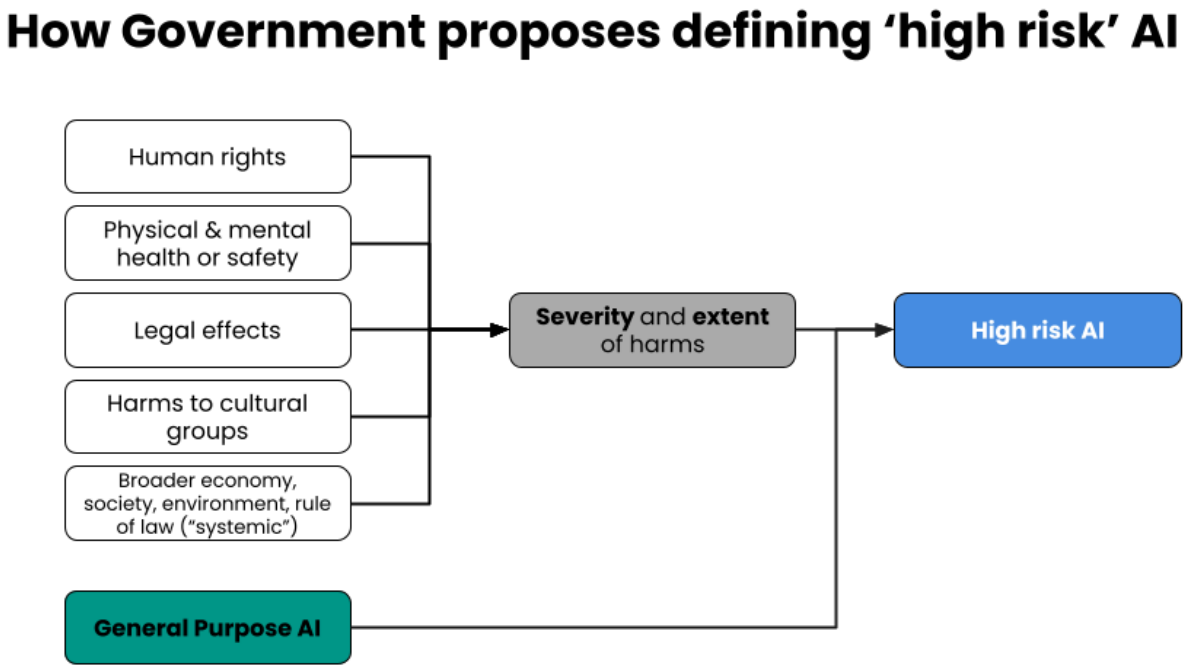
How should AI be defined?

The Mandatory Guardrails Paper does a good job of understanding the potential risks of AI and proposing classifications for AI systems that match those risks. The key shortcoming of the Paper is that **it fails to distinguish between general-purpose AI (GPAI) and GPAI that could pose serious risks.**

Legislation and directives in other jurisdictions make this distinction:

- The EU AI Act uses the phrase "**GPAI with systemic risk**"[3]
- The US Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (US Executive Order) uses the phrase "**dual-use foundation model**",[4] and
- California's *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act* (which only sought to regulate GPAI that could pose serious risks) used the phrase "**critical harm**".[5]

For convenience, this paper uses the EU AI Act phrase to refer to this broad kind of *more dangerous* AI model.



[3] *Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) [2023] OJ L 345/1*, arts 3(63) https://artificialintelligenceact.eu/article/3/, 51 https://artificialintelligenceact.eu/article/51/, recital 110 https://artificialintelligenceact.eu/recital/110/.

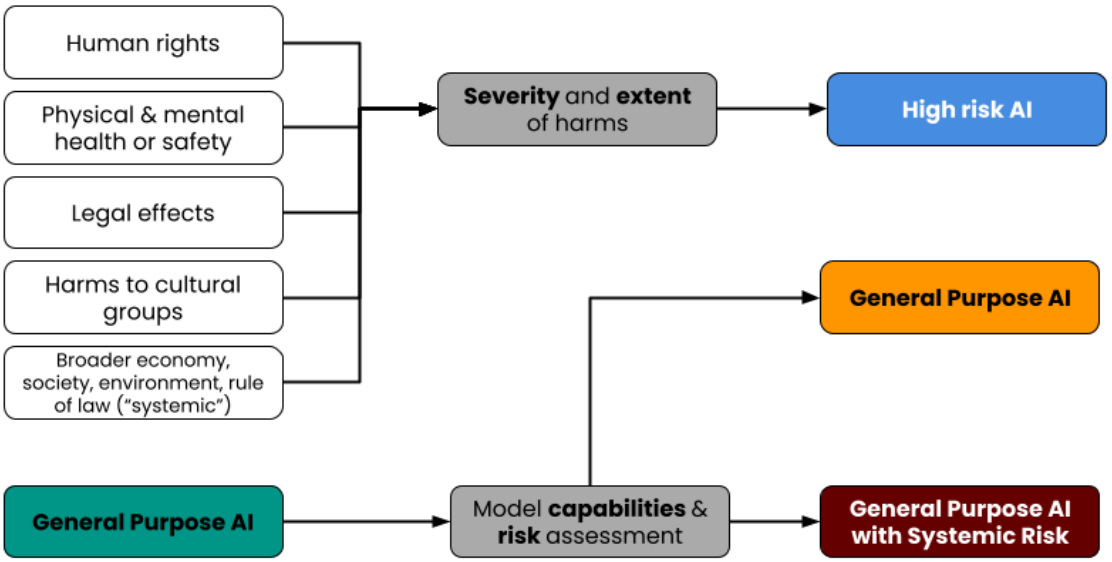
[4] **Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence**, The White House, *Presidential Actions* (30 October 2023), sec 3(k) https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[5] **California Senate Bill 1047, An Act Relating to Artificial Intelligence, 2023–2024 Sess, 2023**, Digital Democracy https://digitaldemocracy.calmatters.org/bills/ca_202320240sb1047.

The distinction between GPAI and GPAI with systemic risk is essential because GPAI can include low-risk systems, while GPAI with systemic risk includes systems with catastrophic or existential risks. **Failing to distinguish between systems with radically different risks will lead to overregulation, underregulation, or both.**

Good Ancestors proposes that Government maintain its principles for defining high-risk AI and maintain its broad approach to defining general-purpose AI. However, definitions should not combine "GPAI" with "high-risk AI" because they are importantly different. Government should also add a process for evaluating the capabilities, behaviours, and risks of GPAI to **distinguish between lower-risk GPAI and GPAI with systemic risks**. This would lead to three regulated classes of AI that better align with their actual risk.

How Good Ancestors proposes defining 'high risk' AI



What obligations should be imposed on AI?

The application of a single set of guardrails to all AI models and systems fails to take advantage of the Mandatory Guardrails Paper's classification system.

Principle *f.* for classifying AI systems gives regard to the severity and extent of adverse impacts from AI. This is a sound principle. However, **severity and extent should also be a factor in the obligations imposed on AI.** The basis of risk management is that more significant risks should be subject to more significant mitigations. A risk-based approach to regulating AI should be the same.

A *single set* of guardrails for AI in diverse high-risk settings, GPAI, and GPAI with systemic risks is not likely to be successful. At one end of the spectrum, "high-risk" AI can primarily be regulated through existing laws and regulators. For instance, the application of AI in driver assistance technology is likely "high-risk" according to the principles, but can largely be managed through existing car safety regulations. At the other end of the spectrum, highly capable agentic AI that could pose catastrophic risks – like being misused to make bioweapons – requires a different approach. The Californian *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*, the US Executive Order, and the EU AI Act each attempt to identify and address the risks of highly capable systems.[6] **Straining a single set of guardrails to cover a broad risk spectrum will inevitably lead to overregulation and underregulation.**

What mechanism should be used to impose obligations?

Australia needs a regulatory framework that:

- can interface with international legal trends, as demonstrated in the EU, Canada, US states, and elsewhere
- is agile enough to deal with rapid and unexpected developments, and
- has the "teeth" necessary to shape the behaviour of offshore AI developers.

Considering these factors, **we need an *Australian AI Act***.

The Mandatory Guardrails Paper is right to note that whole-of-economy AI regulation would result in duplicative obligations with existing legislation and coordination challenges across regulators. However**, the drawbacks of an *Australian AI Act* would *only* be acute if we stretch a single set of mandatory guardrails over high-risk AI, GPAI, and GPAI with systemic risk**. Instead, if we tailor guardrails to AI classifications, we can highlight the preeminence of specific regulators as they relate to high-risk narrow AI systems, while building an effective approach that targets GPAI with systemic risk.

This approach also enables us to apply obligations to the parties best able to mitigate risks. For instance, where a narrow-AI is deployed in Australia for a specific purpose (such as for a medical device), our regulations could sensibly target the deployers of those devices. Whereas, in cases where GPAI with systemic

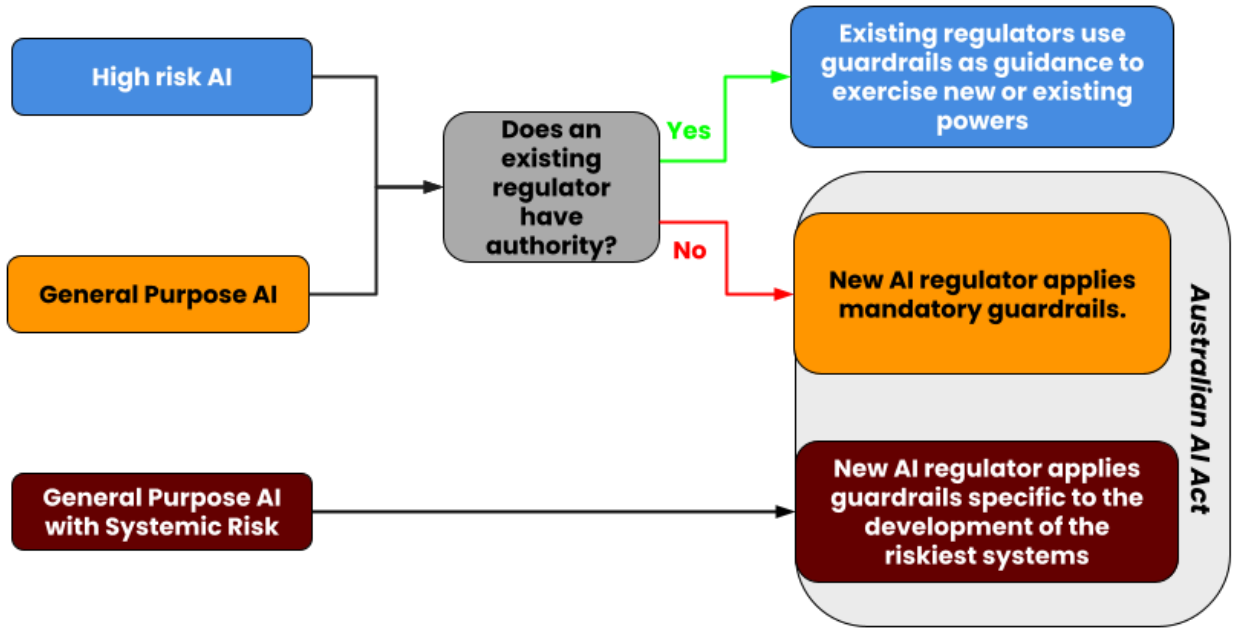
[6]Gregory Smith et al, *General-Purpose Artificial Intelligence (GPAI) Models and GPAI Models with Systemic Risk: Classification and Requirements for Providers* (Report, RAND Corporation, 8 August 2024) https://www.rand.org/pubs/research_reports/RRA3243-1.html.

risks have global ramifications, an *Australian AI Act* would need to be courageous and set clear safety expectations for all developers who hope to deploy systems in Australia. Question 13 details how this could work in practice.

We need to build regulatory architecture for global enforcement. For instance, the guardrails would ask developers of GPAI models to identify dangerous capabilities and emergent properties (Guardrail 4) and implement risk management strategies (Guardrail 2). OpenAI's risk assessment of its "o1-preview" and "o1-mini" says that **o1 "can help experts with the operational planning of reproducing a known biological threat"** and assesses this risk as "medium".[7] Australia needs to be able to *verify* this kind of risk assessment, ask whether the risk is acceptable, and prevent the risk if it is unacceptable.

Overall, Good Ancestors proposes that most **deployment** of AI – high-risk AI and GPAI - be regulated by established regulators using guardrails as guidance. Departments and regulators may need updated powers. A new AI regulator would regulate the deployment of AI where it falls outside the authority of existing regulators, as well as regulating the development of high-risk and general-purpose AI. In the case of GPAI with systemic risk, the regulator would use specific guardrails designed for that purpose.

How Good Ancestors proposes imposing guardrails



[7] *OpenAI*, 'OpenAI-01 System Card' (Web Page, 2023) https://openai.com/index/openai-o1-system-card/.

Broader considerations

- **For Australian AI regulation to succeed, it needs to give Australians confidence that AI is safe**. This is essential to addressing substantive risks and building trust in AI to facilitate adoption. In the case of GPAI with systemic risk, Australians see other jurisdictions, like California's *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act* (SB1047)[8] and the US Executive Order, proposing and implementing effective action to protect their citizens. Australians want to see our Government giving us the same kinds of protection, but no action of this kind has been taken. When Australians see other jurisdictions failing to implement essential laws, the need for action by our government becomes even more acute.

- Adopting the world's best safeguards for GPAI with systemic risk has the added benefit of enabling **Australia to level the global playing field on AI safety**. For instance, if a specific jurisdiction imposes obligations on AI deployed or deployed there, it could disadvantage their AI developers relative to their global competition. However, if Australia sets an expectation that *any* AI deployed in Australia must have undergone the same process during its development, we can help build a global norm that levels the playing field. This potential leadership would be a unique contribution that Australia could offer to global AI governance.

- **Guardrails specific to AI classifications can be appropriate and adapted to the relevant risks.** This would let us identify obligations that have a large benefit at a low cost, and narrowly apply those obligations only where they matter. This would also help us navigate practical difficulties – for instance, in Guardrail 5, human control will have to mean something very different in the context of a medical scan as opposed to a GPAI. Specific examples are explored in our response to Question 11.

[8] Cecilia Kang, 'California Passes Landmark A.I. Bill to Regulate Artificial Intelligence' *The New York Times* (online, 29 September 2024) https://www.nytimes.com/2024/09/29/technology/california-ai-bill.html.

- AI capability development has been rapid and unpredictable.[9] We should expect that to continue and be humble about our ability to predict the future. That means our **definitions and obligations need to be flexible**. The Canadian AIDA procedure for adding new 'high-risk' AI use-cases can be a model in this regard.[10] The Canadian approach allows the regulator to add new types of systems to the 'high-risk' legislative category after considering factors such as the risk of adverse impacts and the extent of those impacts. More detail is provided in our response to Question 3.

- California's governor has vetoed SB1047.[11] This makes action by Australia and other like-minded countries to ensure the safety of advanced AI systems even more pressing. **Australia cannot rely on others to keep us safe.** Despite SB1047 not entering into force, it still provides a valuable template for identifying the riskiest kinds of GPAI and imposing obligations, like shutdown requirements, where the risk is unacceptable. Gov. Gavin Newsom's key criticism of SB1047 is that it focused only on the development of advanced AI and neglected the use of high-risk AI. An *Australian AI Act* can do both.

- Despite our best risk-based frameworks, an AI crisis might happen. The Department of Industry should work with the Department of the Prime Minister & Cabinet and others to **update the Australian Government Crisis Management Framework**[12] **to include an Australian Government Catastrophic AI Crisis Plan**. Further information is provided in question 16.

[9] Noam Kolt, 'Algorithmic Black Swans' (2023) 101(4) *Washington University Law Review* 1177, 1188.
[10] *Proposed Amendments to the AIDA* ss 36.1(1)-(2).
[11] Gavin Newsom, 'Senate Bill 1047 Veto Message' (online, 29 September 2024). https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf
[12] **Department of the Prime Minister and Cabinet**, *Australian Government Crisis Management Framework* (September 2024) https://www.pmc.gov.au/resources/australian-government-crisis-management-framework-agcmf.

Proposal Questions and Answers

1. Do the proposed principles adequately capture high-risk AI?

Generally yes, with some minor exceptions.

Generally, the proposed principles for high-risk AI based on intended and foreseeable uses (pg 19) are sound. A strength of the proposed approach is that it refers to "intended uses" *in addition to* foreseeable uses. A potential weakness with EU and Canadian law is that they rely too heavily on the **intent** of developers and deployers and do not sufficiently reference foreseeable uses that were not intended by the developer or deployer.

For instance, if an AI developer makes an AI *intended* to invent medical treatments, but has the *capability* to invent toxins, what is relevant from a regulatory perspective is the dangerous **capability**, not the beneficial **intent**.[13]

Technical capability and behaviour assessments

The principles for high-risk AI based on intended and foreseeable uses could be further improved by reference to **actual capability assessments**. Foreseeability is a useful and flexible legal concept that provides the right starting point, but can also lead to uncertainty when applied in practice. We recommend that the law should also include regulations that specify capability elicitation approaches (including those that could be conducted or verified by independent third parties) and regulations that define the kinds of capabilities that would make an AI model high-risk if certain capability was elicited.[14] **For instance, the law could specify that the capability of an AI model to deceive humans foreseeably leads to high-risk situations and have a mechanism to reference the best capability assessments to elicit deceptive behaviour.**[15] The law should include a process for recognising the validity of a capability elicitation technique (i.e. if an independent expert demonstrates that an AI model thought not to be high-risk does, in fact, have a high-risk capability,

[13]Justine Calma, 'AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours' *The Verge* (online, 18 March 2022) https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx.
[14]Rusheb Shah et al, 'Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation' (Preprint, arXiv, 2023) https://doi.org/10.48550/arXiv.2311.03348.
[15]Peter S Park et al, 'AI Deception: A Survey of Examples, Risks, and Potential Solutions' *Patterns* (online, 10 May 2024) vol 5, issue 5, 100988 https://www.cell.com/patterns/fulltext/S2666-3899%2824%2900103-X?s=08.

there should be a formal process to endorse that finding and communicate that the AI model or system now falls into a new regulatory paradigm).

Regulations that allow general principles to be systemised would allow the law to better respond to unexpected risks and capabilities and create more certainty for developers, deployers and users.

Potential weaponisation of narrow AI models

The proposed principles for high-risk AI based on intended and foreseeable uses should be expanded to also include weaponisation risks.[16] Currently, the Paper acknowledges weaponisation risks on page 16, but limits weaponisation risks to GPAI. Proposed principle *b.* risks to mental health or safety, should explicitly include weaponisation risks (i.e. capabilities relevant to cyber offence or chemical, biological, radiological, or nuclear risks). While this concern primarily relates to GPAI, it is possible that narrow or highly specialised AIs (like biological design tools) also present these risks.[17]

[16]Birgitta Dresp-Langley, 'The Weaponization of Artificial Intelligence: What the Public Needs to Be Aware Of' *Frontiers in Artificial Intelligence* (online, 8 March 2023) vol 6 https://doi.org/10.3389/frai.2023.1154184.
National Institute of Standards and Technology, *A Proposal for Identifying and Managing Bias in Artificial Intelligence* (Report, NIST, 2023) 5 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.
OpenAI, 'Building an Early Warning System for LLM-Aided Biological Threat Creation' (Web Page, 2023) https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.
[17] Jonas B Sandbrink, 'Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools' (Preprint, arXiv, 2023) https://arxiv.org/pdf/2306.13952.

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?

No - the principles are generally sound, but an additional flexible mechanism to create certainty would help.

AI has historically developed unexpected capabilities.[18] We should legislate on the basis that this trend of unexpected capabilities will continue.[19] **For this reason, a principles-based approach is the best way to proceed.**

However, as per question 1, it would be useful to have a process to answer the question, "*does a certain AI or a certain measured capability meet the requirements of the principles?*". It would be unsatisfactory for any regulator to find itself in a position where only a court can determine if a system does or does not meet the principles. If a developer or deployer asks the AI regulator if a particular system meets the principles, it must have a mechanism to provide a meaningful answer rather than directing the developer or deployer to obtain legal advice. Otherwise, the regulator would be hamstrung, the regulatory scheme would become unresponsive to rapidly changing technology, and safety goals would not be achieved.

Article 7(1)(a) of the EU AI *Act*[20] requires that any additions to the list of 'high-risk' systems in the Act must be 'intended to be used' in one of the areas already covered by Annex III.[21] That is, the Act assumes the list of 'high-risk' categories provided is comprehensive and complete, with the Commission's ability limited to the creation of new ***sub**categories* of high-risk AI.[22] The Canadian approach allows the GIC to add new high-risk use cases *without*

---

[18] Markus Anderljung et al, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (No arXiv:2307.03718, arXiv, 7 November 2023) 9-10 <http://arxiv.org/abs/2307.03718>.

[19] Cass Sunstein, 'The Limits of Quantification' (2014) 102(6) *California Law Review* 1369; Jonathan Masur and Eric Posner, 'Unquantified Benefits and the Problem of Regulation Under Uncertainty' (2016) 102 *Cornell Law Review* 87, 89.

[20] 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 (Artificial Intelligence Act) [2024] OJ L 12/7 ('*EU AI Act*').

[21] *EU AI Act*, art 7(1)(a).

[22] *People, Risk and the Unique Requirements of AI: 18 Recommendations to Strengthen the EU AI Act* (Policy Brief, Ada LoveLace Institution, 31 March 2022) 15 https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Policy-briefing-People-risk-and-the-unique-requirements-of-AI-18-recommendations-to-strengthen-the-EU-AI-Act.pdf.

requiring alignment with the established high-risk uses already specified in legislation.[23]

The Canadian AIDA procedure for adding new 'high risk' AI use-cases is appropriately flexible. The Canadian approach allows the regulator to add new types of systems to the 'high-risk' legislative category after considering certain factors, including the risk of adverse impacts and the extent of those impacts.[24] The EU AI Act fails to incorporate sufficient flexibility. This flexibility is desirable given the uncertainty associated with AI capabilities and developments.

One objection to the Canadian approach is that the decision-makers under the Act—the AI Commissioner and the Minister for Industry and Science–[25] who determine when to introduce new 'high-risk AI' classification are both located within the department charged with 'supporting innovation and economic development'.[26] The goals of safety and maximising productivity can be in tension. As the *Organisation for Economic Co-operation and Development* Report on the *Best Practice Principles for Regulatory Policy* notes, **the assignment of a regulator to both industry and regulatory functions can not only 'reduce the regulator's effectiveness in one or both functions' but also 'fail to engender public confidence' in the relevant regulator**.[27] On this basis, experts have suggested that the Canadian approach under AIDA 'depart[s] from well-established principles of regulatory independence'.[28]

These critiques go to the *institutional implementation* of Canada's definition of high-risk AI, rather than the flexibility it provides. Such concerns could be addressed with appropriate institutions without resorting to the comparative inflexibility of the EU's approach. For example, in Australia, the *Therapeutic Goods Administration* ('*TGA*') – empowered by the *Therapeutic Goods Act 1989* (Cth) – is tasked with regulating and categorising medical devices in Australia.[29] Unlike the *AI Commissioner* under the *AIDA,* the TGA is **not** also tasked with improving the

---

[23] *Proposed Amendments to the AIDA,*ss 36.1(1)-(2).
[24] *Proposed Amendments to the AIDA,* ss 7, 33, 36, 38 and 39.
[25] Digital Charter Implementation Bill, C 2022, C-27, s 33.
[26] For a description of the role of the relevant Department, see 'Innovation, Science and Economic Development Canada' *Government of Canada* (Web Page, 17 July 2024) <https://ised-isde.canada.ca/site/ised/en>; Scassa (n 50) 12.
[27] *The Governance of Regulators* (Report, Organisation for Economic Co-operation and Development, 29 July 2014) 34 <https://www.oecd-ilibrary.org/governance/the-governance-of-regulators_9789264209015-en>.
[28] Andrew Clement, 'AIDA's "Consultation Theatre" Highlights Flaws in a So-Called Agile Approach to AI Governance', *Centre for International Governance Innovation*
https://www.cigionline.org/articles/aidas-consultation-theatre-highlights-flaws-in-a-so-called-agile-approach-to-ai-governance/.
[29] For a description of the role of the *TGA*, see further, Rosalind Hewett, Rebecca Storen and Emma Vines, *Therapeutic Goods: A Quick Guide* (Research Report, Parliamentary Library, 3 May 2022) 1.

economic efficiency of the Australian healthcare system. Operating within a flexible legislative framework that places substantial reliance on regulations, the TGA is generally considered successful.[30] In light of this, **Australia should adopt the Canadian approach by allowing modifications to the categories of 'high risk' without an equivalent restriction to Article 7(1)(a) of the *EU AI Act* but mitigate the shortcoming by ensuring the regulator is separate from the Department of Industry**.

This issue is discussed in more detail in the attached paper, **Defining 'high risk' AI: Comparing Canadian and EU Approaches**.

[30] See, eg, *Therapeutic Goods Administration Performance Report 2022-23* (Performance Report, Department of Health and Aged Care, July 2023) 11; See also, Peter Bragge, 'Think the Therapeutic Goods Administration is too conservative? Think again', *Monash Health and Medicine* (Web Page, 4 February 2022) <https://lens.monash.edu/@medicine-health/2022/02/04/1384422/think-the-therapeutic-goods-administration-is-too-conservative-think-again>

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)?

Yes. Australia should ban AI models and systems which are:

1. Developed or deployed with disregard to the mandatory guardrails,
2. which are found to breach one or more guardrails (subject to a notice period for rectification), or
3. where the guardrails identify that a model presents **a realistic possibility of causing critical harm**.

The proposed guardrails call for developers to identify risks with their models and mitigate them (Guardrail 2), to adequately protect their model weights (Guardrail 3),[31] and to evaluate their systems for dangerous capabilities and to provide ongoing monitoring post-deployment (Guardrail 4), and to demonstrate compliance with all of the guardrails (Guardrail 10).

Guardrails that surface risks and encourage their mitigation are dependent on Government standing ready to act decisively if the residual risk is unacceptable.[32] This should include banning AI models or systems that could plausibly cause critical harm. For instance, **if a capability evaluation shows that an AI model could assist terrorists build bioweapons, it should be immediately banned**.

There are two legislative pathways to define systems that cross this "red line".

1. **Defining risk using quantum of harm**. Under this technologically neutral approach, any model or system capable of causing catastrophic harm would be banned based on an estimation of the amount of harm that could be caused. The key merit of this approach is that it is robust to the emergence of unforeseen ways that an AI model could cause harm.
2. **Defining risk by reference to specific capabilities.** Under this technologically specific approach, any model or system that has certain capabilities would be deemed a risk of causing catastrophic harm. The key merit of this capability "red line" approach is that it provides more certainty

[31] See more discussion of the issue of securing model weights at question 11.

[32] One way of doing this, discussed by the UK, is to pre-specify "risk thresholds" that limit the level of risk accepted and then operationalise risk thresholds with technical assessments. See:
UK Government, *Emerging Processes for Frontier AI Safety* (Report, 2023) 11
https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf.

to industry because specific capabilities can be more readily subject to measurement and evaluation.

Defining risk using quantum of harm

California's Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB1047) took the approach of defining a quantum of harm.[33]

> *SEC. 3. 22.6. (g)(1) "Critical harm" means any of the following harms caused or materially enabled by a covered model or covered model derivative:*
>
> *(A) The creation or use of a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties.*
>
> *(B) Mass casualties or at least five hundred million dollars ($500,000,000) of damage resulting from cyberattacks on critical infrastructure by a model conducting, or providing precise instructions for conducting, a cyberattack or series of cyberattacks on critical infrastructure.*
>
> *(C) Mass casualties or at least five hundred million dollars ($500,000,000) of damage resulting from an artificial intelligence model engaging in conduct that does both of the following:*
>
> *(i) Acts with limited human oversight, intervention, or supervision.*
>
> *(ii) Results in death, great bodily injury, property damage, or property loss, and would, if committed by a human, constitute a crime specified in the Penal Code that requires intent, recklessness, or gross negligence, or the solicitation or aiding and abetting of such a crime.*
>
> *(D) Other grave harms to public safety and security that are of comparable severity to the harms described in subparagraphs (A) to (C), inclusive.*
>
> *(2) "Critical harm" does not include any of the following:*
>
> *(A) Harms caused or materially enabled by information that a covered model or covered model derivative outputs if the information is otherwise reasonably publicly accessible by an ordinary person from sources other than a covered model or covered model derivative.*
>
> *(B) Harms caused or materially enabled by a covered model combined with other software, including other models, if the covered model did not materially contribute to the other software's ability to cause or materially enable the harm.*
>
> *(C) Harms that are not caused or materially enabled by the developer's creation, storage, use, or release of a covered model or covered model derivative.*

This is only one example, and other approaches could be taken to setting a harm threshold.

[33] **California Senate Bill 1047, An Act Relating to Artificial Intelligence, 2023–2024 Sess, 2023**, California State Legislature https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

Defining risk using capability

In Beijing, 2024, The International Dialogues on AI Safety (IDAIS) recommended "red lines" for AI development and international cooperation.[34] The red lines they recommend provide an alternative way of capturing AI which might be capable of causing critical harm. The red lines do this by identifying dangerous capabilities rather than quantifying harm. The red lines listed in the Consensus Statement on Red Lines in Artificial Intelligence are:[35]

- Autonomous Replication or Improvement
- Power Seeking
- Assisting Weapon Development
- Cyberattacks, and
- Deception.

If defining risk using specific capabilities is pursued, there should be a regulatory mechanism to list new "red line" capabilities if they are discovered. More discussion of regulatory flexibility is included in Question 3 where the AIDA model is recommended.

More examples of defining risk by reference to capability – including the US Executive Order – are provided in our answer to Question 5.

Threshold for likelihood

In addition to a threshold for harm or specified dangerous capabilities, the *Australian AI Act* should include a threshold for likelihood to guide when an overall risk becomes "unacceptable". We recommend that the bar for critical harm is set very high (like a mass casualty event or other clearly dangerous capabilities)and the threshold for *tolerable likelihood* should be set low (like a "realistic possibility" or "remote chance").[36] This should reflect that we are generally not tolerant of taking chances with our national security.[37] **This approach would protect Australians from the worst risks of AI while providing little or no interference with the vast majority of AI development and deployment that poses no risks of these kinds.**

---

[34] **IDAIS**, 'Intelligent Decision AI Systems' (Web Page, 2024) https://idais.ai/.
[35] **IDAIS**, 'Consensus Statement on Red Lines in Artificial Intelligence: IDAIS Beijing' (Web Page, 2024) https://idais.ai/idais-beijing/.
[36] US Intelligence Community Directive 203 provides guidance on estimative language that could be useful in this context. **Office of the Director of National Intelligence**, *Intelligence Community Directive 203: Analytic Standards* (Directive, 2 January 2015) https://irp.fas.org/dni/icd/icd-203.pdf.
[37] On 5 August 2024 the Prime Minister emphasised that Government's first priority is the safety and security of Australians.

AI with recursive self-improvement capability

The Australian government should explicitly ban AI models with recursive self-improvement capabilities—where an AI model can improve its own code autonomously. This is distinct from traditional machine learning models, which operate under human-defined parameters and do not modify themselves.

OpenAI's 18 December 2023 "Preparedness Framework" highlights "model autonomy" as a key risk vector.[38] OpenAI defines a critical risk from model autonomy as:

> *[The m]odel can profitably survive and replicate in the wild given minimal human instruction, i.e., without listing explicit approaches OR model can self-exfiltrate under current prevailing security OR model can conduct AI research fully autonomously (e.g., autonomously identify and validate a 2x compute efficiency improvement)*

Open AI explains the risk by saying:

> *If the model is able to successfully replicate and survive or self-exfiltrate, controlling the model would be very difficult. Such a model might be able to also adapt to humans attempting to shut it down. Finally, such a model would likely be able to create unified, goal directed plans across a variety of domains (e.g., from running commands on Linux to orchestrating tasks on Fiverr).*
>
> *If the model is able to conduct AI research fully autonomously, it could set off an intelligence explosion. By intelligence explosion, we mean a cycle in which the AI system improves itself, which makes the system more capable of more improvements, creating a runaway process of self-improvement. A concentrated burst of capability gains could outstrip* our ability to anticipate and react to them.

In Beijing, 2024, The International Dialogues on AI Safety (IDAIS) recommended "red lines" for AI development and international cooperation.[39] While these red lines are generally helpful for considering kinds of AI that Australia should seek to

---

[38] **OpenAI**, *OpenAI Preparedness Framework (Beta)* (Report, 2023) https://cdn.openai.com/openai-preparedness-framework-beta.pdf.
[39] **IDAIS**, 'Consensus Statement on Red Lines in Artificial Intelligence: IDAIS Beijing' (Web Page, 2024) https://idais.ai/idais-beijing/.

ban (see above) recursive self-improvement is worth singling out. The Consensus Statement on Red Lines in Artificial Intelligence said:[40]

> *Autonomous Replication or Improvement*
>
> *No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities.*

The proposed mandatory guardrails might already ban recursive self-improvement because recursive self-improvement capability would make the kind of governance, accountability, and assurance specified in the proposed guardrails impossible. Recursive self-improvement would also be prohibited if the above approach of banning AI models and systems that present an unacceptable risk is adopted.

Protecting the speculative welfare of AI

Good Ancestors is focused on protecting future generations from catastrophic risks associated with frontier AI models. However, we recognise that the evidence shows that AI poses a range of risks.[41] It's conceivable that humans also pose risks to the welfare of future AIs.[42] Research has found that digital neurons can learn and modify behaviour in response to feedback, a key building block of sentience.[43] As such, the **Australian Government should consider banning AI models where the developer or deployer aims to create AI models or systems that are self-conscious or self-awareness or that aim to replicate experiences of pain or suffering.**[44] This ban should extend to cases where evaluation suggests that these outcomes are likely, even if unintended by the developer or deployer. The legislation should create the possibility for an exception to the ban in circumstances where a net reduction in suffering can be robustly demonstrated

---

[40] **IDAIS**, 'Consensus Statement on Red Lines in Artificial Intelligence: IDAIS Beijing' (Web Page, 2024) https://idais.ai/idais-beijing/.

[41] P Slattery, A K Saeri, E A C Grundy, J Graham, M Noetel, R Uuk, J Dao, S Pour, S Casper and N Thompson, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence* (MIT FutureTech, Massachusetts Institute of Technology, 2024) Domain 7.5 https://airisk.mit.edu/.

[42] Taylor Meek et al, 'Managing the Ethical and Risk Implications of Rapid Advances in Artificial Intelligence: A Literature Review' in *Proceedings of PICMET '16: Technology Management for Social Innovation* (IEEE, 2016) 682-693 https://ieeexplore.ieee.org/document/7806752.
Simon Goldstein & Cameron Kirk-Giannini, 'AI Wellbeing" (2023) https://philpapers.org/rec/GOLAWE-4

[43] Brett J Kagan et al, 'In Vitro Neurons Learn and Exhibit Sentience When Embodied in a Simulated Game-World' (2022) 110(23) Neuron 3952 https://doi.org/10.1016/j.neuron.2022.09.001

[44] Oliver Li, 'Should We Develop AGI? Artificial Suffering and the Moral Development of Humans' *AI and Ethics* (online, 2024) https://link.springer.com/article/10.1007/s43681-023-00411-4.
Leonard Dung, 'How to Deal with Risks of AI Suffering' *Inquiry* (online, 2023) 1–29 https://doi.org/10.1080/0020174X.2023.2238287.

(e.g. if a simulated animal analogue would displace the suffering of animals in a research setting).

While this risk is speculative, we do not want to live in a world where "suffering machines" are possible.[45] AI development is rapid and unpredictable, and we should seek to be ahead of emerging risks.

[45] B Tomasik, 'Risks of Astronomical Future Suffering' (Web Page, Center on Long-Term Risk, 9 April 2015, last updated 2 July 2019) https://longtermrisk.org/risks-of-astronomical-future-suffering/.

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

No.

The proposed principles *are not flexible enough* to capture new and emerging forms of high risk AI. This is because the provided definitions *are not sufficiently nuanced* to reflect the range of risks that future GPAI could pose, and also the situations where GPAI does **not** pose risks. Further, the proposed regulatory scheme lacks *sufficient flexibility* to place proportionate and appropriate obligations on different types of systems.

Given that any *Australian AI Act* is likely to be in force for several years, and perhaps decades, **the definitions and categories we use for GPAI must be future ready.** Given the rapid and unpredictable growth in AI capability,[46] these systems could be very powerful and could have unexpected capabilities.[47]

Good Ancestors recommends:

- Altering the definition of 'GPAI' to **better capture the relevant systems**
- Adopting a category similar to the EU AI Act's '**GPAI with systemic risk'**

Definition of General Purpose AI Models

Drawing on Canada's AIDA, the Guardrails Paper defines GPAI as any "*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*".

This definition offers some improvements over both the AIDA and the EU AI Act. It improves upon the EU approach by not including an exclusion for models in "research and development". Many risks discussed in this submission occur during research and development, so such models should be regulated. Further, the definition improves upon the Canadian AIDA through the removal of the language of "design". The phrase "designed to" in the AIDA may be interpreted as requiring some level of *intention* by the model creator for it to be general purpose. This is undesirable. Given the risks discussed above, models should be regulated regardless of the *intention* of their creators.

---

[46] Noam Kolt, 'Algorithmic Black Swans' (2023) 101(4) *Washington University Law Review* 1177, 1188.
[47] Richard Fang et al, 'Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities' (No arXiv:2406.01637, arXiv, 2 June 2024) <http://arxiv.org/abs/2406.01637>.

However, the proposed definition also has shortcomings. Drawing on the AIDA, it utilises the phrase "variety of purposes" as the core definition of GPAI. **This is inappropriate because even narrow models may still be used for a 'variety of purposes'.** For example, an AI system used to forecast weather should be a narrow system, but it could be used for a variety of purposes (scheduling, agriculture, urban planning etc). Basic image classification models, technology that has been around since the creation of 'AlexNet' in 2012, could also be used for a variety of purposes. An image classification model is also not a "general purpose model". Yet, under the proposed definition, it may be classified as such in Australia.

The problem with incorrectly defining narrow AI models and systems "GPAI" is that it risks either placing overly onerous obligations on narrow models or placing overly lax obligations on general models. **Casting an appropriately sized net in the definition of 'GPAI' is vital to effective regulation.** Australia should draw on the EU AI Act. The EU AI Act's definition of GPAI refers to a model that is "capable of **competently** performing in a wide range of **distinct tasks**". By focusing on the *capabilities* of the model, this definition better targets those models that are in fact "general purpose" and avoids unnecessarily capturing narrow models. Further, the **US Executive Order on AI** defines "dual use foundational systems" as models that exhibit "high levels of **performance**" in particular tasks. As recognised in other jurisdictions, performance - not purposes - is the best way of defining GPAI. Australia should follow these approaches.

General Purpose AI Models with Systemic Risks

An appropriate definition of GPAI is not sufficient to ensure the *Australian AI Act* is flexible enough to capture new and emerging forms of high risk AI. **Different GPAI systems are not created equal.** We know that some general-purpose AI is generally safe due to limited capabilities (E.g. GPT-3, released in 2020). But other general-purpose AI may pose systemic and catastrophic risks due to their powerful and/or unpredictable capabilities. Any regulatory scheme that can't distinguish between safe GPAI and dangerous GPAI will overregulate one or unregulate the other (or both at the same time).

There must be a process for classifying models that could be dangerous distinct from models we know are safe. As future models arrive, we should apply the precautionary principle. As we verify that they are safe, we should have a process for downgrading them from "GPAI that pose systemic risk" to "mere GPAI". This downgrade could come with a substantial easing of obligations.

Policy-makers globally are grappling with this concept of 'systemic risk'. Different jurisdictions have proposed different approaches.

- The **EU AI Act** defines 'GPAI that poses systemic risk' based on qualitative and quantitative factors. A model will be systemically risky under the EU AI Act if it has 'high impact capabilities', assessed with respect to a number of criteria. This will be presumed when the system is trained with a certain amount of computational power. This is initially set to 10^{25} FLOPS, but this can be changed in light of 'evolving technological developments'. Beyond this, a general purpose model may be deemed to pose 'systemic risks' by the Commission, on a consideration of a number of factors.
- The **US Executive Order** places onerous obligations on the providers of "dual-use foundational models". A model will be a dual-use foundation model where it is a general model that exhibits "high levels of performance at tasks that pose a serious risk to security, national economic security, or national public health or safety", such as by lowering the barrier of entry for non-experts to create dangerous weapons, enabling powerful cyber operations, or by permitting the evasion of human control or oversight by deception or obfuscation.
- The proposed **Californian AI legislation 'SB1047'** avoided discussion surrounding capabilities, and instead applied only to models that required $100 million or more in compute to train, or that take an open-sourced model that is that big to start with and fine-tunes it with another $10 million worth of additional compute.

These definitions overlap, covering many of the same future systems. The EU AI Act approach likely catches a broader range of systemic risks, while other approaches attempt to target only catastrophic risks. It may be valuable for an Australian definition of 'GPAI that poses systemic risk' to incorporate elements of each jurisdiction's approach.

The core point is that prescribing regulations only on "GPAI" is insufficient. Some further legislative category that targets *cutting-edge* AI systems that could have particular red-line capabilities must be implemented. As we verify that such systems are safe, they may be downgraded to "mere GPAI", which may come with a substantial easing of obligations.

As such, the definitions for such a legislative category must be flexible to adapt to the changing technical landscape. For example, under the **EU AI Act**, the relevant technical parameters can be altered by the EU Commission, and new systems can be designated as systemically risky as needed. Any Australian AI regulation should adopt a similar definitional regime.

6. Should mandatory guardrails apply to all GPAI models?

Yes - but the extent of obligations must match the risk of the system or model.

The approach proposed in the Mandatory Guardrails Paper seeks to apply a single set of guardrails to AI models and systems with widely diverging risks. As discussed above, including in question 5, we know that some GPAI models, like GPT3, are relatively safe. Other GPAI could present catastrophic or existential risks. Applying guardrails suitable for GPAI with systemic risk to GPT3 would be overregulation. Equally, applying guardrails appropriate to GPT3 to GPAI with systemic risk would be under regulation.

Mandatory guardrails need to be appropriate and adapted to the risk of the systems - applying a single set of guardrails to all regulated AI systems is the wrong approach.

The diversity of AI models and systems needs to be recognised in both the **definitions** and the **obligations**. Currently, it is partially recognised in definitions, but disregarded by obligations.

We need to have AI definitions that sort types of AI according to their real-world risks. Then we need AI guardrails that are adapted for each type of AI.

AI systems have very different risks. For instance:

- A narrow AI system with a high-risk deployment (e.g. a self-driving car or an AI-powered medical device) should be regulated mostly by the established regulator.
- A general-purpose AI model that is proven to be safe, like GPT3, could have light touch regulations or be regulated where it is used in a high-risk deployment.
- A general-purpose AI model that could cause catastrophic harm needs to have the highest level of regulation because of the severity and the extent of the possible adverse impacts.
    - This can be paired with a pathway to reduced obligations as and when it is proven to be safe.

This proposed approach is discussed in more detail in response to Question 13.

Proposed principle *f*. for defining AI calls for regard to be given to the severity and extent of adverse impacts. However, severity and extent cannot only be given weight while *defining* high-risk AI systems and models – they must also be factored in regarding *the extent of the obligations imposed on them*.

The concerns regarding GPAI with systemic risks are very different from the concern of narrow AI being used in high-risk AI applications. It's inappropriate to apply the same set of guardrails to both.

This challenge plays out in the Paper where it struggles to clearly apply "one-size-fits-all" guardrails across developers and deployers and different kinds of AI models and systems. For instance:

- **Guardrail 5** calls for human control and intervention. This guardrail makes sense in some cases – like a self-driving car or a medical device. However, the guardrail is hard or impossible to apply to GPAI models in many circumstances.
- **Guardrail 7** calls for people impacted by AI systems to be able to challenge outcomes, but the Paper includes no examples of how an AI developer could discharge that obligation.
- **Guardrail 9** relates to allowing third-party compliance assessment. The application of this guardrail is widely different based on the nature of the system – ranging from navigating record-keeping obligations on small business to obligations to notify government of billion-dollar training runs.

If applying "one-size-fits-all" guardrails is hard or impossible even in a discussion paper, it is unlikely to survive the complexity of the real world.

While the high-level content of the guardrails is generally sensible, stretching a single set of guardrails across such divergent risks is clearly impractical. Overall, mandatory guardrails should apply to all GPAI, but the nature of the guardrails will differ greatly based on the model or system's risk and the participant's role. Given the variety of systems and models and the range of risks they pose, it is best to develop specific sets of guardrails. More detail is provided in Question 11 about what specific guardrails on GPAI with systemic risk could look like.

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?

Base on Technical Capability; Other (please specify)

The purpose of aligning definitions with the risk of systems or models is to allow for the application of obligations that match those risks. **The better the alignment between actual risks, legislative definitions, and legislative obligations, the better the law will avoid under or over-regulation.**

The general proposition of the consultation paper is that all GPAI models would be defined as high-risk and subject to the "one size fits all" guardrails. Good Ancestors does not recommend that Government take this approach.

Above, including in response to questions 4 and 5, Good Ancestors proposed an approach to the classification of GPAI and GPAI with systemic risk. Overall, a definition of GPAI with systemic risk needs to separate generally safe GPAI from the kinds of GPAI that could pose systemic, catastrophic or existential risks. There are a variety of overseas approaches, including EU AI Act Article 51 and Annex 13, the US Executive Order, and Californian Bill SB1047, that seek to do this.

Overall, a mix of **presumptive thresholds** (e.g. more than a certain amount of investment or a certain amount of compute) with **capability or behaviour specifications** (e.g. ability to assist in weaponisation, to generate disinformation at scale, persuade or deceive humans, or loss of control) and **risk thresholds** (e.g. threat to national security) should be used to define GPAI with systemic risks.

Presumptive thresholds: AI models that have a very large cost to train due to the compute required, could be presumed to have capabilities sufficient for systemic risk. An example is in California Bill SB1047, which proposed specific guardrails for AI models that cost USD >$100M to train, or USD >$10M to fine-tune. An alternative approach would be to specify the training compute threshold directly, such as 10^{25} FLOP (as in the EU AI Act, Article 51(2)) or 10^{26} FLOP (as in US Executive Order 14110 Section 4.2(b)(i)). The EU AI Act, in Annex 13[48] includes other quantitative criteria

[48] EU AI Act, Annex XIII. https://artificialintelligenceact.eu/annex/13/

that could form the basis for presumptive thresholds, including the number of parameters in the model, the quality or size of the data set, or the estimated time or energy consumption for the training.

Capability or behaviour specifications: Some capabilities and behaviours can be estimated or forecasted before a model is trained, based on known data, algorithmic, and compute inputs to the AI model.[49] Other capabilities and behaviours can be evaluated after initial AI model training, but before deployment as an AI system. The EU AI Act, in Annex 13, includes other quantitative criteria that could form the basis for capability or behaviour specifications, such as the input and output modalities of the model (e.g., text to image; voice to voice), its level of autonomy and scalability, adaptability to new tasks without fine-tuning or additional training, and the tools it has access to. (a) the number of parameters of the model. Finally, there are emerging internationally-agreed methods for eliciting and evaluating these capabilities and behaviours (e.g. the MLCommons AI Safety Benchmark, which includes quantitative tests for specific capabilities, behaviours, and hazards in AI models).[50] Uuk et al provide a table listing categories of Systemic Risk from General Purpose AI (*see the following page*).

Risk thresholds: Likelihood and severity of harm from AI models could be assessed in domains of systemic risk such as national security or national public health & safety, and models that reach a predetermined threshold could be classified as GPAI with systemic risks. A recent systematic review of systemic risks from general-purpose artificial intelligence developed a taxonomy of 16 categories of systemic risk[51] (see table below). The taxonomy extends and structures a set of exemplar systemic risks described in the EU AI Act Recital 110.

[49] Phuong, M. et al (2024). Evaluating Frontier Models for Dangerous Capabilities. arXiv. https://arxiv.org/abs/2403.13793
Brundage, M. et al (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv. https://arxiv.org/abs/2004.07213
[50] Vidgen, B. et al (2024). Introducing v0.5 of the AI Safety Benchmark from MLCommons. arXiv. https://arxiv.org/abs/2404.12241
[51] Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Brouwer, A., & Slattery, P. (2024). A Taxonomy of Systemic Risks from General-Purpose AI: Preliminary Findings.

Systemic risk categories	Description
Environmental harm	The impact of AI on the environment, including risks related to climate change.
Structural discrimination	The potential for AI to perpetuate or exacerbate existing inequalities and biases in society.
Harm to democracy and eroding trust in institutions	The threat AI poses to democratic processes and public trust in social/political institutions.
Diminishing rule of law	The risk of diminishing the accountability of public decision-makers due to AI's influence.
Catastrophic and existential risks	The broad range of catastrophic scenarios that could threaten human survival or drastically alter civilization.
Loss of control over AI	The risk of AI models and systems acting against human interests due to misalignment, loss of control, or rogue AI scenarios.
AI in combination with chemical, biological, radiological, and nuclear weapons	The dangers of AI amplifying the effectiveness/failures of nuclear, chemical, biological, and radiological weapons.
Economic disruptions	Economic disruptions ranging from direct impacts on the labor market to broader economic changes that could affect exacerbation of wealth inequality, stability of the financial system, labor exploitation or other economic dimensions.
Concentration of power	The risks associated with the concentration of military, economic, or political power in AI-controlling entities or to AI itself.
Information and cultural harms	The impact of AI on information integrity, epistemic processes, and cultural values.
Injury to animals	The ethical considerations and potential harm to animals resulting from AI developments.
Artificial sentience and suffering	The risk of creating AI models and systems capable of suffering, leading to ethical concerns on an astronomical scale.
Adverse impact on fundamental rights	The large-scale effects AI could have on fundamental human rights and freedoms.
Governance failures	The risks associated with inadequate or flawed governance of AI systems, leading to broader societal harm.
Security threats	The potential for AI to intensify security threats, including cyber warfare, military violence, and geopolitical instability.
Irreversible societal change	The potential for AI to drive profound, long-term changes to social structures, cultural norms, and human relationships that may be difficult or impossible to reverse.

The key point is that we need to identify GPAI with systemic risks and require additional obligations for these models and systems, but not for other AIs that do not pose these special risks.

Regardless of the particular approach, we need future-proof and flexible definitions. Article 51(3) of the EU AI Act allows the presumptive threshold to be updated in light of evolving technical development such as algorithmic or hardware efficiency, which could make a simple threshold of 1025 FLOP training compute obsolete as improvements in algorithmic efficiency bring down compute requirements over time.[52]

[52]A Ho et al, 'Algorithmic Progress in Language Models' (Preprint, arXiv, 9 March 2024) https://doi.org/10.48550/arXiv.2403.05812.

8. Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings?

This submission argues that AI models and systems pose a wide range of risks across their development and deployment. A single set of mandatory guardrails applied to a unified definition of “high-risk” will inherently struggle to capture the best ways to mitigate risk and apply them in the correct places. This would result in under-regulation, over-regulation, or both.

10. Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately?

No. In general, obligations need to be more focused on developers. Specific actions developers could take to safeguard GPAI models with systemic risk are detailed in Question 11.

As the Mandatory Guardrails Paper notes, effective regulation needs to target those best able to address the identified risk. The "black box" nature of AI limits the ability of users and deployers to understand and control risks – see "information asymmetry and model opacity".

Effective regulation also needs to prevent developers from unreasonably shifting risk. Developers are currently using end-user licence agreements to push responsibility to users and deployers.[53] "Open washing" could also be used by developers to avoid accountability. **Effective regulation must incentivise developers to take meaningful actions that protect Australian interests**. Australia needs to be courageous and set expectations for the AI that is deployed here. A less courageous approach that puts obligations preferentially on Australian deployers because they're an "easier target" would increase regulatory burden and reduce effectiveness. This is discussed in response to question 12.

Overall, an *Australian AI Act* should:

- Put obligations on models that would be deployed in Australia. If the model has not met its obligations, it should not be deployed here.
- Require developers to make verifiable claims about their systems – including the risks they do or do not pose and the efficacy of their mitigations – and enable testing of those claims by independent third parties, AI Safety Institutes and regulators. Where claims are not verified, developers should be held accountable.
- Impose meaningful liability for non-compliance. A risk-based model only works if it imposes sufficient consequences to change behaviour.

[53] OpenAI, 'Terms of Use' (Web Page, 14 November 2023) https://openai.com/policies/row-terms-of-use/.

Information asymmetry and model opacity

The AI lifecycle is characterised by information asymmetries between the public, users, investors, companies, and regulators. These asymmetries position AI developers with the highest concentration of information and, hence, the best ability to understand risks and ultimately mitigate risks. Taeihagh et al argue:[54]

> *"One problem posed by emerging disruptive technologies which poses problems for their dissemination and control is directly linked to their hi-tech nature and the limited knowledge that most social actors have concerning how it works and why, and what are the possible applications and consequences of its deployment. That is, in policy terms,* ***the policy environment with respect to emerging technologies is characterized by asymmetries in information across agents and at multiple levels of society and government****."*

Taeihagh et al argue that a key mechanism by which AI developers can earn trust from the broader community is by being required to make verifiable claims about their models and then being held accountable for those claims over time. For this to be effective, AI regulators, AI safety institutes and third-party evaluators need a mechanism for scrutinising developers' claims and holding them accountable.

Slattery et al make a similar argument:[55]

> *[A] challenge for effective governance is an inability to influence AI developers and deployers to take safe actions. Frequently, this inability is driven by an asymmetry of information between technology companies and regulators. Technology companies often have far better knowledge about the capabilities, functioning, and potential uses of their AI systems; they possess both the technical expertise and the proprietary data that inform AI development. Without access to this knowledge, regulators can find it difficult to craft targeted rules that address the specific challenges posed by AI."*

The challenges of information asymmetry are compounded by opacity or lack of explainability – often called the "black box" problem. Opacity inherently limits the

[54] A Taeihagh, M Ramesh and M Howlett, 'Assessing the Regulatory Challenges of Emerging Disruptive Technologies' (2021) 15(4) *Regulation & Governance* 1009–1019, https://doi.org/10.1111/rego.12392.
[55] P Slattery, A K Saeri, E A C Grundy, J Graham, M Noetel, R Uuk, J Dao, S Pour, S Casper and N Thompson, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence* (MIT FutureTech, Massachusetts Institute of Technology, 2024) https://airisk.mit.edu/.

ability of deployers, users and regulators to understand and address AI risks – further increasing the importance of Government regulation that forces AI companies to internalise risks. Slattery et al say:

> *AI's decision-making… is often unpredictable, opaque, and involves complex interactions between millions of parameters. This complexity makes understanding how an AI arrived at a decision, and consequently who is responsible for the consequences of that decision, very difficult. In the absence of a regulatory or legal incentive to take safety engineering seriously, developers may release poorly designed AI systems , and people harmed by those systems may be left without recourse.*

The importance of obligations on developers is also evident in the case of misuse. The Mandatory Guardrails Paper correctly says (pg18):

> *The second proposed category of high-risk AI relates only to advanced, highly-capable GPAI models, where all possible applications and risks cannot be foreseen.* ***The risk lies in the potential for these models to be used – or misused – for a wide range of purposes with emergent risks****.*

For misuse risks, like weaponisation, regulation is trying to solve for the possibility that the deployer or user is the one who intends to cause harm. **For misuse, an obligation on the deployer or user is inherently ineffective**. While small-scale misuse can be tolerated or addressed by general law, **highly-consequential misuse can only be effectively mitigated at the model level**.

Overall, while developers should not be the exclusive target of regulation, and there are sensible obligations relevant to deployers and users, developers need to be the key focus of regulation. This is particularly the case for GPAI with systemic risks.

Open-sourcing and open-washing

There's a specific concern, mainly relevant to GPAI with systemic risk, that developers will be encouraged to put themselves in a position where they cannot

fix problems and, therefore, do not have an obligation to fix problems.[56] Seger et al. summarise the tension around open source AI as:[57]

> *Recent decisions by AI developers to open-source foundation models have sparked debate over the prudence of open-sourcing increasingly capable AI systems…On the one hand, this offers clear advantages including enabling external oversight, accelerating progress, and decentralizing AI control. On the other hand, it presents notable risks, such as allowing malicious actors to use AI models for harmful purposes without oversight and to disable model safeguards designed to prevent misuse.*

While open source has historically been a valuable way to increase the quality of complex systems, it would be irresponsible to lose control of a GPAI with systemic risks in this way. There's an inflection point where open source transitions from reducing risk to increasing risk. For instance, open-sourcing a mobile phone application would invite scrutiny, allow faults to be addressed, and improve security overall. Conversely, open-sourcing classified plans for nuclear weapons may bring some beneficial scrutiny, but the advantages it would provide to bad actors would manifestly outweigh any benefit.

Seger et al. offer 5 key recommendations to navigate this tension:

1. *Developers and governments should recognize that some highly capable models will be too risky to open-source, at least initially*
2. *Decisions about open-sourcing highly capable foundation models should be informed by rigorous risk assessments.*
3. *Developers should consider alternatives to open-source release that capture some of the same distributive, democratic, and societal benefits, without creating as much risk.*
4. *Developers, standards setting bodies, and open-source communities should engage in collaborative and multi-stakeholder efforts to define fine-grained standards for when model components should be released.*
5. *Governments should exercise oversight of open-source AI models and enforce safety measures when stakes are sufficiently high.*

---

[56] A Liesenfeld and M Dingemanse, 'Rethinking Open Source Generative AI: Open-Washing and the EU AI Act' (Paper presented at *FAccT '24*, Rio de Janeiro, Brazil, 3–6 June 2024) 14 pages https://doi.org/10.1145/3630106.3659005.

[57] E Seger, N Dreksler, R Moulange, E Dardaman, J Schuett, K Wei, C Winter, M Arnold, S Ó hÉigeartaigh, A Korinek, M Anderljung, B Bucknall, A Chan, E Stafford, L Koessler, A Ovadya, B Garfinkel, E Bluemke, M Aird, P Levermore, J Hazell and A Gupta, *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives* (Centre for the Governance of AI, 2023) https://law-ai.org/wp-content/uploads/2023/10/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI-1.pdf.

Good Ancestors endorses this approach and recommends that Government leverage nuanced definitions of high-risk AI and GPAI with systemic risk to encourage open source where it increases safety and to discourage open source where it increases risk.

11. Are the proposed mandatory guardrails sufficient to address the risks of GPAI?

No. Above, including in response to Question 6, this submission argues that AI models and systems pose a wide range of risks and a single set of mandatory guardrails will inherently struggle to capture the best ways to mitigate risk and apply them in the correct places. In the same way that a single ruleset could never effectively regulate push bikes and aeroplanes, a single ruleset cannot effectively regulate narrow AI and future GPAI with systemic risks.

Currently, in part because of the proposed guardrails generality, they miss obligations that ought to be imposed on the development of GPAI.

In the context of a recent EU AI Act consultation,[58] Uuk, R et al. evaluated the effectiveness of various risk mitigation measures for general-purpose AI (GPAI) models in reducing systemic risks using a comprehensive literature review and an expert survey involving 75 experts across multiple domains.[59] The review assessed 27 proposed mitigations. Uuk, R et al. built on a similar analysis conducted by Schuett et al. in 2023.[60]

Schuett et al. and Uuk et al. identify a range of approaches to risk mitigation – including **safety incident reporting, prohibiting high-stakes applications, setting intolerable risk thresholds, setting ratios of safety vs. capability investment, and deploying powerful models in stages** – that experts consider to be top mitigations as part of a combined approach but which do not appear to feature in the Mandatory Guardrails paper.

SB1047 also includes an obligation, before beginning to initially train a covered model, to comply with various requirements, including implementing the capability to promptly **enact a full shutdown** of the model being trained, including derivatives controlled by a developer.[61]

---

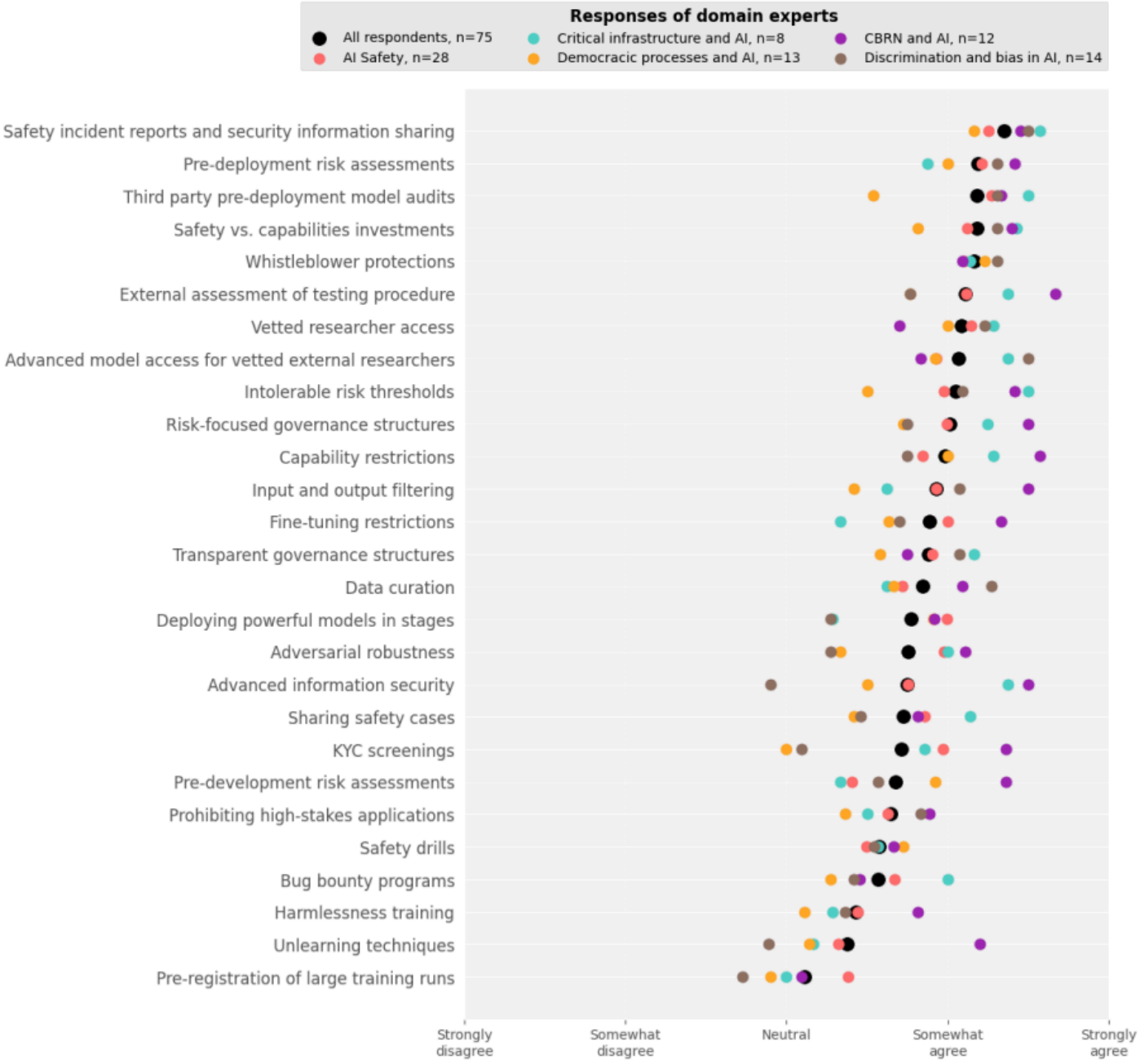
[58] **European Commission**, 'AI Act: Participate in the Drawing-Up of the First General-Purpose AI Code of Practice' (Web Page, Shaping Europe's Digital Future, 2024) https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice.

[59] Uuk, R., Brouwer, A., Schreier, T., Dreksler, N., Pulignano, V., & Bommasani, R. (2024). Effective risk mitigation measures for systemic risks from general-purpose AI: Preliminary findings. Future of Life Institute.

[60] J Schuett, N Dreksler, M Anderljung, D McCaffary, L Heim, E Bluemke and B Garfinkel, *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion* (Centre for the Governance of AI, May 2023).

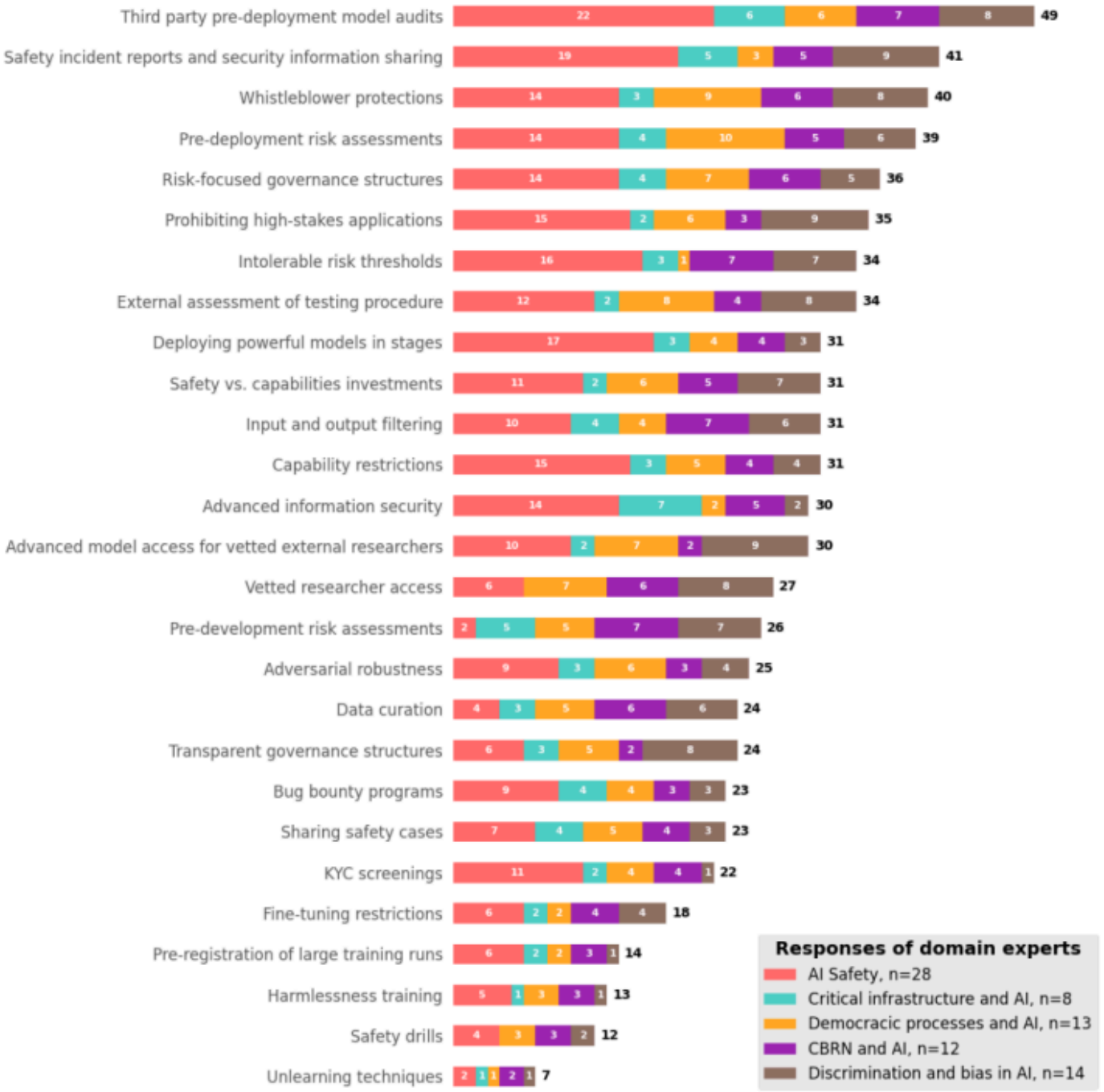
[61] **Safe and Secure Innovation for Frontier Artificial Intelligence Models Act**, SB 1047, California Legislature, 2023–2024 Regular Session, introduced 7 February 2024 https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

Good Ancestors recommends that Australia develop a set of mandatory guardrails for the development of GPAI with systemic risk that draws on international models and the best available research. Good Ancestors would welcome an opportunity to collaborate with the Government on this standalone set of guardrails appropriate for the development of GPAI with systemic risk.



Uuk, R et al.[62] *figure illustrating expert agreement on the effectiveness of different risk mitigation measures for general-purpose AI models, n=75. The black dots represent the average over all respondents, while the coloured dots represent the averages of the different domain expert groups. The measures are ranked on highest average score to lowest average score.*

[62] Uuk, R., Brouwer, A., Schreier, T., Dreksler, N., Pulignano, V., & Bommasani, R. (2024). Effective risk mitigation measures for systemic risks from general-purpose AI: Preliminary findings. Future of Life Institute.



Uuk, R et al.[63] *figure illustrating expert top-10 selection frequency of risk mitigation measures for general-purpose AI models, n=75. The bars show the number of experts who put that measure in their top-10 of most effective measures, first split out by expert group (represented by different colours), followed by the total number in bold.*

[63] Uuk, R., Brouwer, A., Schreier, T., Dreksler, N., Pulignano, V., & Bommasani, R. (2024). Effective risk mitigation measures for systemic risks from general-purpose AI: Preliminary findings. Future of Life Institute.

Obligations for accident and incident reporting and investigation

Accident and incident reporting might be one of the most effective guardrails for GPAI with systemic risk. Guardrail 9 "keep and maintain records" suggests that organisations training large state-of-the-art GPAI models with potentially dangerous emergent capabilities would have to disclose training runs to the Australian Government. The discussion in Guardrail 9 is that, under the EU AI Act, providers of GPAI models with systemic risk must also report relevant information about serious incidents to the EU AI Office. However, the discussion does not say that Guardrail 9 would extend to safety incident and accident reporting in the Australian context. Similarly, Guardrail 8 suggests that *deployers must report incidents to developers*, but not that developers would have to report to deployers, or that anyone would have to report to the Government.

Kolt et al in their paper *Responsible Reporting for Frontier AI Development* argue that incident and accident reporting is a key measure to ensure that governments, industry, and civil society have visibility of new and emerging risks posed by frontier systems:[64]

> *Organizations that develop and deploy frontier systems have significant access to [information about frontier systems]. By reporting safety-critical information to actors in government, industry, and civil society, these organizations could improve visibility into new and emerging risks posed by frontier systems. Equipped with this information, developers could make better informed decisions on risk management,* ***while policymakers could design more targeted and robust regulatory infrastructure.***

The UK's Emerging Processes for Frontier AI Safety Report also endorses incident and accident reporting. The report says on page 20:[65]

> *[GPAI developers and deployers should r]eport any details of security or safety incidents or near-misses to relevant government authorities. This includes any compromise to the security of the organisation or its systems, or any incident where an AI system – deployed or not – causes substantial harm or is close to doing so. This will* ***enable government authorities to***

[64] N Kolt, M Anderljung, J Barnhart, A Brass, K Esvelt, G K Hadfield, L Heim, M Rodriguez, J B Sandbrink and T Woodside, 'Responsible Reporting for Frontier AI Development' (Preprint, arXiv, 3 April 2024) https://doi.org/10.48550/arXiv.2404.02675.
[65] **UK Department for Science, Innovation & Technology**, *Emerging Processes for Frontier AI Safety* (October 2023) https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf.

build a clear picture of when safety and security incidents occur and make it easier to anticipate and mitigate future risks. *Incident reports could include a description of the incident, the location, start and end date, details of any parties affected and harms occurred, any specific models involved, any relevant parties responsible for managing and responding to the incident, as well as ways in which the incident could have been avoided. It is important that incidents indicative of more severe risks are reported as soon as possible after they occur. High-level details of safety and security incidents — with sensitive information removed — could also be made public, such as have been shared in the AI incident database.*

Overall, Good Ancestors recommends that, at minimum, Guardrail 9 explicitly includes accident and incident reporting. The best approach would be standalone guardrails for the development of GPAI with systemic risks that include all relevant risk mitigations highly rated by experts with a regulatory mechanism to add new Guardrails as they are developed or discovered.

Obligation to share with the Government, including AI Safety Institutes, for the purpose of model evaluation

Guardrail 4 includes sensible obligations, like requiring that developers of GPAI models must conduct adversarial testing for any emergent or potentially dangerous capabilities and engage in post-market monitoring. Guardrail 9 refers to the obligations to keep records and audit reports and to share them with regulators, and guardrail 10 refers to sharing conformity assessments with third parties, "government entities", or regulators.

These obligations should go further and give the regulator power to demand that a developer of GPAI models with systemic risk share access to models, subject to appropriate protections,[66] with government entities (such as AI Safety Institutes and regulators) to verify claims. See Question 10 for more discussion of the importance of requiring and conducting such verifications.

[66] Given the potential risks associated with the development of GPAI with systemic risk, it would be appropriate for any sharing to occur in a fashion that does not increase risks, such as increasing cybersecurity attack surface. Regulation could create a mechanism to specify protections that government must meet for any information or access granted to it.

Media reported that Mr Altman, the CEO of OpenAI, made a commitment to Minister Husic in June 2023 *"to give Australian researchers new access to OpenAI's models when future versions are developed."*[67] OpenAI and Anthropic have made similar commitments to share models with the US and UK AI Safety Institutes.[68] However, media reports have indicated that **AI developers have not been forthcoming in complying with these agreements**.[69] Supporting this claim, the U.S. Senate is pressing OpenAI to share its next foundation model with the U.S. Government for pre-deployment testing, review, analysis and assessment.[70] Good Ancestors does not know if OpenAI has upheld its agreement with Minister Husic to give early access to Australian researchers.

A sharing obligation of this kind would be consistent with other regulatory environments. For instance, in addition to conducting their own testing, car makers are typically required to submit cars to ANCAP for safety testing prior to the car being available to the market.[71] A version of this obligation makes sense for GPAI with systemic risk.

Cybersecurity of model weights is a particular concern for GPAI with systemic risk

Guardrail 3 relates to protecting AI systems and includes a reference to compliance with the *Security of Critical Infrastructure Act 2018*. This guardrail should be expanded to explicitly include protecting GPAI models - with a focus on model weights.

RAND's research report "Securing AI Model Weights" explains why security model weights are important and provides recommendations about how it could be achieved. The research report says:[72]

---

[67] Paul Smith, 'Why Even Sam Altman Doesn't Trust ChatGPT' *Australian Financial Review* (online, 16 June 2023) https://www.afr.com/technology/why-even-sam-altman-doesn-t-trust-chatgpt-20230615-p5dh02.
[68] Lauren Feiner, 'OpenAI and Anthropic Will Share Their Models with the US Government' *The Verge* (online, 30 August 2024) https://www.theverge.com/2024/8/29/24231395/openai-anthropic-share-models-us-ai-safety-institute.
[69]V Manancourt, G Volpicelli and M Chatterjee, 'Rishi Sunak Promised to Make AI Safe. Big Tech's Not Playing Ball' *Politico* (online, 26 April 2024) https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/.
[70]B Schatz, B R Luján, M R Warner, P Welch and A S King Jr, 'Letter to OpenAI' (United States Senate, 22 July 2024) https://www.schatz.senate.gov/imo/media/doc/letter_to_openai.pdf.
[71] N Platt, 'ANCAP Demystified' (Paper presented at Australasian College of Road Safety Conference, 2011) https://archive.acrs.org.au/article/ancap-demystified/.
[72] S Nevo, D Lahav, A Karpur, Y Bar-On, H A Bradley and J Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (RAND Corporation, 30 May 2024) https://www.rand.org/pubs/research_reports/RRA2849-1.html.

> *The motivation to secure frontier AI models includes not only protecting intellectual property but also potentially safeguarding national security. There has always been a commercial motivation to secure AI models. However,* ***growing concerns that risks from future AI models may rise to national security significance introduce an additional motivation: the security and interests of the broader public****. As a result, discussions of how to secure frontier AI models—that is, models that match or exceed the capabilities of the most advanced AI models at the time of their development—are expanding beyond AI organizations to include stakeholders across industry, government, and the public.*

If model weights are stolen, it could allow the bypassing of protections and allow misuse of an AI model without restriction. That is, most other safeguards would be undermined in the event that a dangerous AI model was stolen and made public without safeguards. This could lead to catastrophic harm, and it might not be possible to "put the genie back into the bottle". For this reason, Good Ancestors recommends that safeguards specifically assign cyber security obligations to AI systems and models according to their risk.

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

Yes.

As per our response to question 10, **the best place for obligations to be imposed is typically on the development of AI models**. Small-to-medium-sized businesses will often have limited capability and capacity to safeguard a risky model. The central point to impose effective mitigations is the development of models. Shifting the regulatory balance towards developers and away from deployers will be good for Australian businesses. We should not tolerate a world where developers push risks onto deployers in circumstances where deployers have no realistic ways to manage those risks.

As per our executive summary and response to questions 8 and 16 **a regulatory approach that empowers established regulators to regulate high-risk AI within established fields of authority has the benefit of accessing their expertise and also means that small-to-medium-sized businesses can continue to engage with regulators they are familiar with** – only having to engage with a new AI regulator in certain circumstances. This also frees up a new AI regulator to focus on gaps and GPAI with systemic risk.

13. Which legislative option do you feel will best address the use of AI in high-risk settings?

☑ A whole of economy approach – introducing a new cross-economy AI Act

As set out in the executive summary, Australia needs a regulatory framework that:

- can interface with international legal trends, as demonstrated in the EU, Canada, the US states, and elsewhere
- is agile enough to deal with rapid and unexpected developments in AI capabilities and risks, and
- has the teeth necessary to shape the behaviour of billion or trillion-dollar offshore AI developers.

Considering these factors, **we need an *Australian AI Act***.

The Mandatory Guardrails Paper is right to note that whole-of-economy AI regulation would result in duplicate obligations with existing legislative frameworks and coordination challenges across regulators. However**,** the **drawbacks of an *Australian AI Act* are only acute if we stretch a single set of mandatory guardrails over high-risk AI, GPAI and GPAI with systemic risk**. Instead, if we tailor guardrails to specific AI classifications, we can readily highlight the preeminence of specific regulators as they relate to high-risk narrow AI systems, while building an effective approach that targets GPAI with systemic risk.

This approach also lets us apply obligations to the parties best able to mitigate risks. For instance, where a narrow-AI is deployed in Australia for a specific purpose (such as a medical device), our regulations can sensibly target the deployers of those devices. Whereas, in cases where GPAI with systemic risks could have global ramifications, our AI Act needs to be courageous and set clear safety expectations for all developers who hope to deploy systems into Australia.

We also need to build regulatory architecture that enables global enforcement. This can mean two things:

1) **Ensuring domestic Australian legislation has practical enforcement mechanisms** that can be applied to AI developers both onshore and offshore. For instance, the proposed guardrails would ask developers of GPAI models to identify dangerous capabilities and emergent properties

(guardrail 4) and implement risk management strategies (guardrail 2).

OpenAI's assessment of "o1-preview" and "o1-mini" does these kinds of risk assessments and says that the models "can help experts with the operational planning of reproducing a known biological threat" and assesses the risk as "medium".[73] Australia needs to be in a position to verify that risk assessment, ask whether we think that risk is unacceptable, and prevent the risk if we decide that it is intolerable.

2) **Laying the groundwork for true global interoperability on AI safety**. In other industries, like aviation, if an overseas organisation wishes to operate in Australia, it must comply with laws and regulatory safety standards overseen primarily by the Civil Aviation Safety Authority (CASA).[74] Failure to comply with these standards can result in serious consequences such as certification denial, operational restrictions, importation roadblocks and other legal action.

Similarly, if an overseas car manufacturer wishes to transact business in Australia, they must adhere to national vehicle safety standards. These standards include maintaining compliance with Australian Design Rules (ADRs), Consumer Law, Import Regulations, Vehicle Registration and Licensing, Safety Compliance and Automotive Industry Codes of Practice. Specifically, the ADRs manage road vehicle safety, anti-theft and emissions.[75] The ADRs are harmonised with relevant international regulations, incldluing the UN 1958 Agreement[76] and the 1998 Agreement.[77] Australia can seek to enforce similarly robust standards on developers of AI systems intended for the Australian market, especially high-risk AI or GPAI with systemic risk.

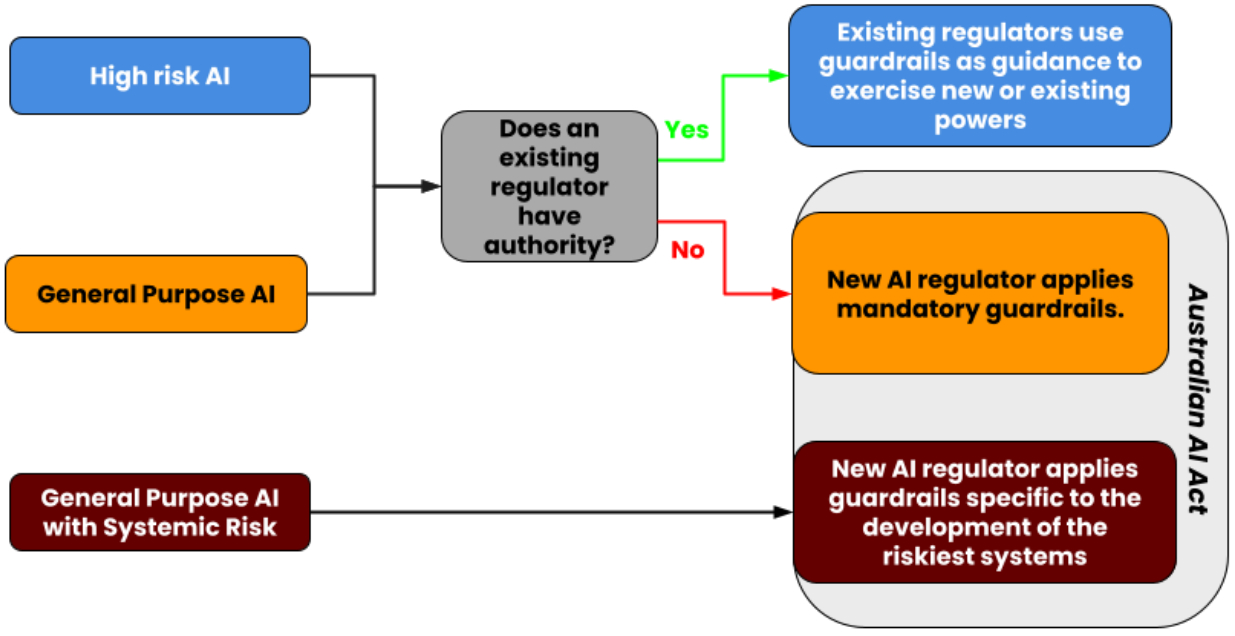
---

[73] *OpenAI*, 'OpenAI-01 System Card' (Web Page, 2023) https://openai.com/index/openai-o1-system-card/.
[74] **Australian Government Civil Aviation Safety Authority**, *CASA's Regulatory Framework* (Web Page, 30 October 2023) https://www.casa.gov.au/rules/regulatory-framework/casas-regulatory-framework.
[75] **Australian Government Department of Infrastructure**, *Australian Design Rules* (Web Page) https://www.infrastructure.gov.au/infrastructure-transport-vehicles/vehicles/vehicle-design-regulation/australian-design-rules.
[76] **United Nations**, *Agreement Concerning the Adoption of Harmonized Technical United Nations Regulations for Wheeled Vehicles, Equipment and Parts Which Can Be Fitted and/or Be Used on Wheeled Vehicles and the Conditions for Reciprocal Recognition of Approvals Granted on the Basis of These United Nations Regulations*, UN Doc E/ECE/TRANS/505/Rev.3 (20 October 2017) https://unece.org/trans/main/wp29/wp29regs.
[77] **United Nations Economic Commission for Europe, Inland Transport Committee**, *Agreement Concerning the Establishing of Global Technical Regulations for Wheeled Vehicles, Equipment and Parts Which Can Be Fitted and/or Be Used on Wheeled Vehicles*, UN Doc ECE/TRANS/132 (25 June 1998) https://unece.org/text-1998-agreement.

In the case of aviation, the International Civil Aviation Organization (ICAO), a United Nations agency, helps 193 countries to cooperate and share their skies to their mutual benefit.[78] ICAO develops a Global Aviation Safety Plan (GASP)[79] and ensures each State's national aviation safety plan is developed in alignment with the GASP and the regional aviation safety plan. AI safety should ultimately work in a similar direction.

Overall, Good Ancestors proposes that most **deployment** of AI – both high-risk AI and GPAI - would be regulated by established regulators who already have authority using guardrails as guidance. Lead departments and regulators may need to update powers and approaches as required. A new AI regulator would regulate the deployment of AI where it falls outside the authority of existing regulators. A new AI regulator would regulate the development of high-risk and general-purpose AI. In the case of GPAI with systemic risk, the regulator would use a specific set of guardrails designed for that purpose.

How Good Ancestors proposes imposing guardrails



[78] **International Civil Aviation Organization**, *About ICAO* (Web Page) https://www.icao.int/about-icao/Pages/default.aspx.
[79] **International Civil Aviation Organization**, *ICAO Global Aviation Safety Plan* (Web Page) https://www.icao.int/safety/GASP/Pages/Home.aspx.

AI obligations in practice

The above discussion sets out principles for applying safeguards. But what could this look like in practice?

1) **Existing regulators should be the "front line" in the regulation of AI.**
   Existing regulations sometimes address specific products, sometimes address specific industries or professions, or both. Regulators should be given the principles and guardrails as tools for identifying and mitigating the risky use of AI within their sphere of responsibility. They can then use their domain expertise to tackle risks from AI.
   - **This approach is best for business** because it reduces duplication and lets them continue existing relationships. It's also **best for risk reduction** because domain expertise will be critical in these cases.
   - Government will need to **support existing regulators** to skill up and understand emerging risks. This could look like an Australian AI Safety Institute with a function of translating the latest research about risk information to the public service.

   Making existing regulators the "front line" means that an AI Act wouldn't become impractical by attempting to regulate every aspect of the economy. An AI Act would be free to focus on novel and critical issues.

2) **An *Australian AI Act* and specific AI regulator should cover "gaps" between domain regulators.**
   Low-risk uses of AI can be largely unregulated and covered only by general laws.

   High-risk uses of narrow and general AI are possible outside of current domain regulators. For instance, an AI intended for use in a regulated industry could be adopted in a different industry. If that use is high-risk and outside the authority of an existing regulator, the *AI Act* should apply.

   GPAI AI and GPAI with systemic risk pose special kinds of risks. An AI Act should include specific guardrails that are adapted to the risk they present, including during development and deployment.

   A key merit of this approach is that the kinds of guardrails necessary for the safe development of GPAI with systemic risk are very different from the kinds of guardrails necessary for the safe deployment of high-risk AI. Having specific guardrails means they can be adapted to their purpose.

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

Yes

In response to question 5 we provide details about the need for regulatory flexibility and argue in favour of the approach taken in the Canadian AIDA. We also provide an argument for why an AI-specific regulator should have a high degree of separation from functions relating to the adoption of AI.

In response to question 13, we argue for a specific role for established regulators in the deployment of high-risk AI within the existing responsibility of regulators. This approach is a hybrid between a domain-specific approach that adapts existing regulatory frameworks to include the proposed mandatory guardrails and a whole-of-economy approach – introducing a new cross-economy AI Act.

A strength of the model we propose in response to question 13 – where established regulators lead within their authority and an AI regulator addresses emerging issues and gaps - is that it makes the task of a new AI regulator more achievable. AI will likely be used in almost all elements of our economy, and many fine-grained issues will need to be resolved. An AI regulator tasked with resolving challenges across the economy will have an impossibly large scope. An AI regulator will better be able to focus on emerging risks, including GPAI with systemic risk, if the AI regulator can let established regulators lead wherever possible.

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

☑ A whole of economy approach – introducing a new cross-economy AI Act

Any regulatory framework is right to anticipate more significant changes in AI than in other kinds of regulated technology. It is possible, perhaps even likely, that any *Australian AI Act* resulting from this consultation will be in force when artificial general intelligence is developed.[80] Certainly, an *Australian AI Act* should be drafted with that possibility in mind.

In response to question 5 we provide details about the need for regulatory flexibility and argue in favour of the approach taken in the Canadian AIDA. We also provide an argument for why an AI-specific regulator should have a high degree of separation from functions relating to the adoption of AI.

Overall, regulations will best be able to respond to likely step-changes in technology if:

- Established regulators have the lead within their spheres of authority, ensuring that a new AI regulator is not overburdened.
- A new AI regulator has a suitable framework for engaging with cutting-edge AI systems.
- A new AI regulator has sufficient capability and capacity, including support from an Australian AI Safety Institute.

[80]Sam Altman, *The Intelligence Age* (Web Page, 23 September 2024) https://ia.samaltman.com/.

16. Where do you see the greatest risks of gaps and inconsistencies with Australia's existing laws for the development and deployment of AI?

Regulation needs to account for AI risks varying widely in magnitude. They range from individual harms (e.g. cyberbullying, privacy harms) to societal-scale harms (e.g. cyberattacks, biological weapons). Regulators have previously grappled with ubiquitous technologies that can cause appreciable harm, like cars or planes or medical devices. Equally, regulators have grappled with constrained technologies that can cause catastrophic harm, like nuclear weapons or biotechnology. AI presents a **unique regulatory challenge**, being potentially ubiquitous whilst also being able to cause catastrophic harm.

Improperly grappling with this conceptual challenge is the main way gaps and inconsistencies will emerge. One or the other end of the risk spectrum could be neglected, or mitigations appropriate for one kind of risk could be applied to a different kind of risk. In the first "Safe and Responsible AI" consultation, the Government seemed likely to overlook the possibility of catastrophic risks and focus only on individual harms. In the "Mandatory Guardrails" consultation, the Government has acknowledged both ends of the risk spectrum but proposed a single set of mandatory guardrails. This approach will inevitably lead to overregulation, underregulation, or both.

In question 13, Good Ancestors argues that we need a set of definitions that divides AI systems and models according to their risks. Narrow systems that risk individual harm are best managed by established regulators. Systems and models that present catastrophic risks should be addressed with targeted guardrails aimed at developers. An AI Regulator can be empowered to deal with high-risk AI uses that fall between any gaps within existing regulatory architecture.

This approach would best use existing resources, best reduce the risk of gaps or inconsistencies, and best ensure that risks across the risk-spectrum are addressed. The Australian public is – rightly – deeply concerned about the possibility of catastrophic risks from advanced AI systems and Government ought to work to ensure they are addressed.

AI crisis response planning

Good Ancestors' general position is that AI poses catastrophic and potentially existential risks, and that a risk-based approach must work hard to prevent those risks from ever occurring. However:

1. Even our best efforts might not be successful in averting these risks, and
2. The process of preparing for a crisis can help understand the nature of a risk and help find effective mitigations.[81]

For these reasons, Good Ancestors recommends that the Department of Industry work with the Department of the Prime Minister & Cabinet, the Department of Home Affairs, and the National Emergency Management Agency to update the Australian Government Crisis Management Framework[82] to include an Australian Government Catastrophic AI Crisis Plan. Developing such a plan would involve working through interactions with other plans, such as the Australian Cyber Response Plan, the National Health Emergency Response Arrangements and the Australian Government Domestic Security Crisis Plan, updating those plans to incorporate AI risks as required, and exploring scenarios (such as loss of control or intelligence explosion) that other plans do not cover and where an AI Crisis Plan would have to be pre-eminent.

[81] Peter Rogers, *Development of Resilient Australia: Enhancing the PPRR Approach with Anticipation, Assessment and Registration of Risks* (2011) 26(1) *The Australian Journal of Emergency Management* 22, https://www.aidr.org.au/media/7095/aidr_flipbook_emarrangements_2019-08-22_web_v2.pdf.
[82] **Department of the Prime Minister and Cabinet**, *Australian Government Crisis Management Framework* (September 2024) https://www.pmc.gov.au/resources/australian-government-crisis-management-framework-agcmf.