Safe and responsible AI in Australia

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

Submission from elevenM

4 October 2024

ABN 77 610 787 534

Contents

Introduction

Trust is a critical dependency for sustained digital innovation. Without trust, adoption falters and progress stalls. This is why at elevenM, we believe that the establishment and maintenance of trust must guide all innovation and is therefore a non-negotiable for technology policy.

Our challenge is that today, Australians don't trust AI. Only one in five Australians say that they are comfortable with government agencies using AI to make decisions about them, and even fewer Australians (one in six) are comfortable with businesses using AI to make decisions about them.[1]

If we want AI adoption to proceed, we must address the trust deficit through tangible, outcome-driven practices. So far, businesses have been slow to move. Although the majority of organisations (78%) agree on the importance of responsible AI outcomes, only a small minority (29%) have taken concrete action to implement responsible AI practices.[2]

AI safety guardrails have the potential to go a long way to reduce the trust deficit, but they must be comprehensive, flexible and designed to engage the public (much like a consumer protection regime). In other words, in addition to being effective in protecting people from harm and prioritising societal wellbeing, our regulatory framework must be seen and understood by the public as achieving those objectives.

With this in mind, we are grateful for the opportunity offer the following comments on the government's *Proposals paper for introducing mandatory guardrails for AI in high-risk settings*.

[1] See Office of the Australian Information Commissioner, *Australian community attitudes to privacy survey 2023* (2023) page 78. Available at https://www.oaic.gov.au/__data/assets/pdf_file/0025/74482/OAIC-Australian-Community-Attitudes-to-Privacy-Survey-2023.pdf.
[2] See Fifth Quadrant and National Artificial Intelligence Centre, *Australian Responsible AI Index 2024*, page 32. Available at https://www.fifthquadrant.com.au/content/uploads/Australian-Responsible-AI-Index-2024-Full-Report.pdf.

About elevenM

elevenM is a specialist AI, privacy, cyber security and data governance consultancy. Our mission is to build trust in an online world.

Our team comprises experts in complementary disciplines such as operations and technology, strategy, public policy, law, risk and compliance, IT risk, supplier risk, data governance and cyber security.

Members of our team combine technical and legal qualifications with extensive experience in the field. We work hand in hand with our clients to understand their businesses and identify effective and efficient solutions which are suitable for them — not only for today but for the constant changes coming over the horizon.

We work closely with organisations in the public and private sector to implement AI ethics, privacy and security programs, including improving transparency, delivering training and awareness initiatives, managing risks, remediating after breaches, conducting impact assessments, assessing vendors and third-party supplier frameworks, embedding ethics, privacy and security by design and more.

elevenM leads the consortium delivering SAAM, one of the Federal Government's new AI Adopt centres, which will guide Australian small to medium sized businesses through the design and implementation of AI solutions, with a focus on workflow and governance solutions.

For more information about elevenM, visit our website.

Submission

Defining high-risk AI

Questions	1. Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove? Please identify any: • low-risk use cases that are unintentionally captured • categories of uses that should be treated separately, such as uses for defence or national security purposes. 3. Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more defined approach, with a list of illustrative uses, needed? • If you prefer a list-based approach (similar to the EU and Canada), what use cases should we include? How can this list capture emerging uses of AI? • If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?
elevenM position	We support a list-based approach, which aligns with international approaches and provides clearer boundaries to the scheme with respect to use cases that are *not* covered. If a principles-based approach is adopted, a clear threshold for 'high risk' must be set in addition to the factors set out in the Proposals Paper. Practical guidance and examples will be required to clarify the threshold, as well as active oversight and enforcement to ensure it remains well defined and consistently applied across the economy.

Certainty of scope should be a priority

It is critical that the scope of application for any mandatory guardrails is clear and understandable – for developers, deployers, end users and those affected by AI systems.

We are concerned that a purely principles-based approach, even if supported by detailed guidance and examples, will leave uncertainty and grey areas where companies are not certain if they need to comply or alternatively, exercise a wide discretion in their interpretation of the principles. Variability also means that the public (end users and customers) are not certain what to expect.

In the context of privacy regulation, a principles-based regime has proven adaptable over time, but has resulted in significant and sustained uncertainty in the scope of application of the law. For example, the privacy invasive but widespread industry practice of 'data enrichment' – companies enhancing profiles of their customers by collecting data from third parties such as unrelated companies, loyalty programs and data brokers – relies on an uncertain and untested interpretation of what is 'unreasonable or impracticable' as well as a

strained interpretation of when an individual is 'reasonably identifiable'.[3] Data enrichment practices are largely inconsistent with non-binding guidance from the regulator, but have developed and flourished in a privacy grey area due (at least in part) to a lack of funding and appetite on the part of the regulator to pursue such an edge case.

Even with an adequately funded regulator, clarification of principles-based standards by means of litigation is slow and expensive. To avoid this, we consider that regulation should, wherever possible, clarify or limit the application of a general risk standard. A list-based approach can do this by ruling certain activities in our out, prior to the application of the general risk standard. This provides a clearer demarcation of which systems and use cases would be covered by the mandatory guardrails, allowing businesses to plan and innovate with more confidence about which rules apply, and providing end users with greater certainty about their risks, contributing to public trust.

A list-based approach would present lower compliance costs for business by requiring risk assessment only of systems that fall within the pre-defined categories. By contrast, a purely principles-based approach, or a 'principles plus examples and guidance' approach requires businesses to assess all AI systems to determine whether they reach the risk threshold at which the mandatory guardrails apply.

As the Proposals Paper notes, a list-based approach would be consistent with the emerging international consensus approach, matching the approaches adopted in the European Union under the Artificial Intelligence Act (AI Act) and proposed in Canada under the Artificial Intelligence and Data Act (AIDA).

In order to avoid capturing low risk use cases that fall within the designated high-risk categories, a list of excepted low risk use cases within the high-risk categories (such as appears in the EU AI Act) would also be required. Following the European model, this could operate as follows:

- Any use case falling within a listed high-risk category must conform to the guardrails.
- Use cases falling within a high-risk category that can be shown not to pose a high risk (applying the principles proposed in the Proposals Paper) are not required to conform to the guardrails, but developers and/or deployers must document this assessment and make it available to regulators on request.
- Further exemptions could apply to certain specified use cases within the high risk category that are determined to be low risk, such as where the AI system is intended to perform a narrow procedural task, or to detect decision-making patterns or deviations from prior decision-making patterns but not replace or influence previously completed human assessments.[4]

[3] See Kemp, Katharine, Australia's Forgotten Privacy Principle: Why Common 'Enrichment' of Customer Data for Profiling and Targeting is Unlawful (September 20, 2022). Available at SSRN: https://ssrn.com/abstract=4224653 or http://dx.doi.org/10.2139/ssrn.4224653.
[4] See EU AI Act, Article 6.

A more clearly defined risk threshold is required

The Proposals Paper lists proposed principles to which regard would be given when designating an AI system as 'high-risk'. The proposed principles outline the types of impacts that must be considered, and provide that the severity and extent of those adverse impacts will also be relevant to the determination of whether a use case is high risk. However the paper does not specify *how much adverse impact* is required for a system to be designated as high risk. For example, how much risk of adverse impact to an individual's physical or mental health or safety is required before a use case is considered high risk? Is a remote risk of minor injury to a single person sufficient? Is a likelihood of serious harm[5] required?

Whether a principles-based or list-based approach is adopted, a threshold for 'high risk' must be set in addition to the factors set out in the Proposals Paper.

Banning high risk use cases

Question	4. Are there high-risk use cases that government should consider banning in its regulatory response (for example, where there is an unacceptable level of risk)? If so, how should we define these?
elevenM position	We support targeted bans on use cases or technologies that present unacceptable risks to human rights or public interests.

It is not appropriate to risk manage or apply guardrails to irredeemably harmful use cases. Bans have the benefit of sending a clear and simple message about the bounds of acceptable behaviour. Clear messages can help engender trust and may encourage more people to use AI with confidence that the most concerning applications will under no circumstances be permitted.

Bans are appropriate where use cases or technologies present unacceptable risks to human rights or public interests without any countervailing benefits. Criteria or requirements for identifying applications or technologies to be banned should be no different to the criteria on which applications or technologies are risk-rated. Unacceptable risks associated with an AI system might emerge for all users, or only for certain groups that are more susceptible to harm, such as children.

Bans should be applied as narrowly and specifically as possible and may be timebound or subject to regular reviews. The risks associated with any AI application or technology will always be highly dependent on context, and any prohibition on a class of applications or technologies risks capturing beneficial uses as well.

Bans on a small number of very high-risk activities or technologies (like social scoring or facial recognition technology in certain circumstances) are likely to be necessary to ensure alignment with international human rights standards, as well as interoperability with key technology trading partners such has the EU. We see such alignment and interoperability as

[5] The threshold triggering mandatory notification obligations in respect of a data breach, under the Privacy Act 1988.

more likely to provide a net benefit to trade and export for Australia's tech sector by minimising compliance costs when engaging with a major market.

Approach and coverage

Questions	8. Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings? Are there any guardrails that we should add or remove?
elevenM position	We are generally supportive of the approach and coverage of the voluntary and proposed mandatory guardrails as well as the steps taken to align the guardrails with international standards. We recommend that the mandatory guardrails include an obligation to take reasonable steps (including notification where appropriate) to respond to incidents where there are grounds to suspect that that a high-risk AI system has or could cause serious harm. We recommend that voluntary guardrail 10 (stakeholder engagement) be included in the mandatory guardrails in addition to proposed mandatory guardrail 10 (conformity assessments).

Overall approach and international interoperability

We are generally supportive of the approach and coverage of the voluntary and proposed mandatory guardrails as well as the steps taken to align the Guardrails with international standards.[6] Interoperability with global standards and key technology trading partners such has the EU provides significant benefit to trade and export for Australia's tech sector by minimising compliance costs when engaging with major foreign markets. We have seen this play out in the privacy context, where countries such as Israel have deliberately aligned with the EU's GDPR to provide significant benefits for tech trade and exports. By contrast, Australian businesses have faced barriers to trade as a result of key gaps in Australia's privacy regulation, which have prevented our law from being recognised by European regulators as offering 'adequate' privacy protection.

Incident response and notification

With some partial exceptions,[7] the Guardrails do not appear to include any requirements with respect to incident response where there are grounds to suspect that a high-risk AI system has or could cause serious harm.

[6] elevenM has covered the Voluntary AI safety standard on our website https://elevenm.com.au/blog/breaking-down-the-voluntary-ai-safety-standard/ and podcast https://elevenm.com.au/podcast/episode/119-breaking-down-the-voluntary-AI-safety-standard.
[7] We note the requirement in Guardrail 8 for deployers to report adverse incidents to developers, and notification obligations that might arise under other legislation and are incorporated under Guardrail 3 (e.g.: with respect to data breaches).

Deployers of AI systems should be required to have effective incident response processes, to assess instances where there are reasonable grounds to suspect that use of a high-risk AI system resulted (or could have resulted) in serious harm, to take appropriate steps to prevent such harm and to notify any relevant supervisory authority and affected individuals. The EU AI Act and Canada's proposed AIDA both include incident response obligations of this kind. Such a requirement would also be consistent with consumer expectations established by other notification schemes such as for data breaches.

Stakeholder engagement

It is critical for AI developers and deployers to identify and engage with their stakeholders, to understand their needs and circumstances and how they experience of AI systems. Stakeholder engagement is particularly critical for high-risk systems that may have significant impacts on stakeholders' rights and interests.

Voluntary guardrail 10 expresses this requirement well and expresses the importance of stakeholder engagement in developing organisation-level AI strategy and approach as well as at the level of individual AI system deployments. However, this guardrail has been omitted from the proposed mandatory guardrails.

We are concerned that the inclusion of stakeholder engagement in the voluntary guardrails but not in the proposed mandatory guardrails may be interpreted as a suggestion that stakeholder engagement is only necessary for low-risk AI systems, or that conformity assessments demonstrating a system's compliance with the other guardrails are an effective substitute for engagement with affected stakeholders. Neither message is consistent with the intent of the guardrails.

We recommend that the mandatory guardrails include an additional guardrail requiring stakeholder engagement, as per voluntary guardrail 10.

Distribution of responsibility across the AI supply chain

Questions	10. Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately? For example, are the requirements assigned to developers and deployers appropriate?
elevenM position	We support the statement of principle in the Proposals Paper that accountability for guardrails should be distributed based on which actors are best equipped to address risks. In order to avoid accountability gaps, which would be damaging to public trust, parties developing and deploying AI systems respectively should be required to take reasonable steps to ensure that actors upstream and downstream in the AI supply chain have complied (or will comply) with the guardrails. After all, as the world of cyber breaches has shown, the supply chain is only as strong as the weakest link. As the guardrails are further developed, specific accountability for each element should be assigned to developers and/or deployers, or to particular stages within the AI supply chain or system lifecycle. Guardrails and/or

	accompanying guidance should indicate where they may require different actions at different stages of the supply chain (e.g.: testing 'in the lab' vs in a specific operational context).

Avoiding accountability gaps

We agree with the statement of principle in the Proposals Paper:

> *The guardrails should be distributed according to which actors are best equipped to address risks associated with a particular stage of development. This takes into consideration an actor's access to critical information such as training data, and their ability to effectively intervene and change an AI system.*

There is justifiable concern from developers and deployers that they may be held accountable for matters beyond their operational scope and capabilities. However, this principle comes into tension with the harm reduction objective of the mandatory guardrails if it permits developers and deployers to avoid accountability for harms that were ultimately within their control to prevent, even if only by declining to provide or use a given AI system for a problematic use case or by ensuring that the system and its risks are properly understood and able to be managed (mitigated, avoided) by the customer. Public confidence in AI systems will not be well served by accountability gaps or debates about liability between developers and deployers when a system brings about a bad outcome.

We consider that parties developing and deploying AI systems should be required to take reasonable steps to ensure that actors upstream and downstream in the AI supply chain have complied (or will comply) with the guardrails:

- Deployers are ultimately accountable for the impacts of the systems that they chose to deploy and should be accountable for ensuring that mandatory Guardrails are complied with for those systems. Conformity assessments and certifications undertaken by developers (Guardrail 10) should be able to provide deployers with assurance that Guardrails have been met for matters beyond their operational scope or capability.

  For example, while testing a model for bias and ensuring a minimum standard of performance would usually be within the domain of the developer, deployers should (at minimum) be required to satisfy themselves that appropriate testing has been done by the developer, and that performance of an AI model is appropriate for their use case.

- Developers should be accountable for taking reasonable steps to ensure that mandatory Guardrails are complied with by deployers of their systems. This could be done by monitoring use or by including requirements in terms of service.

  For example, while the specifics of a deployment would ordinarily be within the domain of the deployer, a developer should (at minimum) be required to take reasonable steps to ensure that they license their AI system only for use within the scope for which it has been developed and tested and is known to perform adequately.

More granular definition and allocation of obligations

As the guardrails are further developed, specific accountability for each element should be assigned to developers and/or deployers. In some cases, it will be necessary for different versions of guardrail requirements to be drafted for the different roles. For example, testing should be conducted by both developers and deployers, but may be conducted in different ways and with different focuses. It may be more appropriate for developers to test performance 'in the lab' against standardised datasets or industry benchmarks for anticipated use cases, while deployers' testing might focus on performance in a specific operational context, with the specific data or customer demographics that the system will see in practice.

Reducing the burden for small-to-medium sized businesses

Questions	12. Do you have suggestions for reducing the regulatory burden on small-to-medium sized businesses applying guardrails?
elevenM position	The regulatory burden for SMEs can be addressed by: • Development of tailored resources that simplify guardrail requirements, set out model approaches and provide guidance and support, such as SAAM, elevenM's AI Adopt centre. • Enabling AI system vendors to address guardrail compliance on behalf of SME procurers.

Ask SAAM. Safe AI made easy

elevenM is leading the consortium delivering SAAM, one of the Federal Government's new AI Adopt centres. SAAM stands for the Safe AI Adoption Model, a set of simple tools and practical resources to help Australian SMEs navigate AI risk and apply appropriate governance. It is based on and will draw from best practice AI safety frameworks, including the Australian Voluntary AI Safety Standard and will incorporate mandatory guardrails when enacted.[8]

We expect that initiatives like SAAM will bring down the cost of compliance for SMEs over time by simplifying guardrail requirements into a smaller list of actions that are more achievable within an SME. However, even with high quality guidance and support, we expect that full implementation of the mandatory guardrails will be beyond the reach of many SMEs.

Embedding guardrail compliance into AI system packages for SMEs

Rather than expecting SMEs to establish and maintain the range of internal functions required to meet the guardrail requirements, the mandatory guardrail framework should include a mechanism to allow AI system vendors (who may be developers or deployers) to

[8] For more information about SAAM, visit saam.com.au.

provide product packages which incorporate compliance with mandatory guardrail elements that would otherwise be the procuring SME's accountability.

One approach might be through model contractual clauses which would permit a vendor of an AI system to assume responsibility for compliance with the guardrails with respect to their system, provided the purchaser complies with certain usage restrictions and minimum governance requirements (assigns an accountable person, refers complaints to the vendor, etc). This would allow AI system vendors to certify guardrail compliance for a system and use case within certain pre-defined parameters, then sell the system to multiple SME customers without the need for each customer to recertify (provided the customer use falls within the pre-certified parameters).

Alternatively, or in addition, the guardrail framework could permit flexibility with respect to how and by whom guardrail requirements are met. This could allow, for example, a vendor of a biometric identification system create a product targeted at Australian SMEs that supports compliance with the guardrails by that SME by:

- Including as a condition of purchase that the purchaser adopts a model accountability framework and assigns an accountable person in relation to the system (Guardrail 1)
- Licensing the system only for strictly defined use cases in specific industries and geographical locations, for which the vendor has tested the system (Guardrail 4) and completed risk assessments and implemented appropriate risk mitigations (Guardrail 2). The vendor may monitor use of the system to ensure ongoing performance (Guardrail 4) as well as to enforce compliance with usage restrictions.
- Ensuring appropriate security, privacy and data governance measures are applied in the development and deployment of the system (Guardrail 3)
- Providing model communications providing appropriate transparency regarding the system for end users (Guardrail 6) and other entities (Guardrail 8)
- Providing human oversight (Guardrail 5) and complaint processes (Guardrail 7) as a service.
- Maintaining relevant records (Guardrail 9) and certifications (Guardrail 10) to demonstrate compliance with the guardrails

Selecting the right regulatory model

Questions	13. Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach? 14. Are there any additional limitations of options outlined in this section which the Australian Government should consider? 15. Which regulatory option/s will best ensure that guardrails for high-risk AI can adapt and respond to step-changes in technology? 16. Where do you see the greatest risks of gaps or inconsistencies with Australia's existing laws for the development and deployment of AI? Which regulatory option best addresses this, and why?
elevenM position	Option 3 (a whole of economy approach) is our preferred regulatory approach:

- A whole of economy approach aligns to the ways in which AI is deployed and used as a technology, and so would establish a more uniform standard with less duplication of effort.
- A whole of economy approach with a strong central oversight authority is most likely to achieve the government's regulatory objectives to protect people from harm, promote societal wellbeing and enable innovation for economic benefit.

Aligning to the ways AI is deployed and used

AI is a family of technologies that can be implemented in an extremely wide variety of use cases across the economy. Often, single AI systems will be deployed in multiple contexts across different sectors, or even in multiple contexts within a single organisation. The same facial recognition system, for example, may be used in retail, manufacturing or banking sectors, for security, safety, or staff management purposes.

A domain specific approach would see the same system subject to slightly different requirements for each domain or use, enforced by different regulators with different emphasis and interpretation of the requirements. This is likely to lead to duplication, complexity and unnecessary costs for compliance and enforcement. For a company with operations across multiple regulatory domains, deployment of a single AI system under a domain specific approach might require consideration of multiple different versions of the Guardrails, and engagement with multiple regulators.

A framework approach would establish a more uniform standard but would require a similar duplication of effort between domain specific regulators – requiring multiple regulators to build internal AI capability and enforce the same set of standards with respect to the same technology in different contexts. Such duplication could be managed to some extent through cooperation between regulators (such as via the Digital Platform Regulators Forum) but would still present a significant and unnecessary overhead.

Both domain specific and framework approaches would likely be difficult to access for individuals harmed by AI, as they would have to be aware first of a breach of the standards, and second of the correct domain specific regulator for their complaint.

By contrast, a whole of economy approach would provide:

- A single set of standards for AI technology, with consistent and efficient enforcement across the economy.
- Establishment of a central authority to consolidate AI expertise in government and monitor and update the guardrails as technology and best practice develops. This central authority could lend supporting expertise to other regulators dealing with the impacts of AI systems within their traditional regulatory domains.
- A better match to the business model of AI developers, providing a central authority and single standard against which a system can be certified then deployed across the economy.

Achieving the government's regulatory objectives

In order to improve adoption and enable innovation, we must improve public trust in AI. To do this, we must not only be effective in protecting people from harm and prioritising societal wellbeing, but our regulatory framework must be seen and understood by the public as achieving those objectives.

A whole of economy approach with a strong central oversight authority is most likely to be seen and understood by the public as a meaningful regulatory intervention. In addition to signalling governmental priority and intent, a whole of economy approach is more likely to move the needle on public trust by providing:

- A simpler regulatory framework that will be easier for the public to understand
- A single front door for individuals concerned about how they have been impacted by AI.
- A single accountable regulator with a mandate to monitor for and address emergent and society-scale risks, which may arise across or outside of existing regulatory domains.
- A central, authoritative voice on AI safety to support public awareness.

Contributors

Jordan Wilson-Otto · Principal · LinkedIn

Jordan is an expert in privacy regulation, policy development and program management, with deep experience in privacy compliance, regulatory investigations, public policy, and open data. Since joining elevenM Jordan has managed complex privacy project development and implementation and helps our clients ensure they are at the forefront of privacy practice and compliance. Prior to elevenM, he was Assistant Commissioner for Operational Privacy and Assurance at the Office of the Victorian Information Commissioner, where he established and led the investigations and assurance function. Jordan has held leadership roles at GovHack and the Victorian Society for Computers and the Law, and holds a Bachelor of Laws and a Master of Laws from University of Melbourne.

Arjun Ramachandran · Principal · LinkedIn

Arjun is elevenM's communications lead and specialises in providing strategic advice and communications expertise to help organisations build trust in their cyber security and privacy programs, internally and externally. He has extensive experience in crisis communications, board and executive engagement and training, and advising organisations in how to embed privacy and cyber security culture. Arjun worked as a journalist at the Sydney Morning Herald and in senior media relations roles across government and the private sector before establishing the cyber outreach and advocacy program for the Commonwealth Bank of Australia, leading the bank's cyber security strategy, incident response communications and security awareness programs. Arjun holds a Master of Arts in Journalism and a Bachelor of Information Technology from the University of Technology, Sydney.

Melanie Marks · Director · LinkedIn

As privacy practice lead of elevenM, Melanie works with Australia's most prominent brands to drive innovation and manage privacy and data governance risks. In the field for over 20 years, Melanie brings deep expertise and knowledge to clients at all stages, from start-ups to large corporates. She is also committed to training the next generation of privacy professionals. Prior to elevenM, Melanie ran privacy governance programs for CBA and the National eHealth Transition Authority. She is a former President of the International Association of Privacy Professionals in the ANZ region, served on advisory boards for Information Governance ANZ and Hello Sunday Morning and is an Expert Advisor to LexisNexis on privacy and data protection. Melanie holds a Master of Industrial Property Law from the University of Technology, Sydney, a Bachelor of Media (Hons) and a Bachelor of Laws from Macquarie University. She is a Graduate of the Australian Institute of Company Directors.