Responsible AI

Thoughtworks Australia Submission to the Department of Industry, Science and Resources Consultation on Supporting Responsible AI

Thoughtworks appreciates the opportunity to participate in the Australian government's consultation on AI regulation and governance.

Thoughtworks is a leading global technology consultancy that integrates strategy, design and software engineering to enable organisations and technology disruptors to thrive. For over 30 years, we've been at the forefront of digital innovation and have vast experience creating adaptable technology platforms, designing world-class digital products and harnessing the power of data and AI to unlock new sources of value.

Since Thoughtworks started in 1993 in Chicago we've been at the forefront of technology innovation. Now we are some 11,500 people across 51 offices in 18 countries. Thoughtworks Australia was incorporated in 1999 and employs over 400 people. We are known in the industry globally and in Australia for being pioneers and thought leaders in software engineering as well as for the quality of our delivery services.

For 30 years, Thoughtworks has witnessed technology rapidly intertwine with government, society and business to create benefits and opportunities as well as inequalities and perils. Consistent throughout has been our belief that technology and software can be delivered fairly when principles, values, and regulation guide decision-making and behaviour.

Thoughtworks has recently appointed a Chief AI Officer and has produced several reports relevant to this consultation, which inform our answers to the questions posed.

/thoughtworks

First, is our series of White Papers on AI that emerged from an observation that artificial intelligence suffers from hype built on problematic assumptions about what AI can and cannot do. Instead of computer-centric outputs, Thoughtworks emphasise an “intelligent empowerment” or “humanity augmented” approach that places the AI-based computational capability in the service of human experts and their reasoning capabilities. This series explores some of the problematic assumptions in AI narratives and the benefits of having a human relationship-first approach. Some papers discuss information security risks of generative AI through information leakage and vulnerability introduction, proposing practical governance solutions that mitigate these risks.

Second is the Responsible Tech Playbook that explores ways of working that align technology and business behaviour with society’s and individual’s interests. It explores and considers the values, unintended consequences and negative impacts of tech, and actively manages, mitigates and reduces risk and harm. We consider notions of ethics, individual and human flourishing, social structures, inclusivity and equity, civil liberties and democracy. If adopted, the Playbook’s principles and tools can help ensure that world-changing innovations like AI have the right kind of transformational impact on our lives, and that technology doesn’t exploit us — it supports us.

Third is The State of Responsible Technology, a 2023 report we released in collaboration with the MIT Technology Review. This report highlights a deeply complex and heterogeneous field where there are no easy answers that can be universally applied. It also observes that while some business leaders express trepidation about pending regulation, others cite it as important industry guidance. Thoughtworks is of the latter view.

Thoughtworks is using AI on a daily basis. We’ve helped clients apply AI to help them make more sustainable decisions, model the impacts of planned sustainability efforts and identify inefficiencies to progress towards their climate action goals. In one initiative, we used an AI-augmented approach to develop a prototype for Charge Point Operators (CPO). This ‘solver’ tool assists in the design of charging infrastructure networks for electric vehicles (EV). Our approach augmented the planning process with AI, to provide charge site recommendations optimised for utilisation and return on investment. We partnered with a company to create a plugin for social media that helps to filter abusive language and trolling, specifically for those at risk of trolling and gender violence, enabling users to make tech more responsible, even when platforms and vendors themselves don’t implement adequate controls.

General observations on the Safe and Responsible AI in Australia Discussion Paper

We need both broad principles around automated decision making at scale, and then more technology-specific rules. We believe that regulatory governance will be far more effective than a purely voluntary approach.

Our first principles are that human oversight and accountability should always remain in the decision making process, and that transparency of source data and model training is essential.

Several reasonable and complementary elements are necessary for a risk-based approach for managing AI risks in Australia:

Quantitative methods only: A simple risk matrix, register, RAG status or maturity model will not suffice. We suggest further elements to comprise a framework.

Impact Assessments: These are important to identify and mitigate potential harms, especially for medium and high-risk AI applications. Assessments should be proportional to the risk level - more comprehensive for higher risks. Peer review by independent experts could provide greater scrutiny for high-risk AI.

Notices: Notices are valuable to inform individuals when AI/automation is used in ways that significantly impact them. With awareness, it is easier for people to understand AI-enabled decisions or challenge them if needed. Plain language notices should be required for medium and high-risk applications.

Human Oversight: Meaningful human oversight is critical, especially for high-risk AI applications with high potential impacts. Assessments can identify where human oversight is most valuable to minimise risks. However, human oversight needs to be carefully designed considering human limitations. For some lower-risk applications, human oversight may be less necessary or feasible.

Explanations: Explanations are essential for transparency and trust. Individuals affected by medium and high-risk applications should be provided specific explanations about the logic and factors influencing AI-enabled decisions. Access to technical documentation may also be warranted for high-risk applications.

Training: Recurring training proportional to risk level makes sense to ensure proper oversight. For high-risk applications, verification of completed training could be required.

Monitoring & Documentation: More intense internal monitoring and documentation requirements for higher-risk applications are sensible. Independent external audits of monitoring for high-risk AI could provide greater accountability.

Such a framework could help support responsible AI practices in Australia if implemented appropriately based on proportionality and context. Public consultation will help refine the specifics.

Question 1 - Definitions

While there is a need for generic labels in order to broadly refer to a phenomenon, when these labels refer to technical tools there is the potential for confusion and conflation, which is currently happening in much of the discourse around Artificial Intelligence (AI).

In order to create meaningful legislation or frameworks that govern the use of AI, the technology must be clearly and realistically defined in technical terms, rather than referred to under a single label like “AI” which is often co-opted for marketing purposes to refer to a broad range of technologies. We recommend taking a more nuanced view of the technologies involved when it comes to considering the risks and impacts of these technologies from a governance standpoint.

AI is a branch of computer science that aims to create systems capable of performing tasks that would typically require human intelligence. These tasks include learning and adapting to new information or environments, understanding human language, recognizing patterns, solving problems, and making decisions.

AI can be categorised into two main types:

Narrow AI: These are systems designed to perform narrow tasks (e.g., facial recognition or internet searches) and can only operate under limited constraints. At present, all of the AI that we encounter on a day-to-day basis is narrow AI.

General AI: These are systems that possess the ability to perform any intellectual task that a human being can. They can understand, learn, adapt, and implement knowledge in different domains. This type of AI is currently theoretical.

AI technologies use different approaches to try to understand and learn from data. The most common is **machine learning,** where algorithms are trained on a large amount of data and then apply this training to new data.

Deep learning, a subset of machine learning, uses neural networks with many layers (hence the "deep") to carry out the process of machine learning. The more parameters these models contain, the more powerful, less understandable, and more risky they become.

Generative AI, another subset of AI, uses these deep learning techniques to understand and learn from data and create new, original outputs similar to the learned data.

Composite AI is a term which recognises that specific solutions are particularly suited for certain tasks, while solutions with broader applicability may be also composed from individual solutions. By this definition, AI also includes statistical and probabilistic techniques, graph analytics, simulation, optimisation, etc, individually or in composition with each other and machine learning.

2. What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

Lack of transparency and explainability: Existing laws do not expressly require transparency around AI/ADM or explain how systems operate and make decisions. New regulations should mandate transparency and documentation requirements for such technologies where they are used in high-risk contexts or by governments.

Algorithmic bias and discrimination: While anti-discrimination laws offer some protection against current algorithmic harms, they can be difficult to enforce against machine-amplified biases. New laws expressly prohibiting algorithmic discrimination, as well as measures for its reporting, detection, and enforcement, are necessary in this context.

Spread of misinformation/disinformation: Generative AI tooling, in particular, has significant potential to be abused to produce mis- and disinformation at a large scale. Broader laws may be needed to address this use of AI tooling, which may include a requirement to disclose the provenance of AI-generated content.

Privacy risks from large language models: Specific regulations around data collection and consent could mitigate privacy risks posed by large language models trained on massive datasets. We note the recent Privacy Act Review from the Attorney General, and the consequential report released this year. With the Privacy Act already back on the drawing-board, now is the right time for the Government to take this opportunity to examine enhanced privacy protections and breach enforcement measures that address information derived from (and used) by AI.

Validity and safety risks: Sector-specific safety regulations may be needed as AI permeates high-risk fields like healthcare and transport. These regulations should include mechanisms to maintain systems-level resilience to poor AI performance, and impose sufficient liability on model providers or developers for harm caused by the negligent implementation of an AI-based system. A body such as Standards Australia could assist by developing responsible AI standards and experimental protocols.

Risk of industrial displacement: Adoption of AI may impact the ability of workers or organisations to benefit from their own output, or may reduce opportunities for workers to acquire valuable skills. Use of these technologies must come with appropriate legal protections for impacted workers and industries.

3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

Maintain and evolve Australian institutions for Responsible AI - The Responsible AI Network is an exemplar and should be supported by companion institutions. Such independent bodies should exist to provide services such as conducting AI testing and audits, providing guidance on safety-critical AI, and connecting practitioners and industry with expertise in the responsible development of AI.

Expand educational programs - Fund additional university scholarships or sponsor more AI ethics challenges to grow Australia's talent pool in responsible AI and raise awareness. This should include deeper investment in the humanities and inclusion of humanities education in computer science disciplines, as many issues connected to AI raise significant ethical and human rights concerns.

Create a public AI registry for government AI use - Require government agencies to publish details of public-facing AI systems they use, in order to improve transparency and oversight.

Launch a Responsible AI rating system - A voluntary rating scheme may help Australians identify AI products and services that meet defined transparency and responsible development criteria, in addition to providing incentive for the development of responsible and safer systems.

Sponsor research into explainable/auditable AI - the current generation of deep learning models produce results that cannot be reasoned about. Research into explainability for these models is in its infancy and must progress so that the risks of bias and output validity in AI systems can be meaningfully mitigated.

Prohibit autonomous weapons systems - Support multilateral efforts to prohibit and regulate autonomous weapons systems, ensuring meaningful human control, oversight and decision making regarding the use of force.

Provide support for workers and industries that have been disproportionately impacted by AI technologies - Adoption of AI may impact the ability of workers to benefit from their own output, may reduce opportunities for workers to acquire valuable skills, and may disproportionately disadvantage those in creative industries. Consider support programs to maintain valuable creatives and skilled workers, and comprehensive transition pathways for impacted professions.

4.Do you have suggestions on the coordination of AI governance across the government? Please outline the goals any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

We suggest that a strong program of ethics, education, research, transparency, and public engagement initiatives be established in order to complement regulatory measures to create an AI governance ecosystem, aligned with Australian values and interests, that is coordinated and consistent. As part of this program, we recommend the following:

Establish a national AI ethics advisory council comprising government, industry, academia, and civil society representatives. This body should exist to identify emerging AI risks, provide ethical guidance, and foster coordination between industry and different branches of the government.

Develop an AI accountability framework outlining best practices for organisations to assess, document, and mitigate AI risks. Algorithmic impact assessments should be included as a part of this framework. It is strongly recommended that such frameworks are mandatory for relevant applications of AI technology to high risk use cases and sectors, especially government applications. Any such frameworks should be compatible with existing industry frameworks and regulations to encourage ease of adoption.

Sponsor challenges and incentivise research into the development of responsible AI, thorough testing methodologies, and practical techniques to minimise algorithmic bias. Facilitate data-sharing frameworks and computing resources for AI researchers. More comprehensive data access assists the development of fair and representative models, and investment in this field is likely to yield technical insights that will inform future governance.

Fund civil society groups to act as watchdogs, test AI systems, and advocate for public interests. Independent oversight should complement government efforts and ensure that the governance systems include robust checks and balances.

Support education programs to build AI expertise within government, regulators, and civil servants, in order to enable effective oversight of related issues. Firmly include the arts, philosophy, social sciences, and humanities in the curriculum.

Encourage professional AI associations to institute voluntary certification programs for developers that signal a practitioner's commitment to ethical AI and ADM practices.

5. Are any governance measures being taken or considered by other countries (including any not discussed in this paper) relevant, adaptable, and desirable for Australia?

As technological developments have international impacts, regardless of geography, it is imperative that Australia participate in international collaboration on AI research ethics, risk assessment frameworks, testing methodologies, and labelling standards. We should have an active part in the development and coordination of global AI governance measures, and leverage existing international governance measures where they are relevant. As this discussion is robust, there is much we believe Australia can draw on and adapt for our own use, especially localised versions of governance mechanisms focused on algorithmic transparency, auditing, explainability, and accountability.

The EU's AI Act presents a solid framework for approaching regulatory issues regarding AI. Its re-statement of importance (and possible pre-requisite of) the GDPR, the proposal for a Critical Entities Resilience Directive, would impose cybersecurity and due diligence requirements for entities like AI providers critical to the economy and society. A similar approach could enhance the reliability and safety of impactful AI systems. Additionally, the EU's proposed restrictions on specific uses of biometric identification and social scoring systems could inform similar limitations in Australia.

In the United States, the proposed Algorithmic Accountability Acts require impact assessments, risk management procedures, and external audits for specific public agency automated decision systems. Elements of this approach could foster accountability in Australian Government AI use. The US AI Bill of Rights principles could be a starting point to develop similar guidance appropriate for the Australian context. Further examples can be found in the US state-based legislation that governs AI use in specific sectors such as employment and education at state levels, demonstrating an approach to context-specific, targeted legislative intervention.

Australia may also be able to draw on elements of the UK's laws, in particular the obligation for public agencies to publish Algorithmic Transparency Reports that are intended to explain when and how algorithms are used in government decision-making. The creation of oversight bodies for AI and specific AI regulatory agencies in the UK (Centre for Data Ethics and Innovation) and Canada (Advisory Council on Artificial Intelligence) should also serve as an example for Australia to emulate.

In addition to the region-specific examples cited above, Australia should join the international collaboration efforts on the prohibition of fully autonomous weapons systems, and lead by example.

6. Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

Thoughtworks believe some differentiated approaches may be appropriate for public versus private sector use of AI technologies:

The public sector should be subject to more stringent transparency, accountability, and human oversight requirements. This higher standard helps build public trust.

Mandatory algorithmic impact assessments could be required for all public sector AI projects but be voluntary for the private sector in certain applications.

Blanket requirements to publish technical details of AI systems may stifle commercial innovation, so that limited transparency requirements may be more appropriate for the private sector.

The public sector could be obligated only to procure AI systems adhering to specific safety or ethical standards, while standards could remain voluntary for the private sector.

A Responsible AI assurance or certification regime could be mandated for public sector systems but be voluntary and market-driven for the private sector.

The private sector may need more flexibility in using AI for efficiency gains, so human oversight requirements could be less stringent than for high-risk public sector applications. The responsibility for compliance might become a lagging indicator and move to a post hoc audit.

AI use for national security, defence, and law enforcement should warrant customised, government-specific requirements and oversight.

The public sector should exemplify responsible and transparent AI use, so more stringent oversight, explainability, and accountability requirements may be justified. Private sector obligations could focus more on providing innovation flexibility while addressing potential consumer harms. Multi-stakeholder consultation can determine suitable tailored approaches for each sector.

7. How can the Australian Government further support responsible AI practices in its own agencies?

The Australian Government should lead by example on responsible AI development and deployment. Investing in oversight, transparency, documentation, coordination, and staff competency is essential to the deployment of systems that are intended to serve the Australian public, and to establish public trust.

We recommend that the government incorporate the following approaches, or applicable elements thereof, for use in its own agencies:

- Mandate algorithmic impact assessments for all AI projects to evaluate risks and mitigation strategies before, during, and after a project's development.
- Require plain language explanations of agency AI systems to improve public transparency.
- Institute human-in-the-loop checks for high-risk public sector AI applications.
- Create an interdepartmental AI coordination office to develop consistent AI usage and development policies, and to oversee appropriate validation of systems before they are deployed.

- Establish an independent oversight body to audit agency AI use, investigate complaints, and enforce policies.
- Implement AI procurement standards to acquire systems adhering to ethical principles and technical robustness.
- Develop minimum AI competency requirements for public servants overseeing AI systems.
- Publish annual algorithmic transparency reports detailing the government's use of AI.
- Appoint dedicated AI accountability officers within agencies to monitor internal use.
- Fund research into AI auditing methods and tools tailored to public sector contexts.
- Require agencies to maintain detailed documentation of their AI systems and make them available for oversight.
- Create a public AI ethics advisory council to engage civil society on government AI initiatives.

9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:

a. where and when will transparency be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

It should be noted that deep learning models are, by their nature, opaque. While transparency of training data, methodologies, algorithms and performance is desirable, no amount of transparency will reveal the inner workings and reasoning processes behind the decisions these models make. This fact should inform every decision to implement AI technologies in a product or process.

If, having considered the above, it is still believed to be useful to implement an AI-based solution in any given context, then thorough documentation and transparency should be a requirement at all stages of development, as there are many points at which influence may be exerted on the eventual output of a model. The efficacy of any framework, governance model, or regulatory mechanism put in place around AI technologies is fundamentally dependent on such transparency in order to function in any meaningful sense.

In order to be as effective as possible, documentation and evidence should be provided for all stages of the AI life-cycle, including:

- data sourcing, labelling, and preprocessing of foundation models
- design choices, architectural features intended to ensure safety, and training approaches during model development
- testing scenarios, model limitations, and plans for monitoring post-deployment
- model revisions, monitoring insights, and processes to remedy harmful failures post-deployment
- the entities and individuals responsible for each phase of development

This information should be available for all models in use in public sector applications, as well as for private sector applications above specified risk thresholds or particular use cases.This should go hand-in-hand with requirements matching these use cases that mandate algorithmic impact assessments, subject to public/expert review, as well as mechanisms for public inspection of operational models, audit logs, and ongoing monitoring. Given that this field is rapidly developing, it is important to pilot a range of implementations with stakeholders in order to determine how this should function in practice and what may be necessary for these measures to evolve adequately over time.

10. Do you have suggestions for

a. Whether any high-risk AI applications or technologies should be banned completely?

b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

This is a complex issue that requires careful consideration of both risks and benefits. Some thoughts:

a) Complete bans could be warranted for highly unethical AI applications with limited societal value, like

- Indiscriminate biometric surveillance systems that violate privacy/civil liberties.
- AI that is designed to target and harm human beings autonomously.
- Highly manipulative AI aimed at exploiting vulnerable groups.
- Uncorrectable flawed/unsafe AI in ultra-high-risk scenarios like surgical robots.

b) Criteria to identify categories of AI warranting bans could include:

- The risk of irreversible deprivation of human rights or significant human harm.
- The absence of sufficient use case validity, utility, or countervailing benefit.
- The existence of safer technological alternatives to achieve the purposes.
- An unacceptable susceptibility to misuse, bias, or error.
- Insufficient remediation, reversibility, or containment of adverse impacts.
- Overriding existing legal/ethical prohibitions if deployed.

However, bans may be premature for many categories of AI with both beneficial and concerning applications. Contextual prohibitions in limited high-risk use cases may be more prudent than outright bans. For example, while general biometric surveillance could be banned, uses like identifying traffickers at airports could remain legal. Multi-stakeholder deliberation and impact analysis should inform restrictions. Any bans should also be regularly re-evaluated as technologies and use cases evolve.