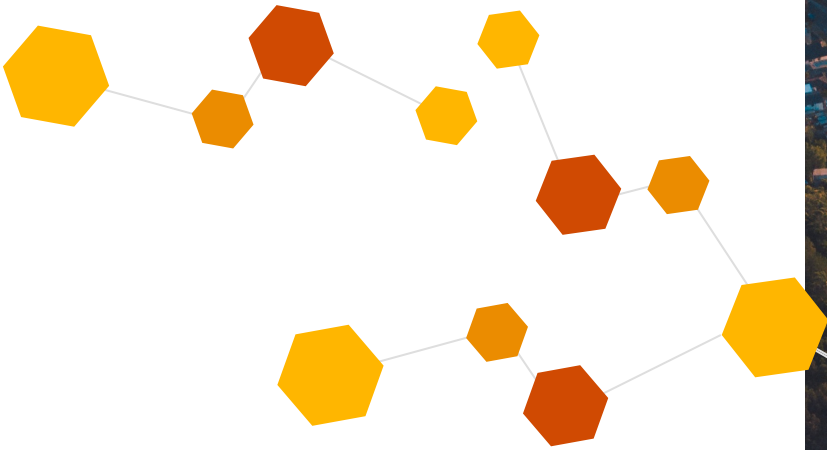


# Safe and responsible AI in Australia

PwC's response to industry consultation  
July 2023

**Artificial Intelligence**  
Accelerate responsibly.



# Contents

Executive summary	1
Defining the scope for AI regulation	3
Strengthening AI accountability measures	6
Improving transparency across the AI value chain	9



# Executive summary

Recent breakthroughs in research and development and a rapid uptick in global investments for Artificial Intelligence ('AI') have the potential to accelerate transformations in the way we work, produce information and interact with technology.

AI-enabled transformations represent a significant opportunity for Australia's innovation and growth agenda. There is limited up-to-date data available for 'sizing the prize' in Australia, however back in 2018 a report commissioned by the CSIRO projected \$315 billion in added economic value from AI by 2028.<sup>1</sup>

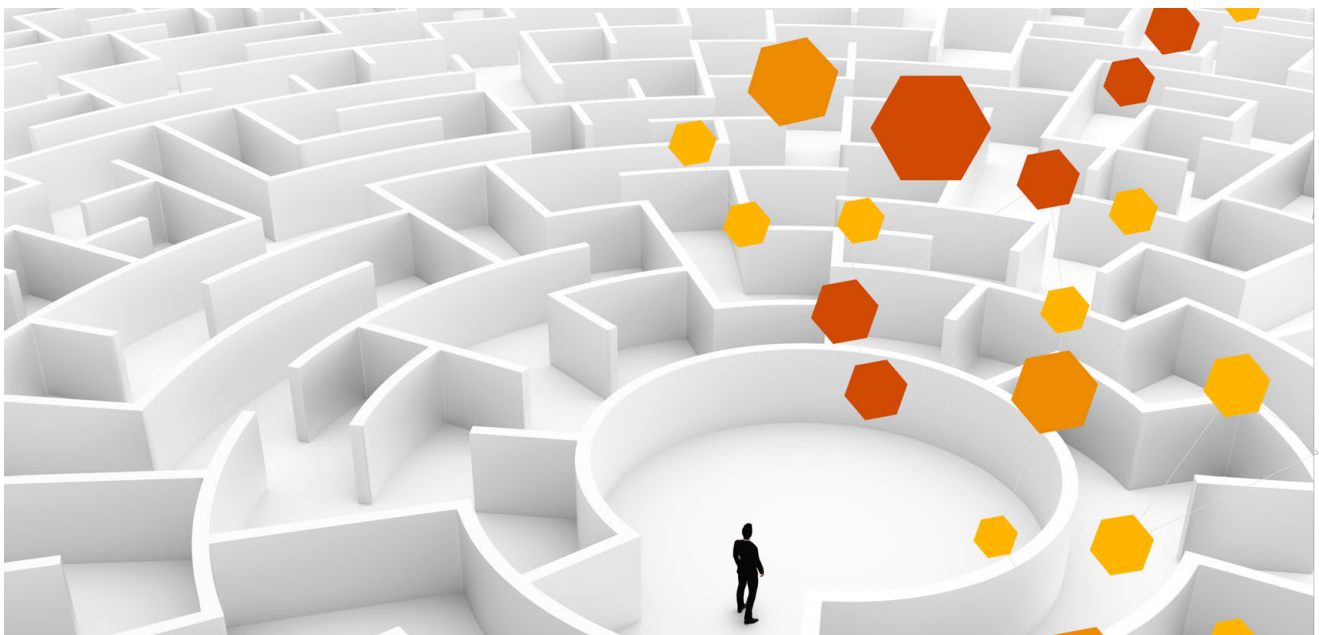
Importantly, to realise this growth, Australian businesses need to be sufficiently confident and competent in developing, procuring and applying AI safely and responsibly, and Australian consumers and employees need to be willing to trust it.

Today, less than 1 in 20 Australian businesses are well-versed (mature) in navigating the risks of AI, fewer than 4 in 10 consumers trust AI and only half of employees feel positively towards it.<sup>2</sup> Confidence in new technology often influences its rate of adoption, and adoption rates for AI in Australia remain relatively low.<sup>3</sup>

From working across a broad range of industries, we have observed several common concerns and challenges for the adoption of AI in the enterprise:

- Gaps in the knowledge of how AI works, the risks associated with using it and managing its use in-line with social and ethical expectations.
- Difficulty interrogating and explaining the outputs of models or providing the level of transparency that regulators might expect in the absence of specific 'black letter law' for AI.
- Difficulty foreseeing the damage that could arise in the event of faults, errors or oversights in the way that AI systems are designed or applied.
- Unclear expectations, which can vary between sectors, for the appropriate assignment of accountability, ownership and liability for the outputs of AI models, especially where AI systems are created through complex and interdependent value chains.

For the reasons above, and more, we find that a number of organisations are not ready to apply AI safely and responsibly across their core products, services and processes.



<sup>1</sup> AlphaBeta Advisors (2018) Digital Innovation: Australia's \$315b opportunity, report to the CSIRO Data61.

<sup>2</sup> Fifth Quadrant (2023) Research shows a worrying lack of action towards responsible AI. University of Queensland (2023) Trust in Artificial Intelligence. PwC (2023) Global Workforce Hopes and Fears Survey.

<sup>3</sup> Department of Industry, Science and Resources (2023) Safe and responsible AI in Australia.

We believe that a substantial and timely uplift in organisations' maturity and public trust in AI will be necessary for Australia to unlock the full benefit of AI-enabled transformations.

There are three key areas where we believe that further policy measures could support this:

### Defining the scope for AI regulation

Both the definition of 'AI systems' and a risk-based approach to classification are pivotal for shaping the breadth and impact of AI regulations in Australia.

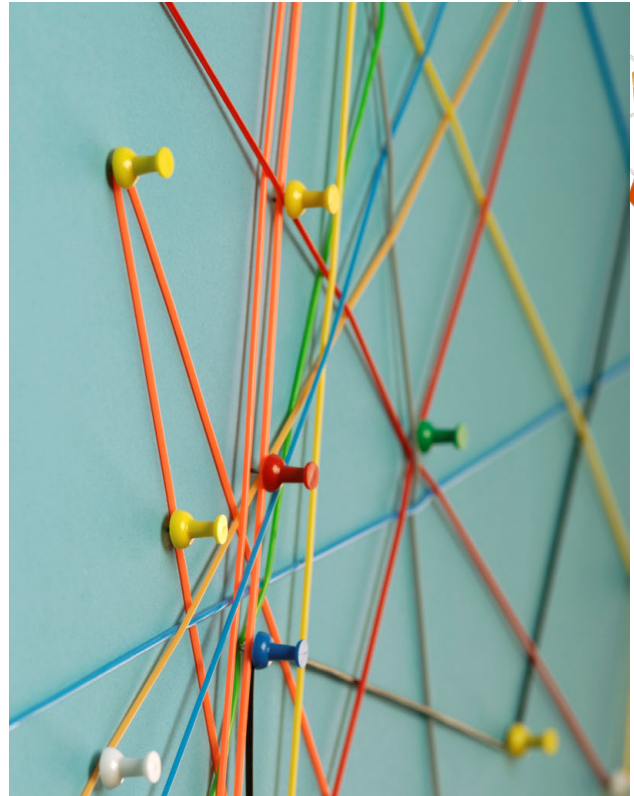
We believe that the defining criteria for AI regulation should go beyond how it is developed and functions, and should also address the way that an AI system is being applied. We also support DISR's proposal for a risk-based governance approach – one that is centred around the impacts that AI systems have on people, society and ecosystems – so that regulatory efforts are directed towards the AI systems which pose the highest risks.

It will also be important that the definition for 'AI systems' is globally harmonised. The definition supplied in the Discussion Paper is adapted from ISO/IEC 22989, however this definition has not been adopted by key global bodies, such as the EU Council, the Organisation for Economic Co-operation and Development ('OECD'), the US National Institute of Standards and Technology ('NIST') and the US National Telecommunications and Information Administration. Other globally aligned definitions may be more fit-for-purpose.

### Strengthening AI accountability measures

We believe that mandatory AI governance principles, with independently assured management reporting obligations for higher risk AI systems, could help to accelerate the adoption of AI in Australia by building confidence and consensus among organisations and consumers.

These governance measures could be modelled off the 'gold standard' accountability systems that are used globally to build trust in financial information. This model has already been adopted for other non-financial use cases such as environmental sustainability and cybersecurity accountability and reporting.



### Improving transparency across the AI lifecycle

Prescriptive transparency disclosures for AI systems may be challenging in practice due to the varied ways in which AI systems are used. A potential alternative would be to adopt a principles-based approach for transparency. The specifics of how organisations adhere to this principle could be determined by the organisations themselves or through sector-specific guidance and standards.

Any prescriptive transparency requirements, such as mandatory notices for individuals subjected to AI decision-making, should be reserved for higher risk AI systems that could pose serious harm to an individual's rights and opportunities.

\*\*\*

There are other important measures – beyond governance and regulation – that could also help to drive the safe and responsible use of AI in Australia. For example: research and development incentives, innovation grants and investments in training and skills - early childhood through to tertiary. Policy measures beyond AI governance and regulation are not explored in this paper.

# Defining the scope for AI regulation

Both the definition of 'AI systems' and a risk-based approach to classification are pivotal for setting the scope of regulatory requirements and the potential magnitude of regulatory impact on the Australian economy.

For instance, a broad definition for AI systems would likely need to be paired with a risk-based tiering approach to limit the burden of regulatory obligations for organisations working with lower risk AI systems. On the other hand, risk-based tiering may be unnecessary if the threshold for regulation is more targeted and the level of adverse impact arising from AI systems is incorporated into the definition itself (i.e. forming a comprehensive definition for 'regulated AI systems').

Definitions and risk tiering approaches are discussed together in this section due to their interdependence.



*Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?*

Question 1

The definition for AI systems that was proposed in the Discussion Paper may result in traditional probabilistic and non-deterministic statistical techniques (that have long been commonplace) becoming in-scope for AI regulation. This could create a significant overhead for both regulators and regulated entities without necessarily addressing the targeted AI risks and challenges outlined in the Discussion Paper. For example, Monte Carlo simulations that aid in financial forecasting could be interpreted as 'in scope' under the proposed definition.

Inversely, the proposed definition currently excludes systems with 'explicit programming', however, deterministic (rules-based) decision engines, such as expert systems, can pose a substantial risk of harm to people and society even though they are produced with explicit programming. For example, a system that advises a patient on whether they should seek medical advice based on their symptoms.

We believe that the definition – or supplementary remarks – that form the basis for regulation should encapsulate how AI is used and applied, rather than limiting the definition to the AI system's development methods and functionality. By doing so, it may be possible to demarcate technology systems involved in AI processing based on the specific risks of harm that the regulatory framework seeks to address.

For example, the *US Blueprint for an AI Bill of Rights* contains an exclusion of 'passive computing infrastructure' which does not 'meaningfully impact individuals' or communities' rights, opportunities, or access'.<sup>4</sup>

<sup>4</sup> The White House (2022) *Blueprint for an AI Bill of Rights*.



We believe that there is also an opportunity to clarify the scope of the term ‘system’. The proposed definition references an ‘engineered system’ which we understand to be the ‘technology solution’. The risks of AI harm can often arise from how technology is applied and used, rather than only how it was designed and developed. The definition could be broadened to encapsulate the ecosystem of processes, practices and behaviours that influence how the technology solution is applied.

### Global interoperability and harmonisation

The definition for AI that was supplied in the Discussion Paper is adapted from the definition of an ‘AI system’ used in ISO/IEC 22989. For global interoperability, we generally support alignment with international standards, however, at the time of writing, there is yet to be widespread global adoption of ISO/IEC 22989.

For example, the definitions of AI adopted by the OECD, NIST, in the latest proposed amendments to the EU AI Act and in the US Blueprint for an AI Bill of Rights do not currently align with the ISO/IEC 22989:2022 definition, which means that the global interoperability and harmonisation and optimisation benefits may be limited.<sup>5</sup>

On 14 June 2023, the EU Council and Parliament opted to align with the definition applied by the OECD and NIST. Specifically:

*“a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations,*

*or decisions that influence physical or virtual environments.”*

This definition forms what may currently be the most widely adopted global baseline.

### Other definitions

The Discussion Paper proposes definitions for other related concepts, such as machine learning, generative AI and automated decision-making.

The proposed definition of *Machine Learning* refers, somewhat recursively, to patterns derived from training data using machine learning algorithms. It is unclear whether the definition should be used to identify a process through which models are optimised, or a type of algorithm that generates predictions based on input data. Internationally, there has been long-term misalignment in the distinction between machine learning and artificial intelligence and, in many cases, we observe these terms being used synonymously.

The proposed definition of *Generative AI models* may not sufficiently reflect how generative AI systems function and could be broadly interpreted as covering any system that randomly generates content. We believe that it will be important to encapsulate the core concept of machine learning into this description to differentiate from other methods of content generation.

Overall, a nationally agreed definition for machine learning and generative AI models may not be required if accountability and governance measures are scoped at the broader level of ‘AI systems’.

<sup>5</sup> OECD (2019) Scoping the OECD AI Principles. European Parliament (2023) Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). The White House (2022) Blueprint for an AI Bill of Rights.

*Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach? What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome? What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?*

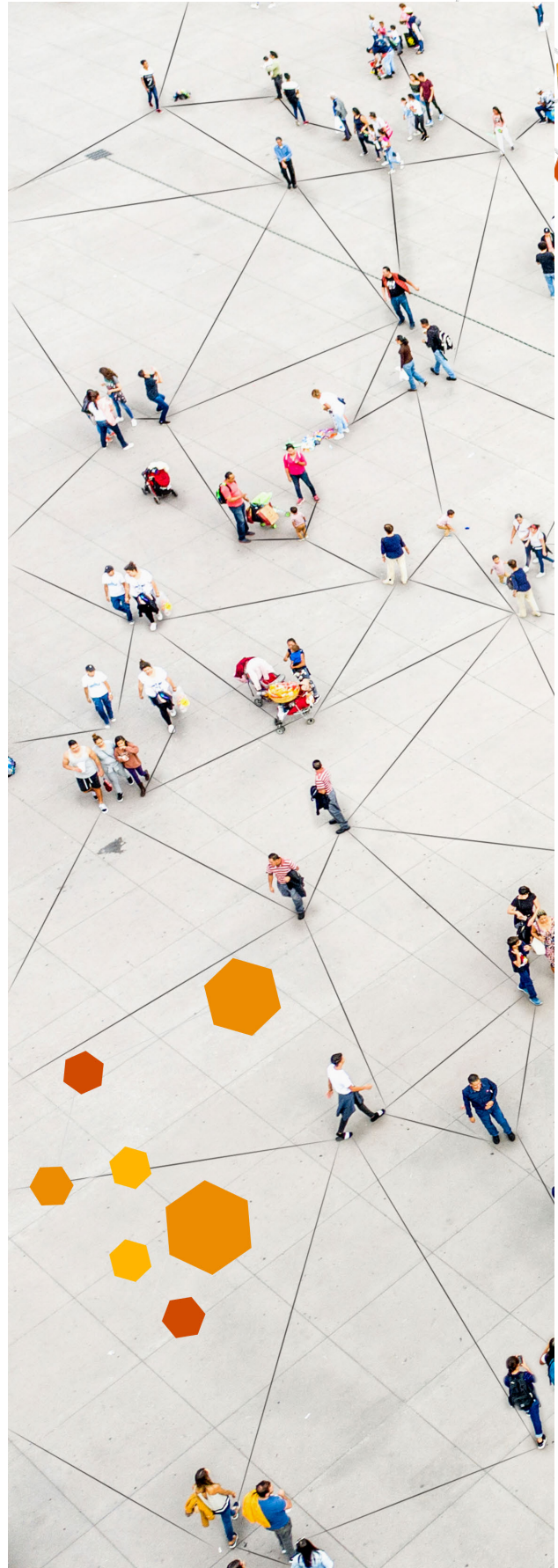
Question 14, Question 15, Question 17

A risk-based approach for classifying AI systems is consistent with approaches proposed in other jurisdictions, such as the EU and Canada, and could enable better targeting of governance and oversight efforts towards the AI systems with the most serious risks of harm.

The adoption of a risk tiering system will be particularly important if Australia adopts a broad definition for AI systems, as it will help to avoid overly burdensome compliance requirements for organisations that develop, procure and apply AI systems that only pose a limited risk of public harm; for example, an AI system that detects anomalies in cloud computing spend.

*Attachment C* in the Discussion Paper contemplates a three-tier risk scale that varies based on the 'risk of harm' including example use cases. AI use cases will vary with time and – in addition to examples – it will be important that the categories of public harm (e.g. harm to people, society and ecosystems) and the maximum acceptable thresholds for each risk tier are defined clearly. These thresholds will have real practical implications on AI adoption in Australia and the benefits and costs will need to be balanced carefully. For example, the proposed explainability obligations at higher risk tiers could result in certain types of deep learning models becoming *de facto* prohibited for higher cost of failure use cases, even if they happen to be more precise.

Given the broad and diverse applications of AI in Australia, there may be a need for more tailored, sectoral level classification guidance for consistent interpretation and application of the regulation within each sector. This could include, for example, applications of AI systems for health diagnosis and treatment, providing financial advice, approving insurance claims or assessing lending eligibility.



# Strengthening AI accountability measures

*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?*

## Question 2

AI accountability measures help to provide clarity of expectations and appropriately balance the promotion of AI innovation against the need to protect people, society and ecosystems from unacceptable harm.

Not dissimilar to the way we build trust in financial information, we believe that trust in AI systems and the data that feeds them could ultimately be achieved through a combination of a management assertion on compliance with an AI framework, and independent assurance on management's assertion.

We believe that this model could help to accelerate the adoption of AI in Australia by building confidence and consensus among organisations, investors and consumers.

## Australia's current regulatory environment

AI systems in Australia are currently regulated by a patchwork of existing obligations, including those related to intellectual property, employment, surveillance, human rights and anti-discrimination, privacy, corporations and consumer protection laws.

Examples of existing mandatory obligations that may apply to AI systems under Australian laws include (but are not limited to):

- Transparency and consent requirements for data collection and use under the Privacy Act and Australian Privacy Principles.
- Liability for AI products which cause injury or property damage due to safety defects.
- Refraining from misleading and deceptive marketing practices under the Australian Consumer Law.



*(continued)*


- Refraining from unjustified discriminatory treatment of individuals on the basis of sex, sexual orientation, age, disability, race or ethnicity.

While existing legal principles which apply to the governance of AI systems are in law, their interpretability and applicability to AI systems is still evolving and there are few examples of enforcement of these obligations in the context of AI technologies. Limited precedent and minimal regulatory guidance exist to provide confidence to the market that organisations are meeting their compliance requirements appropriately.

## Designing an AI accountability framework

In developing an AI accountability framework, we recommend that policy makers look to the financial reporting ecosystem as the gold standard in ensuring the reliability of, and market confidence in, company-specific information. The financial markets trade and rely on the information reported, enabled by trust in the clear roles and responsibilities of each of the parties – ranging from regulators, standard setters, investors, businesses, and auditors. The emerging sustainability ecosystem is leveraging the baseline established by financial reporting and we





believe that this model could also be emulated in support of the trusted and sustainable use of AI.

An AI accountability framework should centre around binding AI governance principles. Existing (non-binding) frameworks, such as Australia's AI Ethics Framework, could potentially be evolved for this purpose. Alternatively, principles from global best practice frameworks, such as the NIST AI Risk Management Framework, could be adopted.

Organisations should then be able to operationalise the principles with the support of risk management frameworks. For example, the Committee of Sponsoring Organizations of the Treadway Commission (COSO) Framework was established in 1992 for organisations to exercise oversight in designing, implementing, and conducting internal control. As such, COSO or a similar framework could be applied by companies and their boards to establish a system of control for adhering to the AI governance principles.

### **Managing the impact of regulation on the Australian economy**

It is important that any governance and policy measures are practical for all stakeholders in Australia's AI ecosystem. An AI regulatory response needs to be fit-for-purpose not only for global and local corporations, but also for researchers, entrepreneurs and small to medium enterprises (SMEs). SMEs alone make up almost one-third (32%) of Australia's total economy.<sup>6</sup>

To assist in streamlining compliance for organisations, we believe that AI accountability measures should be implemented as consistently as possible across related risk domains such as security, privacy and data governance. Furthermore, for organisations that operate across different sectors and local jurisdictions, a nationally consistent approach (in contrast to a sector-led or state/territory-led approach) could help to provide the required coordination and unification. This is particularly important for Australia due to the relatively small market size (in comparison to other

countries) and the need to ensure local compliance costs are not disproportionate to market opportunity.

Similar to the Privacy Act, a turnover threshold could be applied to reduce the cost of compliance for SMEs that do not operate in higher risk industries, handle higher risk data or undertake higher risk business activities; however, if this approach is adopted, the points of intersection between regulated corporate entities and the obligations of unregulated entities will be important to manage closely.

Our experience working across other risk classes (e.g. cybersecurity and privacy) has demonstrated that the interconnectivity across supply chains, where large corporates consume services from smaller (potentially unregulated) entities, has proven to become a point of weakness for managing risks associated with technology and data. There is a key role for big business to play in supporting SMEs.

Mechanisms such as third-party assurance reporting could be useful for regulated entities to establish trust in (potentially unregulated) service providers. Existing networks of trusted business advisors (e.g. accountants and tax agents) could also be engaged - with the appropriate training, funding and incentives - to help establish a higher level of awareness and literacy of good practices for AI governance among unregulated entities and SMEs.

### **Appointing a suitable regulatory body**

To design, operationalise and enforce the AI accountability measures, a suitably empowered and adequately resourced regulatory body will be required. The oversight body will also require deep and independent technical expertise.

We suggest that the body provides a mechanism for organisations to request for binding rulings to be issued by the regulator to help provide clarity around the correct interpretation and application of the AI governance principles. This will enable incremental refinement of legal principles as the technology matures and provide a means of disseminating those principles across the market, without the need for frequent legislative reform processes.

---

<sup>6</sup> Australian Small Business and Family Enterprise Ombudsman (2020). Small Business Counts December 2020



## Management reporting

Where AI systems could result in a high risk of public harm, management reporting for compliance against the AI governance principles could help to further build trust by demonstrating that organisations' internal control environments for AI systems are well-designed and operating as intended.

Reporting standards should be developed to provide users with an instant 'shorthand' regarding the reliability of an AI system, even if the user does not have foundational knowledge of the subject matter or the related disclosure requirements. Such reporting standards could be designed to enable 'digital first' management reporting, such as an authenticated trust icon on an AI system, signed digitally by the organisation.

It will be important that any management reporting regime that is introduced balances the effort and cost of reporting against the value it delivers to users and consumers to avoid creating a 'tick box exercise' for organisations.

## Independent assurance

Although management ultimately has responsibility for the reliability of the AI systems that it develops and applies, our global research shows that key external stakeholders, such as investors, have more confidence in the information that they use when it has been independently assured.<sup>7</sup>

Investors have high expectations of the organisations that provide assurance and – based

on our research<sup>6</sup> – we believe that all AI systems auditors' should have:

- deep expertise in the technical design, development, deployment and operation of AI systems,
- extensive experience in attestation methodologies with appropriate quality management systems in place,
- professional licensure requirements that mandate the achievement of educational and technical competency and continuing professional education, and
- internal frameworks that reinforce the assurer's independence, integrity and objectivity.

The extent of testing and independent assurance requirements should be determined in alignment with a risk-based approach, such as the one discussed in the 'Defining the scope for AI regulation' section of this response.

The guidelines for independent assurance should also be prescribed as standards to provide a universal and consistent baseline for the level of confidence that users should expect. Use of the terms 'audit' or 'assurance' without reference to a generally accepted standard can fail to convey the level of effort applied, the scope of procedures performed, the level of assurance provided, or the qualifications of the provider, among other shortcomings.

---

<sup>7</sup> PwC (2021) Global investor survey: The economic realities of ESG.



# Improving transparency across the AI value chain

*Given the importance of transparency across the AI lifecycle, please share your thoughts on (a) where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI, and (b) mandating transparency requirements across the private and public sector, including how these requirements could be implemented.*

## Question 9

A prescriptive approach to transparency disclosures for AI systems is challenging in practice due to the varied ways in which AI systems are used across an organisation's value chain. For example, consumers may not expect disclosures on outputs from AI systems if those outputs have been through detailed, expert human review. However, it would be incredibly complex to define and enforce a threshold for human oversight and review of AI outputs that – when met – removes the requirement for disclosure.

To strike a balance between the cost of compliance and the adequacy of AI protections, we believe that any prescriptive transparency requirements, such as mandatory notices for individuals subjected to AI decision-making, should be reserved only for higher risk AI scenarios that could potentially impact individuals' rights and opportunities.

For other scenarios, a principles-based approach could enable organisations to identify and implement suitable solutions themselves (supported with sector-specific guidance and standards), that align with the expectations of information recipients and impacted individuals.

## The importance of transparency in AI

Transparency is a critical element of effective AI governance and risk management. The principle of transparency enables individuals to 'understand when they are being significantly impacted by AI' as well as ensuring system testers, operators and users of AI systems can understand how the system functions, troubleshoot issues and address root causes of errors.<sup>8</sup> The ability to transparently understand when and how AI systems affect individuals also underpins other ethical AI principles, such as contestability and accountability.

Providing transparency to individuals affected by the outputs of AI systems affords respect and meets social expectations about the way institutions and organisations interact with individuals. Having visibility of the rationale for decisions enables individuals to understand the factors which were taken into account and to challenge the validity of those factors, such as where assumptions may have been made or facts are incorrect. Where an AI system is used to determine whether an individual may obtain an opportunity or access services, transparency also assists individuals to understand eligibility criteria, how to take steps to meet those criteria and provides a feedback loop for the continuous improvement of AI systems.

Transparency becomes even more critical when there are multiple parties contributing to an AI system across the value chain. For example, entities that broker and process raw data, AI solutions providers and consultations, cloud service providers, platform providers, research institutions and end-user service providers. This requires a clear chain of transparency disclosures as well as a clear delineation of responsibility for each party.

When transparency and disclosures are sufficiently embedded throughout the AI value chain, it enables a transfer of trust from party to party.

---

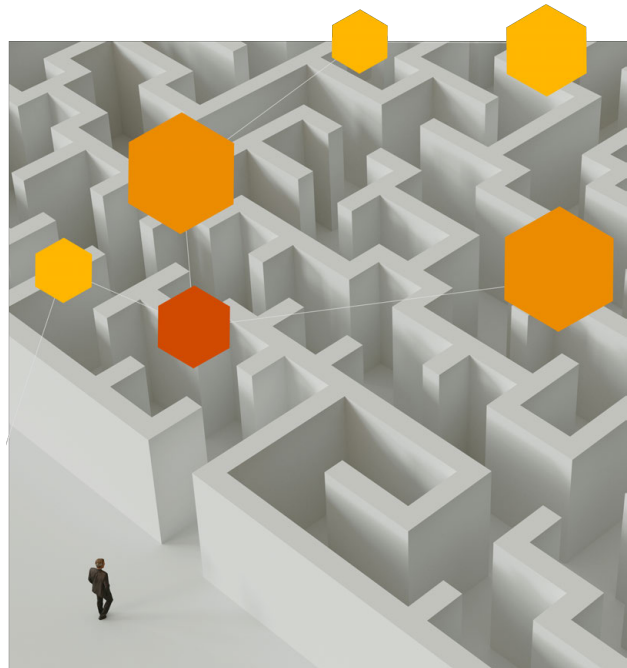
<sup>8</sup> Department of Industry, Science and Resources (2019) Australia's AI Ethics Framework – Australia's AI Ethics Principles

## Strategies for improving transparency in AI

We believe that a combination of the following requirements could help to improve AI transparency in Australia:

- Document and disseminate key design decisions that contextualise outputs. This should include critical information about potential biases, knowledge limits and conditions under which the AI system may fail to behave as intended.
- Provide notice to users at the point at which data is collected for processing, consistent with existing obligations under the Privacy Act and APPs 3 and 5 when collecting personal information. Notice should apply to both data used for AI model training as well as AI inferencing. We acknowledge that this approach may only be practicable when data is collected directly from an individual.<sup>9</sup>
- Provide notice to users at the point at which they begin interacting with AI. This could involve notice statements prior to commencing interactions with AI systems that could reasonably be mistaken as human-generated, such as conversational bots, phone calls and visual simulations.
- Publish AI product transparency reports, which explain system limitations and failure modes, representativeness of training data, trade-offs inherent in model design or which may be configurable, and scenarios where fall-back methods may need to be relied upon.
- Provide post-hoc explanations summarising the most relevant factors that contributed to a given AI system output. These explanations should be human-readable and convey the key variables which determined the outcome of the AI system's analysis. The explanation should also inform the user about avenues that may be available to dispute the result and seek human review of the outcome.

Not all of these requirements should be binding for all AI systems.



## Alignment to risk tiering

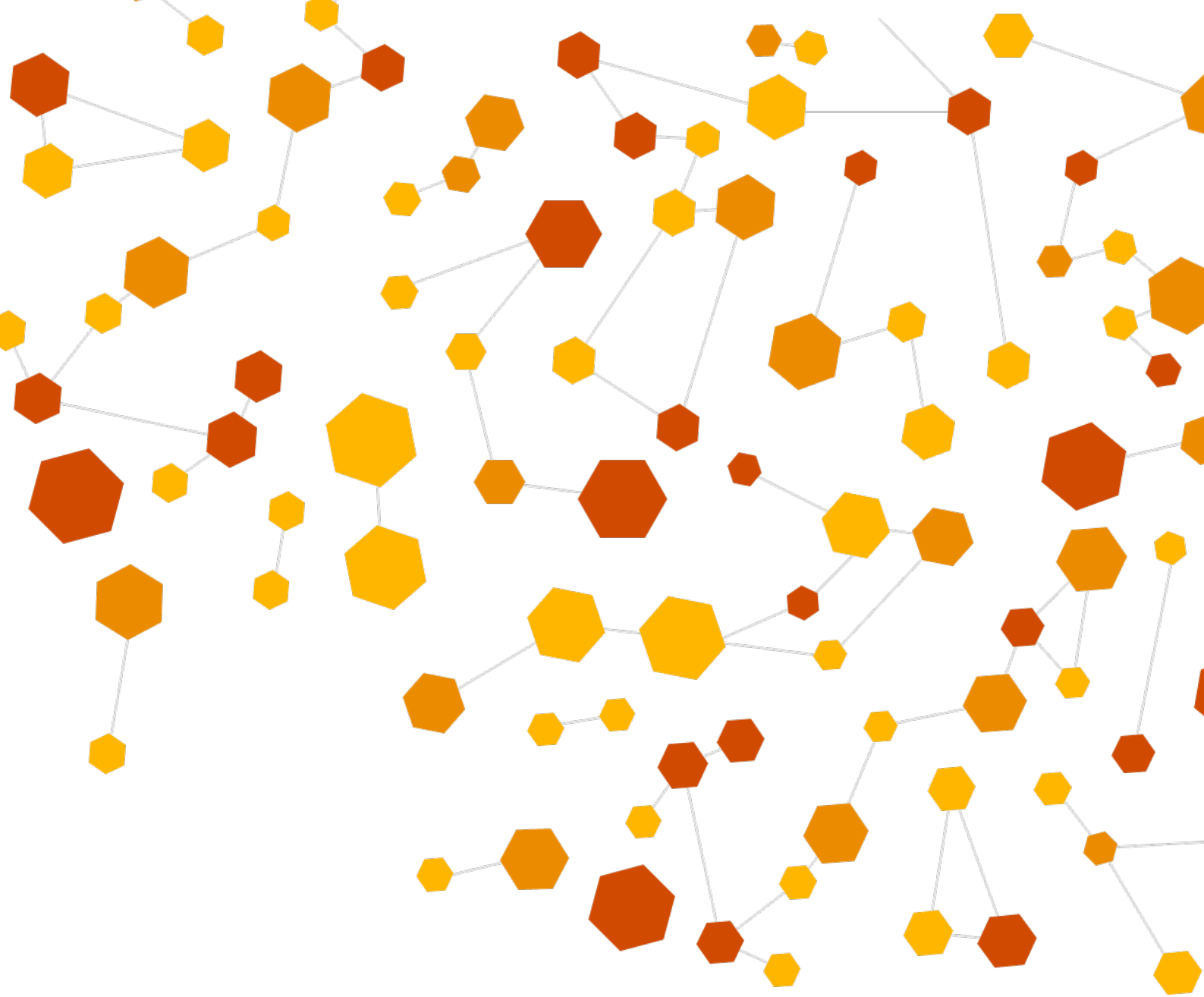
Not all AI systems and use cases have equivalent needs for transparency and explainability, and therefore only some – or none – of the methods outlined to the left might apply to a particular AI system. For example, a model that is used to render system-generated graphics in a computer game won't require the same transparency disclosures as a model that is used to review and approve a person's home loan application.

It is important to note that enforcing strict transparency and explainability requirements for all AI systems would mean that some branches of AI (such as generative AI) could become *de facto* prohibited for all use cases, due to the inability to assess the 'inner alignment' of large language models.<sup>10</sup>

Transparency and explainability are most relevant when AI tools are used in ways which have impacts on individuals' opportunities or access to services, and should be aligned to the risk-based approach described in the 'Defining the scope for AI regulation' section of this response.

<sup>9</sup> Note, that the obligation to collect directly from individuals and provide notice about what their data will be used for, only applies to the extent that it is reasonable to do so. Under the current drafting of the Australian Privacy Principles (APP 3.6) it is not always mandatory for an APP entity (excluding APP agencies) to obtain information directly from an individual, and APP entities may not be required to give notice that they have done so.

<sup>10</sup> Wolf et al (2023) Fundamental limitations of alignment in large language models.



[www.pwc.com.au](http://www.pwc.com.au)

© 2023 PricewaterhouseCoopers. All rights reserved. PwC refers to PricewaterhouseCoopers (Australia Partnership), and may sometimes refer to the PwC network. Each member firm is a separate legal entity. Please see [www.pwc.com/structure](http://www.pwc.com/structure) for further details.

This content is for general information purposes only, and should not be used as a substitute for consultation with professional advisors. Liability limited by a scheme approved under Professional Standards Legislation. At PwC Australia, our purpose is to build trust in society and solve important problems. We're a network of firms in 158 countries with more than 250,000 people who are committed to delivering quality in assurance, tax and advisory services. Find out more and tell us what matters to you by visiting us at [www.pwc.com.au](http://www.pwc.com.au).