Australian Government
Department of Industry, Science and Resources

Safe and responsible AI in Australia

Discussion paper

June 2023

| **consult.**industry.gov.au/supporting-responsible-ai

Contents

1 Introduction

Artificial intelligence (AI) is delivering significant benefits across the economy and society. As an enabling capability, AI is optimising and augmenting many aspects of our lives, including by:

- supporting diagnosis and early detection of health conditions in our hospitals
- expediting travellers at airports through the use of SmartGates
- supporting personalised learning and teaching in remote areas.

AI is unique because it can take actions at a speed and scale that would otherwise be impossible. The speed of innovation in recent AI models are posing new potential risks and creating uncertainty about their full implications, giving rise to public concerns.

While global investment in AI is increasing, adoption rates of AI across Australia remain relatively low[1]. One factor influencing adoption is the low levels of public trust and confidence of Australians in AI technologies and systems.[2]

Building public trust and confidence in the community will involve a consideration of whether further regulatory and governance responses are required to ensure appropriate safeguards are in place. A starting point for considering any response is an understanding of the extent to which our existing regulatory frameworks provide these safeguards. These existing regulations include our consumer, corporate, criminal, online safety, administrative, copyright, intellectual property and privacy laws.

As new technologies or new ways of doing business bring about new potential risks, these regulatory frameworks can and are reviewed and adjusted. For example, the Attorney-General's Department has released a report reviewing the *Privacy Act 1988* (Privacy Act Review) to ensure it is fit for purpose in the digital era.[3]

While Australia already has some safeguards in place for AI and the responses to AI are at an early stage globally, it is not alone in weighing whether further regulatory and governance mechanisms are required to mitigate emerging risks. Our ability to take advantage of AI supplied globally and support the growth of AI in Australia will be impacted by the extent to which Australia's responses are consistent with responses overseas. However, the early responses of other jurisdictions vary.

Some countries like Singapore favour voluntary approaches to promote responsible AI governance. Others like the EU and Canada are pursuing regulatory approaches with proposed new AI laws. The US is consulting on how to ensure AI systems work as claimed, and the UK has released principles for regulators supported by system-wide coordination functions. G7 countries in May 2023 agreed to prioritise collaborations on AI governance, emphasising the importance of forward-looking, risk-based approaches to AI development and deployment.

There are strong foundations for Australia to be a leader in responsible AI. For example, Australia:

- has world-leading research capabilities in AI, and we are early movers in fostering trusted use of digital technologies
- established the world's first eSafety Commissioner to safeguard Australian citizens online and was one of the earliest countries to adopt a national set of AI Ethics Principles
- is a signatory to the OECD's AI Principles, which encourages organisations to reflect ethical practices and good governance when developing and using AI.

The Australian Government has also recently announced further measures in the 2023-24 Budget to support the responsible use of AI, building on these previous initiatives.

---

[1] OECD.AI, 'A sharp increase in AI-related venture capitalist investments could transform global economies and shape the future of artificial intelligence', *OECD.AI website*, 2021, accessed 22 May 2023; Our World in Data, 'Annual global corporate investment in artificial intelligence, by type', *Our World in Data website*, 2022, accessed 22 May 2023; The Productivity Commission, *5-year Productivity Inquiry: Australia's data and digital dividend*, Volume 4 Data and Digital Dividend, p 11, 2023.
[2] N Gillespie, S Lockey, C Curtis, J Pool and A Akbari, *Trust in Artificial Intelligence: A Global Study*, The University of Queensland and KPMG Australia, p 14, 2023.
[3] A final report was made public on 16 February 2023.

This consultation will help ensure Australia continues to support responsible AI practices to increase community trust and confidence. This paper builds on the recent rapid research report on generative AI delivered by the government's National Science and Technology Council (NSTC).

Discussions about governance responses to mitigate risks from fast-paced technologies like AI are often framed around balancing potential risks with fostering innovation and adoption. However, these are not mutually exclusive. Proportionate and timely governance responses, regulatory or otherwise, will build the public trust needed for our economy and society to reap the full benefits of these productivity-enhancing technologies.[4]

1.1 Scope of this paper

This paper seeks advice on steps Australia can take to mitigate the potential risks of AI. Recognising that many related Australian Government initiatives are already underway, we are seeking system-wide feedback on actions that can be taken across the economy on AI regulation and governance. Accordingly, the paper does not provide an in-depth analysis of all the laws applicable to AI. However, it does:

- provide an overview of existing domestic governance and Australia's broader regulatory framework
- provide an overview of recent (and ongoing) international developments
- seek feedback on whether further governance and regulatory responses are needed in Australia.

The focus of this paper is to identify potential gaps in the existing domestic governance landscape and any possible additional AI governance mechanisms to support the development and adoption of AI. Feedback on this paper will inform consideration across government on appropriate responses. This will help support coordinated and coherent responses, recognising that these issues are cross-cutting and related to a broad range of interests.

The paper focuses on governance mechanisms to ensure AI is used safely and responsibly. These mechanisms can include regulations, standards, tools, frameworks, principles and business practices.

This paper does not seek to consider all issues related to AI, for example the implications of AI on the labour market and skills, national security and intellectual property. It also does not consider military specific AI uses. Although AI that may have both military and civilian uses is within scope of the paper. This 'dual-use' of AI will require continued engagement across government.

1.2 Definitions

The paper uses the term 'governance' to include the regulatory and voluntary mechanisms to address potential risks.

In engaging with these issues, some countries use the term 'regulation' to include both:

- voluntary mechanisms to encourage a particular set of behaviours and actions, such as principles, guidelines and voluntary standards
- regulatory mechanisms, which impose formal legal obligations.

There is no single agreed definition of AI. This paper uses the key definitions below, which are based on the International Organisation for Standardization (ISO) definitions.

[4] 'In Australia, trust is a central driver for widespread acceptance of AI'. See The Productivity Commission, *5-year Productivity Inquiry: Advancing Prosperity*, Volume 4, p 83, 2023.

Figure 1: Key definitions used in this paper[5]

Technologies

Artificial intelligence (AI) refers to an engineered system that generates predictive outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation.

Machine learning are the patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes.

Generative AI models generate novel content such as text, images, audio and code in response to prompts.

Applications

A **large language model (LLM)** is a type of generative AI that specialises in the generation of human-like text.

Multimodal Foundation Model (MfM) is a type of generative AI that can process and output multiple data types (e.g. text, images, audio).

Automated Decision Making (ADM) refers to the application of *automated systems* in any part of the decision-making process. Automated decision making includes using automated systems to:

- make the final decision
- make interim assessments or decisions leading up to the final decision
- recommend a decision to a human decision-maker
- guide a human decision-maker through relevant facts, legislation or policy
- automate aspects of the fact-finding process which may influence an interim decision or the final decision.

Automated systems range from traditional non-technological rules-based systems to specialised technological systems which use automated tools to predict and deliberate.

[5] The definitions of 'Artificial Intelligence (AI)', 'machine learning', 'algorithm' are based on the respective ISO definitions (ISO/IEC 22989:2022); The definitions of 'generative AI', 'a large language model (LLM) and 'multimodal foundation model (MfM)' are based on the definitions in Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). *Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs)*. Australian Council of Learned Academies; The definition of 'Automated Decision Making (ADM)' is based on the definition in Commonwealth Ombudsman, *Automated Decision-Making: Better Practice Guide*, Commonwealth Ombudsman, Australian Government, 2020.

For more detailed technical definitions, see the ISO's definition of terms related to AI (ISO/IEC 22989:2022).

Although the focus of the paper is AI, where relevant it draws linkages to related applications such as automated decision-making (ADM). Although ADM may in some instances use AI technologies, in other cases it will not. Even where it does not use AI technologies, risks and challenges associated with ADM may also be mitigated by governance arrangements considered in this paper.

Your input on these definitions will ensure they are appropriate for identifying the types of AI technologies and techniques that may materially impact individuals and societal groups. A broad definition of AI is intended that includes any products or services using AI techniques. These techniques range from simple rules-based algorithms guided by human-defined parameters to more advanced applications like neural networks.

2 Opportunities and challenges

As with all technologies, emerging technologies such as AI bring new opportunities but also new challenges.

2.1 Opportunities

The safe and responsible deployment and adoption of AI presents significant opportunities for Australia to improve economic and social outcomes. AI has been identified as a critical technology in Australia's national interest.[6] In its recent 5-year *Productivity Inquiry* report, the Productivity Commission (PC) identified AI as one of the transformative digital technologies that can help to drive productivity growth in Australia including through the support it provides for the production and adoption of robotics. McKinsey has estimated that automation, including AI, could cumulatively add between $1.1 trillion and $4 trillion to the Australian economy by the early 2030s.[7]

AI technologies are already deployed across our economy and society. Examples include:

- hospitals using AI to consolidate large amounts of patient data and help analyse medical images
- AI tools to help evaluate and optimise engineering designs to improve building safety
- AI-enabling improvements and cost savings in the provision of legal services.[8]

New opportunities will arise as the technology evolves. Given the speed of innovation driven by increasing investment, together with the rapid emergence of open-source systems, many of these opportunities are not yet fully understood. The NSTC's Rapid Response Information Report on generative AI notes that the opportunities presented by large language models (LLMs) and multimodal foundation models (MFMs) are almost impossible to accurately forecast over the next decade.[9]

2.2 Challenges

Despite these benefits, the increased application of AI raises the potential for significant risks.

Like other technologies, AI can be used for positive or harmful purposes. There are many examples and concerns around AI being used for potentially harmful purposes, such as:

- generating deepfakes to influence democratic processes or cause other deceit[10]
- creating misinformation and disinformation[11]
- encouraging people to self-harm.[12]

Inaccuracies from AI models can also create many problems. These include unwanted bias and misleading or entirely erroneous outputs such as 'hallucinations' from generative AI.[13]

[6] Department of Industry, Science and Resources (DISR), 'List of Critical Technologies in the National Interest', *DISR website*, 2023, accessed 22 May 2023.
[7] Taylor et al., *Australia's automation opportunity: Reigniting productivity and inclusive income growth*, McKinsey & Company, p 46, 2019, accessed 22 May 2023.
[8] Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI.* This report defines generative AI as taking its name from its capacity to generate novel content, as varied as text, image, music and computing code, in response to a user prompt.
[9] Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI.*
[10] K Hiebert, 'Democracies are Dangerously Unprepared for Deepfakes', *Centre for International Governance Innovation website*, 27 April 2022, accessed 2 May 2023.
[11] A Satariano and P Mozur, 'The People Onscreen are Fake. The Disinformation is Real.', *New York Times website*, 7 February 2023, accessed 2 May 2023.
[12] J Turc, 'Unconstrained Chatbots Condone Self-Harm', *Medium website*, 8 April 2023, accessed 2 May 2023.
[13] Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI. p 10.*

Algorithmic bias is often raised as one of the biggest risks or dangers of AI. It was a major focus of the Australian Human Rights Commission's *Human Rights and Technology Report* in 2021.[14] Algorithmic bias involves systematic or repeated decisions that privilege one group over another. Examples of discrimination against individuals based on race, sex or other protected categories are well-publicised. These include:

- racial discrimination where AI has been used to predict recidivism which disproportionately targets minority groups[15]
- educational grading algorithms favouring students in higher performing schools[16]
- recruitment algorithms prioritising male over female candidates.[17]

Bias can occur when datasets used to train a model or algorithm are not comprehensive. This can lead to disproportionate impacts on vulnerable groups from AI, including First Nations people, as they are not properly represented in datasets. This may be because the datasets reflect historical biases, or are either too small or only include some relevant data.[18] Bias can also result from how the model is designed, defined and how users interpret its results.

People designing or implementing AI or ADM systems need to be aware of how unwanted bias can be introduced. They need to design, test and validate their systems to correct for bias and potential harms especially where vulnerable groups and individuals are involved. Where AI developers cannot correct for or mitigate unwanted bias, they should either:

- reconsider the appropriateness of deploying the AI system at all
- find alternative data, scale back or revisit their objectives, and then carefully train and test their models again.

Rich, large and quality data sets are a fundamental input to AI. AI systems depend on these training datasets to allow algorithms to be designed, tested and improved. However, access to and application of these datasets have the potential for individuals' data to be used in ways that raise privacy concerns. Privacy protection laws and access to quality data must be carefully balanced to enable fair and accurate results and minimise unwanted bias from AI systems.

AI is unique because it can take actions at a speed and scale that would otherwise be impossible. This speed and scale at which AI can be deployed (to generate benefits as well as cause potential harm) is one of the most significant policy challenges prompting calls for greater regulatory action.

Other risks identified in the NSTC Report concern technical aspects of AI systems.[19] These include system accountability and transparency, and the validity and reliability of data used to train models for their intended purpose**.** For example, the use of large datasets from overseas may fail to capture the location-specific factors required to train AI models to predict bushfires in Australia.

Transparency in AI is also an important challenge across different stages of the AI lifecycle. AI developers and designers can allow validation and demonstrate trustworthiness by being transparent about the acquisition, collection, storage, maintenance and application of data sources. Transparency is important for AI buyers to ensure they are aware of the function of AI in what they are buying, and any flow-on risks or limitations. Transparency can help ensure appropriate accountability, risk

[14] Australian Human Rights Commission (AHRC), 'Human Rights and Technology: Final Report', *AHRC website*, 2021, accessed 5 May 2023.
[15] K Hao, 'AI is sending people to jail - and getting it wrong', *MIT Technology Review website*, 21 January 2019, accessed 14 April 2023.
[16] D Kolkman, 'What the world can learn from the UK's A-level grading fiasco', *LSE Impact Blog*, 26 August 2020, accessed 14 April 2023.
[17] C Hanrahan, 'Job recruitment algorithms can amplify unconscious bias favouring men, new research finds', *ABC News website*, 2 December 2020, accessed 14 April 2023.
[18] Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making*, UK Government, p 26, November 2020, accessed 20 May 2023.
[19] The NSTC's Rapid Research Report referred to three broad risk categories for generative AI: technical system risks, contextual and social risks, and systematic social and economic risks. See Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI. p 9.*

mitigation and responsibility for liability is applied appropriately across AI vendors and buyers along the value chain.

At the end of the value chain, consumers or individuals may not know they are using AI-enabled products or services, or that they have been affected by ADM. Without this knowledge, individuals can't fully appreciate the potential risks or act to protect themselves. In the case of ADM systems, individuals are not prevented from challenging decisions or seeking a review of adverse decisions. However, they are hampered in effectively establishing a case or expressing their view unless they understand how the decision was made and on what basis. The *Australian Community Attitudes to Privacy Survey 2020* prepared for the Office of the Australian Information Commissioner (OAIC) showed:

- 84% of respondents believed people should have a right to know if a decision affecting them is made using artificial intelligence technology[20]
- 78% believed individuals should be told what factors and personal information are considered by the algorithm and how these factors are weighted.[21]

An additional concern raised in the NSTC report is that ownership of large, rich datasets by certain entities or corporations may pose barriers to potential competitors entering or expanding into the market. This can also lead to imbalances between individuals or smaller organisations and the larger or more economically powerful organisations developing and deploying sophisticated AI.

[20] Office of the Australian Information Commissioner (OAIC), *Australian Community Attitudes to Privacy Survey 2020*, OAIC, Australian Government, p 87, 2020, accessed 20 May 2023.
[21] Office of the Australian Information Commissioner (OAIC), *Australian Community Attitudes to Privacy Survey 2020*, OAIC, Australian Government, p 87, 2020, accessed 20 May 2023.

3 Domestic and international landscape

To inform governance mechanisms for safe and responsible AI in Australia, it is useful to consider relevant developments in Australia and internationally. Considering existing domestic mechanisms helps identify any potential gaps, whilst international mechanisms can provide ideas for possible domestic consideration.

3.1 Domestic environment

Navigating the current regulatory landscape

In Australia, the potential risks of AI are currently governed by both general regulations (laws that apply across industries) and sector-specific regulations. These laws are administered by a range of regulators.

The most relevant general regulations include:

- data protection and privacy law
- Australian Consumer Law
- competition law
- copyright law
- corporations law
- online safety
- discrimination law
- administrative law
- criminal law
- the common law of tort and contract.

Examples of sector-specific regulations include those for:

- therapeutic goods
- food
- motor vehicles
- airline safety
- financial services.

These are areas where the government has deemed specific sector-specific laws are necessary. Sector-specific regulations need to be well designed to avoid duplicating economy-wide regulations while filling in any gaps appropriate to AI.

Many of these regulatory regimes, general and sector-specific, can and are being used to address potential harms stemming from AI. As with all emerging risks, regulators will consider how their existing regulatory frameworks may mitigate potential risks. They may issue guidance to clarify the application of laws (or administrative and judicial proceedings may provide greater clarity). Reforms to laws may also be considered to achieve desired policy goals. This process of applying or adjusting existing regulatory frameworks is already underway. For example:

- The *Online Safety Act 2021* includes mechanisms to address online safety issues that may involve AI, from cyberbullying, to image-based abuse (including deepfake pornography) and other kinds of material. The eSafety Commissioner has powers to require the removal of illegal and harmful online content, including child sexual exploitation material and non-consensual intimate images of a person, that extend to AI generated material.
- In 2021 the Therapeutic Goods Administration (TGA) implemented reforms to medical devices regulations and the development of accompanying guidance. These clarified the requirements for software and mobile apps used in medical contexts (known as software as a medical device, or SaMD).

- The determination by the Australian Information Commissioner and Privacy Commissioner that Clearview AI, Inc breached Australian privacy law by scraping individuals' biometric information from the web and disclosing it through a facial recognition tool. Clearview AI were ordered to cease collecting facial images and biometric templates from individuals in Australia, and to destroy existing images and templates collected from Australia.[22]
- The development of new laws to provide the Australian Communications and Media Authority (ACMA) with powers to combat online misinformation and disinformation, announced in January 2023, which could also extend to misinformation and disinformation generated using AI technologies.
- The Privacy Act Review. As part of this review, stakeholders raised concerns about the transparency and integrity of decisions made using ADM (see Box 1).

Box 1: Examples of Privacy Act Review proposals addressing transparency

To promote transparency, the *Privacy Act Review report* recommended that entities include information in their privacy policy about whether personal information will be used in ADM which has a legal, or similarly significant effect on an individual's rights (proposal 19.1).

The report also recommended individuals be given a right to request how these decisions are made (proposal 19.3). This would ensure individuals have sufficient understanding about the rationale for automated decisions so they can exercise other rights. These include their rights under privacy, administrative or anti-discrimination law.

The review also considered the privacy risks associated with using high volumes of data to deliver targeted advertising and content on digital platforms. Stakeholders noted that targeting has the potential to cause significant harm:

- when individuals have limited awareness of why and how they are being targeted and have no control over it
- where targeted content and advertising are used to manipulate, discriminate, exclude and exploit individuals based on their vulnerabilities.

To give individuals greater transparency and control, the report recommended that entities be required to:

- provide information about how they target users (proposal 20.9), including on the algorithms and profiling they use to recommend content to individuals
- let individuals easily opt out of receiving targeted advertising (proposal 20.3).

Consultation on the report closed on 31 March 2023. Feedback will be used to inform the Australian Government's response to the review, which will set out the pathway for reforms.

Additionally, there may be opportunities to consider how some of Australia's general regulations, such as anti-discrimination laws, can be used to avoid issues arising from AI applications.

To help the Australian Government understand any potential regulatory and governance gaps in relation to AI in Australia, we are seeking advice from experts and AI practitioners. Your feedback will be highly valued to help build our understanding of the intersections between AI and laws, and to identity any potential gaps.

Box 2 has an example of the possible application of Australian Consumer Law to AI, including any potential gaps.

[22] This determination was appealed by Clearview AI and is currently going through the Administrative Appeals Tribunal. See Administrative Appeals Tribunal of Australia, 'Clearview AI Inc and Australian Information Commissioner [2023] AATA 1069', *AustLII website*, 8 May 2023, accessed 20 May 2023; Office of the Australian Information Commissioner (OAIC), *Clearview AI breached Australians' privacy*, OAIC, Australian Government, 2021, accessed 20 May 2023.

Box 2: Possible application of Australian Consumer Law (ACL) to AI (general explanation)

The ACL applies to all products or services (except financial products and services) supplied to Australian consumers. This includes products and services incorporating or using AI.

Among other things, the ACL sets out basic rights that consumers can expect when they purchase goods or services. These basic rights are called consumer guarantees. These guarantees include that goods must be of acceptable quality, including by:

- being safe
- lasting
- not having any faults
- looking acceptable
- doing all the things someone would normally expect them to do.

For services, these guarantees include that services must be provided with due care and skill, and that they are fit for any stated purpose.

When businesses supply goods or services that don't meet the consumer guarantees, consumers have the right to a remedy. Remedies include a refund, repair, replacement, or cancellation of a service contract. The remedy consumers are entitled to will depend on whether there has been a minor or major failure to meet the consumer guarantees.

The ACL also includes specific provisions relating to the safety of consumer goods and product-related services. Under these provisions, the relevant Australian Government minister may impose a mandatory safety standard or ban where there is a risk of injury to a person. In addition, suppliers must:

- notify the relevant Australian Government minister of a voluntary recall of consumer goods
- comply with a compulsory recall of consumer goods imposed by a federal, state or territory minister
- provide mandatory reports, subject to exemptions, to the relevant Australian Government minister on any death, serious injury or illness associated with the use or foreseeable misuse of consumer goods or product-related services.

The product safety provisions of the ACL only apply to consumer goods and product-related services, not consumer services more generally. The extent to which the product safety provisions of the ACL apply to consumer-facing uses of AI such as generative AI has not yet been considered by a court.

The Federal Court case *Trivago vs the ACCC* is an example of how the ACL, which was drafted without AI in mind, has been applied to algorithmic decision making.[23] Trivago had used an algorithm to display hotel room recommendations. The algorithm gave consumers the impression they were getting the best deal or cheapest rates, which was not the case.

One challenge for the application of some of Australia's laws is that remedies are often resolved or provided after potential impacts have occurred. While these laws can be an effective deterrent, they can be deficient in certain circumstances. For example, where the impacts from AI are systemic or difficult to reverse. Preventative laws can help to limit problems before they arise. Australia's *Online Safety Act 2021*, for example:

- establishes the Basic Online Safety Expectations, which aim to drive greater transparency around industry's actions to improve online safety

[23] Australian Competition and Consumer Commission (ACCC), 'Trivago to pay $44.7 million in penalties for misleading consumers over hotel room rates', *ACCC website*, 22 April 2022, accessed 5 May 2023.

- provides for the development of new online safety industry codes to address illegal and seriously harmful content online.

These initiatives are flexible and can be applied to potential harms (and solutions) stemming from AI. For example, in supporting prevention these initiatives may require proactive detection and demotion of harmful content in algorithm recommendations.

As an enabling technology, AI is increasingly combined with other components and emerging technologies to produce innovative new businesses, products and services. This often means that AI is regulated under multiple laws, increasing the likelihood of possible duplication or conflict between regulatory systems, and associated compliance burdens on AI developers and adopters.

While the domestic regulatory landscape surrounding AI can seem complex, the range of contexts in which AI can be used, and for different purposes, may necessitate context-specific responses. Rules that are suitable for medical sector device regulation, for example, may not be suitable in the education sector.

This consultation does not seek to consolidate or replicate the development of existing general or sector-specific regulations and governance initiatives across the Australian Government. While this consultation is underway, portfolios will continue to explore and consider AI developments specific to their governance area. For example:

- the Education portfolio will continue working with state and territory counterparts on rules to apply to the use of AI in schools
- the Communications portfolio and the eSafety Commissioner will continue to explore the implications of generative AI in the context of online safety.

The focus of this paper is to identify potential gaps in the existing domestic governance landscape and whether additional AI governance mechanisms are required to support the safe and responsible development and adoption of AI.

Australia's governance responses to date

AI-specific governance responses in Australia to date have largely been voluntary. An example of an important step to help build trust and confidence in the use of AI was the release of Australia's AI Ethics Framework in 2019.

The AI Ethics Framework guides businesses and government to responsibly design, develop and implement AI. It consists of 8 voluntary AI Ethics Principles (see Box 3) to ensure AI is safe, secure and reliable. The principles are consistent with the OECD's Principles on AI.[24] They are intended to be best practice and complement - not replace - existing AI regulations and practices.

[24] More than 40 countries are adherents to the OECD's Recommendation of the Council of Artificial Intelligence, which includes principles for responsible stewardship of trustworthy AI. It was formed using the guidance from over 50 members across government, business, academics, and more. The aim is to have governments, businesses and individuals develop and use AI with people's best interests in mind while ensuring that accountability measures are in place for the proper functioning of AI.

Box 3: Australia's AI Ethics Principles

1. **Human, societal and environmental wellbeing:** AI systems should benefit individuals, society and the environment.
2. **Human-centred values:** AI systems should respect human rights, diversity, and the autonomy of individuals.
3. **Fairness:** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
4. **Privacy protection and security:** AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
5. **Reliability and safety:** AI systems should reliably operate in accordance with their intended purpose.
6. **Transparency and explainability:** There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.
7. **Contestability:** When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
8. **Accountability:** People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Many private and public organisations are already adopting ethical principles or similar practices to ensure appropriate accountability and governance mechanisms are in place for AI. These include:

- major tech firms such as Microsoft, Google, Salesforce and IBM[25]
- public organisations such as the US Department of Defense and the Australian Signals Directorate.[26]

The Australian Government is taking steps to boost its practices to support responsible AI in recognition of public expectations that governments model best practice and lead by example. The Digital Transformation Agency (DTA) has issued guidance on public sector adoption of AI as part of its Australian Government Architecture (AGA).[27] Further, the Office of the Commonwealth Ombudsman's *Automated decision-making better practice guide* provides a practical checklist for Australian Government agencies implementing AI and ADM systems.

At a state level, the NSW Government's AI Assurance Framework came into effect in March 2022. It helps government agencies design, build and use AI-enabled products and solutions appropriately. The framework assists project teams using AI to comprehensively identify, analyse, document and mitigate their AI-specific risks to help establish clear governance and accountability measures.[28]

[25] Microsoft AI, 'Responsible AI', *Microsoft website*, 2023, accessed 12 April 2023; Google AI, 'Responsibility: Our Principles', *Google website*, 2023, accessed 12 April 2023; Salesforce, 'How Salesforce Infuses Ethics into its AI', *Salesforce website*, 5 August 2020, accessed 12 April 2023; IBM, 'AI Ethics', *IBM website*, 2023, accessed 12 April 2023.
[26] C Todd Lopez, 'DOD Adopts 5 Principles of Artificial Intelligence Ethics', *U.S. Department of Defense (US DOD) website*, 25 February 2020, accessed 14 April 2023; Australian Signals Directorate (ASD), 'Ethical AI in ASD', *ASD website*, 2023, accessed 14 April 2023.
[27] Digital Transformation Agency (DTA), 'Australian Government Architecture', *DTA website*, 10 October 2022, accessed 14 April 2023.
[28] NSW Government, *NSW Artificial Intelligence Assurance Framework*, NSW Government, March 2022, accessed 15 April 2023.

Some Australian Government laws that expressly authorise the use of ADM systems include safeguards or procedural requirements to address administrative law and practical risks raised by automated decisions.

The academic community is also researching issues arising from AI and AI-enabled technologies, such as facial recognition, to inform public policy. One recent example is the University of Technology Sydney's Human Technology Institute's report *Facial recognition technology: towards a model law*. This report recommends reforms to modernise Australia's laws, including in relation to privacy and other human rights. It outlines a risk-based legislative approach grounded in international human rights law. The Privacy Act Review report proposes further work to consider the UTS model law and the extent to which it could be accommodated into the Privacy Act framework.[29]

As awareness and attention on responsible AI has grown, government and industry-led initiatives continue to emerge.[30] The National AI Centre, funded by the Australian Government and coordinated by CSIRO, recently established the Responsible AI Network (RAIN). RAIN is a gateway for Australian industries to uplift their practice of responsible AI. It does this by:

- bringing together a national community of practice, guided by world leading expert partners
- enabling Australian businesses with best practice guidance, tools and learning modules.

RAIN is centred around 6 core pillars: law, standards, principles, governance, leadership and technology. There is also a large suite of technical standards and work being progressed by the international standards committee responsible for standardisation in the area of AI, ISO/IEC JTC/1 SC42, and the IEEE.[31] This includes technical standards enabling more transparent, explainable and ethical design of AI systems.[32]

Most recently, the 2023-24 Budget provided funding to extend the National AI Centre and its role in supporting responsible AI usage through developing governance and industry capabilities.

In addition, the Australian Government's new Responsible AI Adopt program will provide $17 million to establish centres to help small to medium enterprises (SMEs) adopt AI technologies responsibly. This will elevate and power their businesses to better compete in international and interstate markets.

Attachment A provides an overview of current Australian Government initiatives relevant to the development, application or deployment of AI.

[29] Attorney-General's Department (AGD), *Privacy Act Review Report 2022*, AGD, Australian Government, p 126, 2023, accessed 19 May 2023.
[30] See the Australian Information Industry Association (AIIA) and KPMG, 'Navigating AI: Analysis and guidance on the use and adoption of AI', *KPMG website*, 28 March 2023, accessed 12 May 2023; Actuaries Institute (AI) and Australian Human Rights Commission (AHRC), 'Guidance Resource: Artificial Intelligence and Discrimination in Insurance Pricing and Underwriting', *AI website,* December 2020, accessed 12 May 2023.
[31] International Organization for Standardization (ISO), 'ISO/IEC JTC 1/SC 42 Artificial Intelligence', *ISO website*, 2017, accessed 2 May 2023; Institute of Electrical and Electronics Engineers Standards Association (IEEE SA), 'About Us', *IEEE SA website*, 2023, accessed 3 May 2023.
[32] For example: ISO/IEC AWI TS 6254 describes approaches and methods that can be used to achieve explainability regarding machine learning models and AI systems; ISO/IEC CD TS 12791 provides mitigation techniques that can be applied throughout the AI system life cycle to treat unwanted bias; IEEE P7000-2021 enables organisations to design systems with explicit consideration of individual and societal ethical values, such as transparency, sustainability, privacy, fairness and accountability; IEEE 7001-2021 sets out measurable, testable levels of transparency for autonomous systems.

3.2 International developments

While the regulation of AI remains in an early state globally, there is a developing international direction towards a risk-based approach for governance of AI. The most advanced developments are in the European Union and the United States, while Canada and New Zealand have implemented requirements for government.

The NSTC report highlights the different approaches countries are taking to AI governance and demonstrates the breadth of work underway within this space. Many countries are grappling with similar issues and are developing diverse approaches that range from voluntary to regulatory. The countries discussed in this paper include those referred to in the NSTC report.

Australia continues to engage in bilateral, regional and multilateral discussions with other jurisdictions. In addition, significant multilateral work on AI is being undertaken, including by the OECD, United Nations, World Trade Organisation and the World Economic Forum (WEF).[33] This work is not discussed in detail in the paper but, as it develops, will likely inform national responses.

European Union

The General Data Protection Regulation (GDPR) came into effect in 2018. It regulates the use of personal data in ADM systems 'which produce legal or similarly significantly effects'.[34] It requires that individuals be given:

- prior notice of the use of personal data in ADM, including profiling[35]
- a right to access information about the existence of ADM and 'meaningful information about the logic involved, as well as the significance and the envisaged consequences' of such processing to the individual[36]
- the 'right not to be subject' to certain forms of ADM.[37]

GDPR also requires controllers (i.e. those who determine the purposes and means of processing personal data) to implement measures to:

- enable individuals to obtain human intervention on the part of the controller
- express their point of view
- contest the decision.[38]

This general right to not be subject to ADM does not apply to automated decisions that are contractually necessary, authorised by an EU or member state law, or based on the subject's explicit consent.[39]

In September 2022, the European Commission proposed adapting existing civil liability rules concerning AI (the AI Liability Directive) to alleviate the burden of proof for victims of AI-enabled products or services in liability claims. The aim of the AI Liability Directive is to ensure victims of

[33] The European Commission, for example, is funding the OECD under its Committee on Consumer Policy to undertake a 2-year project on the consumer safety of new technology (including the mental health impacts of AI). The OECD under its *Working Party on Artificial Intelligence Governance Policy* is undertaking a study on mental health in the digital age and developing a framework for monitoring and reporting on AI incidents, along with a range of other relevant work.
[34] European Parliament and Council of the European Union, *General Data Protection Regulation (GDPR) Art 22 - Automated individual decision-making, including profiling*, European Parliament, n.d., accessed 18 May 2023.
[35] European Parliament and Council of the European Union, *GDPR Art 13(2)(f) - Information to be provided where personal data are collected from the data subject*, European Parliament, n.d., accessed 18 May 2023; European Parliament and Council of the European Union, *GDPR Art 14(2)(g) - Information to be provided where personal data have not been obtained from the data subject*, European Parliament, n.d., accessed 18 May 2023.
[36] European Parliament and Council of the European Union, *GDPR Art 15(1)(h) - Right of access by the data subject*, European Parliament, n.d., accessed 18 May 2023.
[37] European Parliament and Council of the European Union, *GDPR Art 22(1) - Automated individual decision-making*.
[38] European Parliament and Council of the European Union, *GDPR Art 22 - Automated individual decision-making*; European Commission (EC), 'Proposal for a Regulation laying down harmonised rules on artificial intelligence', *EC website*, 21 April 2021, accessed 19 May 2023.
[39] European Parliament and Council of the European Union, *GDPR Art 22 - Automated individual decision-making*.

AI-enabled products and services are equally protected as victims of traditional technologies. The directive also aims to:

- reduce legal uncertainty regarding the liability exposure of businesses developing or using AI
- harmonise the national civil liability rules that apply to the development and use of AI across the EU.[40]

The EU *Digital Services Act* (DSA) came into effect in November 2022 and will be wholly applicable in February 2024.[41] The DSA applies to all digital services that connect consumers to goods, services, or content. The Act:

- creates new obligations for online platforms to reduce harms and counter risks online, including how they design services and procedures
- introduces protections for users' rights online
- places digital platforms under a new transparency and accountability framework, including requirements to:
    - provide regulators and researchers access to data, including algorithms
    - publish transparency reports on content moderations decisions and algorithms used for recommendations.

The European Commission is setting up a European Centre for Algorithmic Transparency (ECAT) to support supervision and monitoring of the DSA.[42] The ECAT will provide support with assessments as to whether the functioning of very large online platforms and search engines are in line with the risk management obligations of the DSA. This will ensure a safe, predictable and trusted online environment.

The proposed EU *AI Act* adopts a risk-based approach to the regulation of AI, with differing regulatory requirements for minimal, limited, high and unacceptable risk (see Attachment B for further detail). Minimal risk AI is permitted with no restrictions, while unacceptable risk AI is banned. The European Parliament is scheduled to vote on the proposed Act in the first half of 2023 and the final Act is expected to be adopted by the end of 2023.[43] The EU AI Act will become law after the Council of the European Union (members States), the European Parliament and the European Commission agree on a common version of the text.

European data regulators are increasingly focused on the specific impacts of generative AI. On 13 April 2023, the European Data Protection Board launched a task force to look at privacy concerns related to ChatGPT.

United States of America

The White House Office of Science and Technology Policy released the *Blueprint for an AI Bill of Rights* in June 2022.[44] The non-binding blueprint sets out 5 principles and associated practices to guide the design, use, and deployment of automated systems to protect the rights of the American public.[45] These principles are supported by a technical companion that provides guidance on how to put the principles into practice.

Prior to this, in 2020, the White House issued a Guidance for Regulation of Artificial Intelligence Applications. The guidance establishes a framework for federal agencies to assess potential regulatory and non-regulatory approaches to AI issues. It included principles guiding US agencies on

[40] European Commission (EC), *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*, EC, 2022, accessed 20 April 2023.
[41] European Commission (EC), *Digital Services Act: EU's landmark rules for online platforms enter into force*, EC, 16 November 2022, accessed 20 April 2023.
[42] European Commission (EC), 'European Centre for Algorithmic Transparency', *EC website,* n.d., accessed 20 April 2023.
[43] K O'Connell, C Lim, L Pallaras, D MacRae and B Kidman, 'Developments in the regulation of Artificial Intelligence', *King and Wood Mallesons (KWM) website*, 19 April 2023, accessed 24 April 2023.
[44] U.S. Government, 'Blueprint for an AI Bill of Rights - Making Automated Systems Work for the American People', *U.S. White House website*, n.d., accessed 17 April 2023.
[45] The five principles consist of: (i) Safe and Effective Systems; (ii) Algorithmic Discrimination Protections; (iii) Data Privacy; (iv) Notice and Explanation; and (v) Human Alternatives, consideration and fallback.

whether and how they could regulate AI. Several US agencies have since produced reports on the regulation of AI in their respective sectors.[46]

In 2020 the US released an *Executive Order on AI*. It guides federal agencies to design, develop, acquire and use AI in a way that fosters public trust and confidence while protecting privacy, civil rights, civil liberties and American values.

The US Federal Trade Commission has also released a statement that it would take enforcement action against biased AI systems under section 5 of the Federal Trade Commission Act.[47] The US government Accountability Office also issued a report on key practices to ensure responsible use of AI by federal agencies.[48]

In January 2023, the US Chamber of Commerce's Commission on AI called for the regulation of AI as it found that a failure to do so would harm the economy and constrain the development and introduction of beneficial technologies. In the same month, the US National Institute of Standards and Technology (NIST) released a voluntary *AI Risk Management Framework (AIRMF)*. The framework can be used by organisations to address risks in the design, development, use and evaluation of AI products, services and systems. Although the AIRMF does not explicitly rely on risk categories, it requires businesses to weigh up positive and negative net risks of AI adoption.[49]

In April 2023, the National Telecommunications and Information Administration (NTIA), which advises the President on technology regulation, issued a request for public comment to support its AI-related work.[50] The public comments received will support the development of policies on AI audits, assessments, certifications and other mechanisms that aim to build trust in AI systems. It focuses on four key areas:

1. What kinds of trust and safety testing should AI development companies and their enterprise clients conduct?
2. What kinds of data access are necessary to conduct audits and assessments?
3. How can regulators and other actors incentivise and support credible assurance of AI systems along with other forms of accountability?
4. What different approaches might be needed in different industry sectors - like employment or health care?

In the same month, US Senate Majority Leader Chuck Schumer launched a proposed regulatory framework to deliver transparent, responsible AI while not stifling critical and cutting-edge innovation.[51] The proposed framework requires companies to allow independent experts to review and test their AI technologies ahead of public release. They must also make the test results accessible to users.

At a state level, Alabama, Colorado, Illinois and Mississippi have passed bills that limit the use of AI in their states.[52] For example, state and local agencies in Colorado that use or intend to use a facial

[46] K O'Connell et al., 'Developments in the regulation of Artificial Intelligence', *KWM website*, 19 April 2023, accessed 24 April 2023. Refer to the section on United States regulation of AI in their sectors: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence, Artificial Intelligence and Machine Learning in Software as a Medical Device Action Plan, Using Artificial Intelligence and Algorithms and Trustworthy AI Playbook.
[47] U.S. Federal Trade Commission, *Federal Trade Commission Act: Incorporating U.S. SAFE WEB Act amendments of 2006 (Unofficial version)*, U.S. Government, n.d., accessed 22 April 2023.
[48] U.S. Government Accountability Office (GAO), 'Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities', *U.S. GAO website*, 30 June 2021, accessed 22 April 2023.
[49] National Institute of Standards and Technology (NIST), *AI Risk Management Framework: Second Draft*, U.S. Department of Commerce, 18 August 2022, accessed 24 April 2023.
[50] D Shepardson and D Bartz, 'US begins study of possible rules to regulate AI like ChatGPT', *Reuters website*, 12 April 2023, accessed 26 April 2023; National Telecommunications and Information Administration (NTIA), *NTIA Seeks Public Input to Boost AI Accountability*, U.S. Department of Commerce, 11 April 2023, accessed 26 April 2023.
[51] Senate Democrats, 'Schumer Launches Major Effort to Get Ahead of Artificial Intelligence', *Senate Democrats website*, 13 April 2023, accessed 24 April 2023.
[52] C Kraczon, 'The State of State AI Policy (2021-22 Legislative Session)', *Electronic Privacy Information Center (EPIC) website,* 8 August 2022, accessed 17 April 2023. Refer to Illinois: Artificial Intelligence Video Interview Act which requires employers to notify applicants in videotaped interviews that AI is being used to analyse the interview and consider the applicant's fitness for the position.

recognition service (FRS) are required to file a notice of intent and produce an accountability report. Agencies using FRS are required to:

- subject decisions that produce legal effects to meaningful human review
- conduct periodic training of individuals who operate the FRS
- maintain records sufficient to facilitate public reporting and auditing of compliance with FRS policies.[53]

At the local level, New York City has been consulting on regulations and revisions to these regulations since 2022 to restrict the use of automated employment decision tools by employers and employment agencies. The main requirements are that:

- these tools be subjected to bias audits annually
- the results of the bias audit are published
- employers and employment agencies notify employees and job candidates that such tools are being used to evaluate them.[54]

Washington, DC is also considering a bill that would prohibit the use of algorithmic decision-making in a discriminatory manner that limit 'important life opportunities'. It requires for example:

- notices to individuals whose personal information is used in certain algorithms
- requirements for audit of algorithms for discriminatory impacts
- reporting this information to the Washington Attorney General's office in contexts including employment, housing, healthcare and financial lending.[55]

The California Privacy Rights Act will allow regulations to be developed to grant access and opt-out rights for ADM technology. It will require businesses' responses to access requests to include meaningful information about the logic involved in such decision-making processes.[56]

On product safety, the US Consumer Product Safety Commission has released reports on the use of AI and machine learning technologies in consumer products, and the assessment, testing and evaluation of hazards associated with AI and machine learning in consumer products.[57]

United Kingdom

In 2021, the UK Government published the National AI Strategy, 'a 10-year plan to make Britain a global AI superpower'.[58] The strategy sets out the UK's long-term actions regarding the governance of AI in addition to broader economic actions regarding the AI industry. The strategy proposed the introduction of cross-sector AI-specific principles to enable more consistency across the various sector-specific regulatory regimes.[59]

The UK has also developed the Algorithmic Transparency Standard, a recording standard that helps public sector bodies provide clear information about the algorithmic tools they use and why they're using them.[60] The standard is one of the world's first policies for transparency on the use of

[53] Colorado General Assembly, 'Artificial Intelligence Facial Recognition', *Electronic Privacy Information Center website*, 2022, accessed 17 April 2023.
[54] The Mayor's Office of Operations, 'Automated Employment Decision Tools (Updated)', *NYC Rules website*, n.d., accessed 26 May 2023.
[55] Government of the District of Columbia, *Bill 24-558 - Stop Discrimination by Algorithms Act of 2021*, U.S. Government, 2021, accessed 17 April 2023; D Castro, 'DC's Proposed "Stop Discrimination by Algorithms Act" Would Discriminate Against Algorithms', *Center for Data Innovation website*, 22 September 2022, accessed 17 April 2023.
[56] California Legislative Information, *California Consumer Privacy Act of 2018 - 1798.186(16)*, U.S. Government, n.d., accessed 18 April 2023.
[57] United States of America Consumer Product Safety Commission, *Artificial Intellgenice and Mcahine Learning in Consumer Products*, U.S. Government, 19 May 2021, accessed 19 April 2023; United States of America Consumer Product Safety Commission, *Applied Artificial Intelligence and Machine Learning Test and Evaluation Program for Consumer Products*, U.S. Government, 24 August 2022, accessed 19 April 2023; U.S. Consumer Product Safety Commission, 'Potential Hazard Associated with Emerging and Future Technologies', *U.S. Consumer Product Safety Commission website*, 19 January 2017, accessed 15 May 2023.
[58] HM Government, *National AI Strategy*, UK Government, September 2021, accessed 19 April 2023.
[59] HM Government, *National AI Strategy*, UK Government, p 54, September 2021, accessed 19 April 2023.
[60] Central Digital and Data Office and Centre for Data Ethics and Innovation, 'Collection - Algorithmic Transparency Reports', *Gov.UK website*, 13 January 2023, accessed 17 April 2023.

algorithmic tools in government decision making. It comprises an algorithmic transparency data standard and an algorithmic transparency template and guidance that helps public sector organisations provide information to the data standard.

The standard is part of the UK Government National Data Strategy. The strategy has a commitment to explore an appropriate and effective way to deliver greater transparency on algorithm-assisted decision making in the public sector. The National AI Strategy reiterated this commitment, with an action to conduct research that will help develop a cross-government standard for algorithmic transparency.

Some of the governance initiatives proposed in the strategy have since been completed. The Alan Turing Institute piloted the AI Standards Hub in 2022 to provide organisations with access to educational materials and training on global AI standards and best practice.[61] The Information Commissioner's Office developed the AI and data protection risk toolkit in 2022 to provide practical support to organisations assessing the risks to individual rights and freedoms caused by their own AI systems.[62] The UK Department for Education released a policy paper on the use of generative AI in the education sector in March 2023.[63]

The UK Government published a policy white paper in March 2023 titled *A pro-innovation approach to AI regulation* based on the proposals in the 2021 National AI Strategy. The white paper sets out a framework for responsible development and use of AI in all sectors of the UK economy. The proposed framework is underpinned by 5 principles:

- safety, security and robustness
- appropriate transparency and explainability
- fairness
- accountability and governance
- contestability and redress.[64]

The principles are proposed to be issued on a non-statutory basis and implemented by existing regulators. Following an initial period of implementation, it is proposed that the principles will be legislated to create a statutory duty for regulators to have due regard to the principles. The framework also includes a central coordination function to ensure a coherent regulatory approach.

Canada

The Canadian Directive on Automated Decision Making applies to most of Canada's federal government institutions.[65] It takes a principles-based approach to classifying AI into risk categories. The Canadian Directive uses the following classifications:

- **Low ('level I') risk:** impacts that are reversible or brief
- **Moderate ('level II') risk:** impacts that are likely reversible and short-term
- **High ('level III' risk):** impacts that can be difficult to reverse and ongoing
- **Very high ('level IV' risk):** impacts that are irreversible and perpetual.

The directive requires Canadian Government agencies to classify new systems into 1 of 4 risk categories. Graduated requirements require more intensive algorithmic impact assessments, transparency, quality assurance, recourse and reporting requirements for higher risk systems.

On 16 June 2022, the Canadian Government introduced the *Digital Charter Implementation Act 2022*, a package of laws that will:

[61] AI Standards Hub, 'Training', *AI Standards Hub website*, n.d., accessed 24 April 2023.
[62] Information Commissioner's Office (ICO), 'Our work on Artificial Intelligence', *ICO website*, n.d., accessed 24 April 2023.
[63] Department of Education, 'Policy Paper - Generative artificial intelligence in education', *Gov.UK website*, 29 March 2023, accessed 24 April 2023.
[64] Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, UK Government, March 2023, accessed 20 May 2023.
[65] Government of Canada, 'Directive on Automated Decision-Making', *Government of Canada website*, 25 March 2023, accessed 17 April 2023.

- implement Canada's first AI legislation, the *Artificial Intelligence and Data Act (AIDA*);
- reform Canadian privacy law, replacing the *Personal Information Protection and Electronic Documents Act* with *the Consumer Privacy Protection Act*
- establish a tribunal specific to privacy and data protection.[66]

The AIDA establishes Canada-wide requirements for the design, development, use and provision of AI systems. It prohibits certain conduct in relation to these systems that may result in serious harms or biased outputs.

The Canadian Government is currently considering a suite of regulations on data protection and artificial intelligence. Bill C-27 (and its predecessor Bill C-11) proposes the enactment of the *Consumer Privacy Protection Act*, *Personal Information and Data Protection Tribunal Act*, and the *Artificial Intelligence and Data Act*. The Bill proposes that organisations provide a general account of their use of any automated decision system to make predictions, recommendations or decisions that could have significant impacts.[67] The Bill also proposes that organisations that have used an automated decision system provide an explanation of how the prediction or decision was obtained when requested by the individual affected.[68] The *Artificial Intelligence and Data Act* section of the Bill sets out a risk-based approach to regulating AI systems.[69] Bill C-27 passed the second reading in the House of Commons on 24 April 2023 and will next be considered by the Standing Committee on Industry and Technology.[70]

China

China has numerous laws regarding AI and automated decision-making. The *Personal Information Protection Law* (2021) includes provision on the governance of automated decision-making. The Internet Information Service Algorithmic Recommendation Management Provisions (2022) govern the provision of AI-based personalised recommendation services to users. China subsequently developed a mandatory registration system for recommendation algorithms (the Internet Information Service Algorithm Filing System) to specify what datasets and types of data were used to train the model.[71]

In January 2023, the People's Republic of China's Regulations on the Administration of Deep Synthesis of Internet-Based Information Services (the 'deep synthesis laws') entered into force. These rules govern how companies develop deep synthesis technology such as deep fakes and other AI-generated media.[72]

More recently in April 2023, the Cyberspace Administration of China issued draft rules for public comment to manage how companies develop generative AI products like ChatGPT.[73] These rules are expected to take effect sometime before the end of 2023 and appear to apply more broadly beyond algorithms covered by the 'deep synthesis laws' to include 'models and rules' used to generate content.[74] The draft rules are reported to require service providers to ensure generated content reflect the 'core value of socialism', 'respect social morality and public order', and do not attempt to 'subvert state power' or 'undermine national unity' or produce content that is pornographic, or encourages

[66] M Medeiros and J Beatson, 'Bill C-27: Canada's first artificial intelligence legislation has arrived', *Norton Rose Fulbright website*, 23 June 2022, accessed 17 April 2023.
[67] Parliament of Canada, *Bill C-27 s62(2)(c),* Government of Canada, 22 November 2021, accessed 17 April 2023.
[68] Parliament of Canada, *Bill C-27 s63(3),* Government of Canada, 22 November 2021, accessed 17 April 2023.
[69] Parliament of Canada, *Bill C-27 Part 3,* Government of Canada, 22 November 2021, accessed 17 April 2023.
[70] Parliament of Canada, *Bill C-27,* Government of Canada, 22 November 2021, accessed 17 April 2023.
[71] M Sheehan and S Du, 'What China's Algorithm Registry Reveals about AI Governance', *Carnegie Endowment for International Peace website*, 9 December 2022, accessed 26 April 2023.
[72] J Finlayson-Brown and S Ng, 'China brings into force Regulations on the Administration of Deep Synthesis of Internet Technology', *Allen & Overy website*, 1 February 2023, accessed 19 April 2023.
[73] Rajah & Tann Asia, 'China Issues Draft Administrative Measures for Generative Artificial Intelligence Services', *Lexology website*, 19 May 2023, accessed 19 April 2023.
[74] Davis Wright Tremaine LLP, 'China's Cyberspace Administration Proposes Draft Rules to Regulate Generative AI', *Lexology website*, 11 April 2023, accessed 19 April 2023.

violence, extremism, terrorism or discrimination.[75] In addition, new generative AI products will need to go through a 'security review' before release, and verify users' identities and tracking usage.[76]

China has also introduced laws regulating the use of algorithmic technologies that create a range of obligations for digital service providers. These include requiring that details about significant recommendation algorithms are registered with the Chinese government.[77]

New Zealand

New Zealand has implemented an Algorithm Charter, which classifies algorithms into 3 risk levels.[78] The charter must be applied to algorithms deployed by the New Zealand Government and requirements apply for:

- transparency
- consultation
- data quality
- privacy
- ethics
- human rights
- human oversight.

Singapore

Singapore's Personal Data Protection Commission (PDPC) first developed the Model Artificial Intelligence Governance Framework in 2019 to provide private sector organisations with guidance on how to address key ethical and governance issues when deploying AI solutions. The Model Framework aims to promote public understanding and trust in AI technologies through the practice of good data accountability practices, and transparent communication.[79] The second edition of the Model was released in 2020 to include industry examples of how organisations have implemented AI governance practices.[80]

In 2021, the Monetary Authority of Singapore (MAS) and the National AI Office (NAIO) at the Smart Nation and Digital Government Office (SNDGO) launched the National Artificial Intelligence (AI) Programme in Finance - a sector-specific initiative focusing on developing the capabilities of financial institutions. One of the objectives of the Programme is to improve societal acceptance of AI through sound AI governance. For example, the 'Veritas' initiative within the Programme helps financial institutions utilise AI and data analytics responsibly based on fairness, ethics, accountability, and transparency (FEAT) principles.[81]

On 25 May 2022, Singapore's Information Media Development Authority (IMDA) and the PDPC launched standardised self-testing tools ('AI Verify') to enable businesses to check the implementation of AI models against a set of principles.[82] Ten companies from different sectors and of various sizes have already

[75] E Hale, 'China races to regulate AI after playing catchup to ChatGPT', *AI Jazeera website*, 13 April 2023, accessed 18 April 2023; Rajah & Tann Asia, 'China Issues Draft Administrative Measures for Generative Artificial Intelligence Services', *Lexology website*, 19 May 2023, accessed 19 April 2023.
[76] M Deutscher, 'Regulators in US and China request public comment on AI rules', *SiliconANGLE website*, 11 April 2023, accessed 18 April 2023; A Kharpal, 'China releases rules for generative AI like ChatGPT after Alibaba, Baidu launch services', *CNBC website*, 11 April 2023, accessed 26 April 2023.
[77] M Sheehan and S Du, 'What China's Algorithm Registry Reveals about AI Governance', *Carnegie Endowment for International Peace website*, 9 December 2022, accessed 26 April 2023.
[78] StatsNZ and New Zealand Government, 'Algorithm Charter for Aotearoa New Zealand', *data.govt.nz website*, July 2020, accessed 14 April 2023.
[79] Personal Data Protection Commission (PDPC) Singapore, *Model Artificial Intelligence Governance Framework: Second Edition*, Singapore Government, 2020, accessed 16 May 2023.
[80] PDPC Singapore, *Model Artificial Intelligence Governance Framework: Second Edition*, p 4.
[81] Monetary Authority Singapore (MAS) and Smart Nation & Digital Government Office, 'National programme to deepen AI capabilities in financial services', *MAS website*, 8 November 2021, accessed 16 May 2023.
[82] Personal Data Protection Commission (PDPC), 'Launch of AI Verify - An AI Governance Testing Framework and Toolkit', *PDPC website*, 25 May 2022, accessed 16 May 2023.

tested or provided feedback on AI Verify, and there is currently an open invitation for other companies across industry to pilot the tools.[83]

Thailand

Thailand's *National AI Ethics Guideline* was approved by the Thai Cabinet in 2021. The Guideline was developed by the Digital Economy and Society (DES) Ministry to ensure the development and use of AI technology in Thailand aligns with economic goals and is compliant with law and international standards.[84] The Guideline establishes principles and expectations for different actors (regulators, developers, manufacturers, end users) and provides a basis for procurement-based risk management.[85]

The Thai Cabinet approved the Thailand National AI Strategy and Action Plan (2022-2027) in 2022. One of the key pillars of the Strategy is "preparing Thailand's readiness in social, ethics, law, and regulation for AI application". The Strategy explicitly declares the Thai government's expectation that "at least 600,000 Thai population have awareness of AI law and ethics" and "an AI Law & Regulation is enforced" by 2027.[86]

Italy

In March 2023, the Italian Data Protection Authority (Garante) announced a temporary conditional ban of ChatGPT, raising concerns about private data that had been gathered to 'train' the product.

OpenAI announced on 28 April 2023 that ChatGPT had been reinstated in Italy after it implemented changes to comply with Garante's data privacy conditions, including:

- increased transparency on OpenAI's website about how ChatGPT processes user data
- opt-out rights, including the option to disallow user conversations from being used as training data
- age verification to protect children under 13 in Italy from accessing ChatGPT
- a notice that makes users aware that ChatGPT could produce inaccurate information about 'people, places or facts'.[87]

Indonesia

Indonesia requires all tech companies to apply for licences to operate in Indonesia under Regulation No. 5 of 2020 on Private Electronic System Operators (MR5).[88] All private digital services and platforms are required to register with the Ministry of Communication and Information Technology to avoid being blocked by internet service providers.[89] The regulation requires tech companies to comply with government requests to access user data and to almost immediately take down online content that is 'unlawful' or may 'disturb public order'.[90]

---

[83] Infocomm Media Development Authority (IMDA), 'Singapore launches world's first AI testing framework and toolkit to promote transparency; Invites companies to pilot and contribute to international standards development', *IMDA website*, 25 May 2022, accessed 16 May 2023; Personal Data Protection Commission (PDPC) Singapore, 'Singapore's approach to AI Governance', *PDPC website,* 2022, accessed 16 May 2023.

[84] A Sharon, 'Thailand Draft Ethics Guidelines for AI', *OpenGov website*, 4 November 2019, accessed 15 May 2023; OneTrust DataGuidance, 'Thailand: MDES releases AI ethics guidelines', *OTDG website*, 28 October 2019, accessed 16 May 2023; OECD.AI Policy Observatory, 'Thailand National AI Strategy and Action Plan', *OECD website*, 2022, accessed 16 May 2023.

[85] Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI. p 27.*

[86] National Electronics and Computer Technology Center (NECTEC), 'The Cabinet approved the (Draft) Thailand National AI Strategy and Action Plan (2022 - 2027)', *NECTEC website*, 30 July 2022, accessed 16 May 2023.

[87] Garante Per La Protezione Dei Dati Personali (GPDP), 'ChatGPT: OpenAI reinstates service in Italy with enhanced transparency and rights for european users and non-users', *GPDP website*, 28 April 2023, accessed 3 Mary 2023; K Chan, 'OpenAI: ChatGPT back in Italy after meeting watchdog demands', *Associated Press website*, 29 April 2023, accessed 3 May 2023; Deutsche Welle (DW), 'Italy lifts ban on ChatGPT after data privacy improvements', *DW website*, 29 April 2023, accessed 3 May 2023.

[88] C Guntur Lebang and G Priyandita, 'Indonesia's controversial tech licensing scheme', *Australian Strategic Policy Institute (ASPI) website*, 9 August 2022, accessed 26 April 2023.

[89] Guntur Lebang et al,, 'Indonesia's controversial tech licensing scheme'.

[90] S Strangio, 'Indonesia Prepping Strict New Rules for Online Platforms: Report', *The Diplomat website*, 24 March 2022, accessed 26 April 2023; F Potkin and S Sulaiman, 'Indonesia preparing tough new curbs for online platforms', *Reuters website*, 23 March 2022, accessed 26 April 2023.

In addition, Indonesia requires all businesses that operate in Indonesia to be registered under a risk-based licensing system. The risk classification of a business correlates with the level of regulatory requirements it is required to meet.[91] Generally, lower risk business activities have less regulatory requirements.

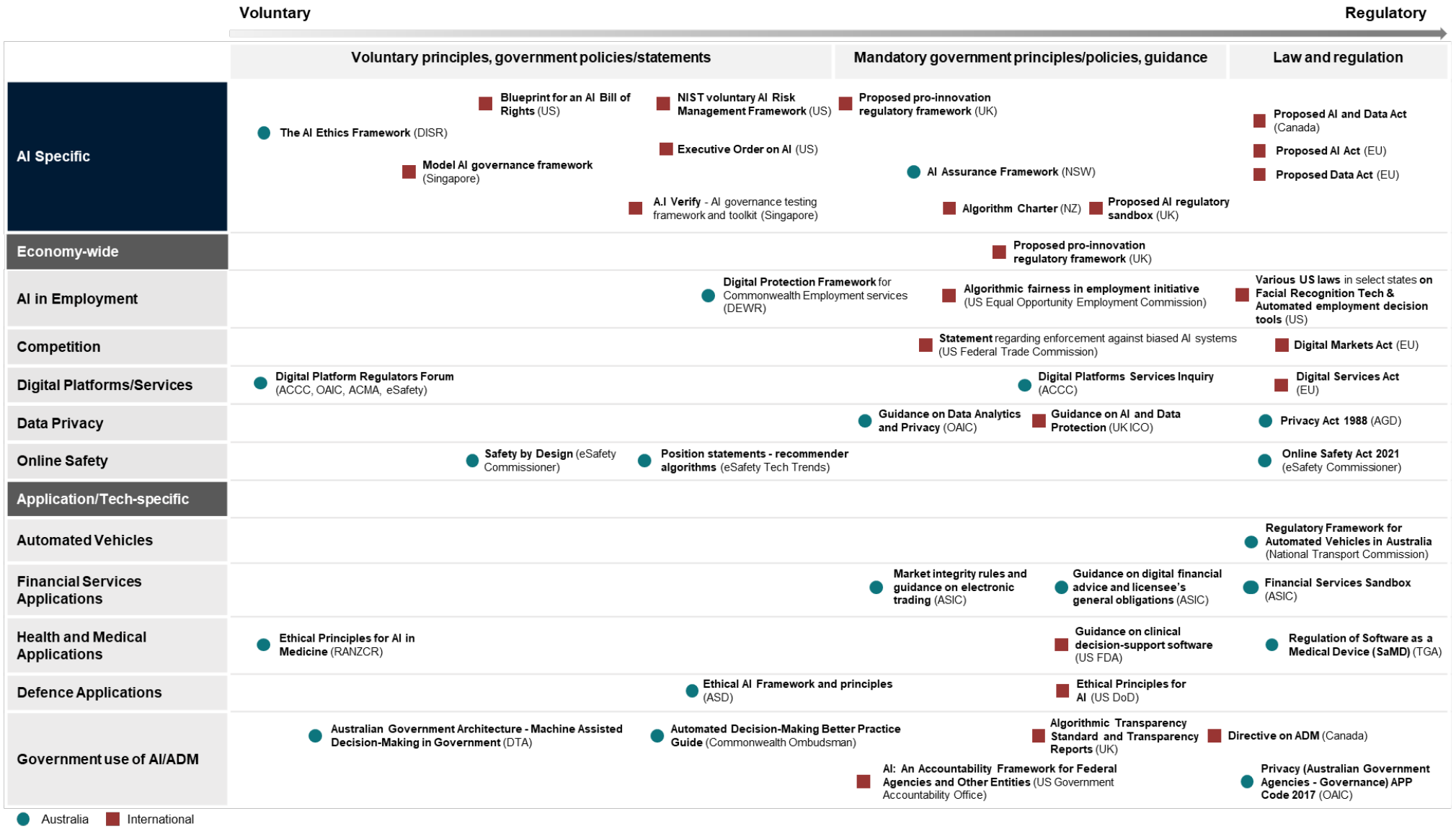
3.2.1 Mapping of the domestic and international environment

Figure 2 illustrates the breadth and clustering of international and domestic governance-related initiatives affecting AI. It is not an exhaustive list and not all of the initiatives listed are discussed in the paper. Readers are encouraged to draw their own observations from this diagram, but a few observations are as follows:

- **Complexity:** There is a complex tapestry of governance actions being taken. These range from regulating the technology or its application in specific fields or in generic regulatory regimes (such as competition and data privacy).
- **Competition issues:** There is increasing attention by governments on potential anti-competitive conduct by large digital platforms, many of which use some types of AI as part of their services.
- **Government use of AI:** There is also a growing focus on initiatives to ensure accountability, transparency and minimum standards when governments use AI.
- **Guidance is common:** Regulators in higher-risk areas such as privacy, medical devices and online safety have issued guidance on the interactions of data analytics and privacy and contemporary issues such as recommender algorithms.
- **High-risk settings:** Peak bodies, key organisations and government departments in other higher-risk settings (including defence) have agreed to ethical principles for AI. These include the US Department of Defense, the Australian Signals Directorate and the Royal Australia and New Zealand College of Radiologists.
- **Select countries:**
    - Singapore has a clear preference for guidance and practical tools such as the 'A.I. Verify' toolkit to encourage and assist AI governance practices.
    - The US is taking multi-pronged approaches with voluntary approaches such as the AI Bill of Rights and the NIST AI Risk Management framework. But it is also one of the more advanced countries in having individual states laws on AI tools used in employment settings and restrictions on the use of facial recognition technology.

[91] InCorp Editorial Team, 'Business License in Indonesia: Risk-Based Classification Approach', *In.Corp website*, 15 July 2021, accessed 26 April 2023. The risk classifications are: low-risk business, medium-low risk business, medium-high risk business, high-risk business.

Figure 2: Domestic and international governance responses to support the safe and responsible deployment of AI



4 Managing the potential risks of AI

Many countries are considering responses to emerging risks from AI. Australia's current approach to date relies on a combination of:

- **a broad set of general regulations** that are mainly technology neutral (for example, consumer protection, online safety, privacy and criminal)
- **sector-specific regulation** (for example, therapeutic goods, financial services, food safety and motor vehicle safety)
- **voluntary or self-regulation initiatives** such as ethical principles for AI that provide guidance to businesses and governments to responsibly design, develop and implement AI.

The extent of potential risks from advances in AI, such as generative AI, remains uncertain. However, with the rapid acceleration of the development of AI applications, such as ChatGPT, and indications of increased capability, it is time for Australia to consider whether further action is required to manage potential risks while continuing to foster uptake.

Through this consultation and ongoing engagement with the community, we want to ensure that Australia has the right governance settings to respond to the rapid development of AI. This system-wide consultation will help support a coordinated and coherent response from the government to emerging issues. As the PC noted in its recent report, greater coordination between policymakers and regulators on diverse, complex and quickly evolving technologies will help avoid a piecemeal regulatory environment that can be a barrier to adopting these productivity-enhancing technologies.[92]

Drawing on the international initiatives discussed above, Figure 3 shows a range of possible responses to the governance of AI along with their strengths and limitations. Responses are mapped across the spectrum from voluntary to regulatory. Many of the possible responses would enable those deploying AI to improve their ethical and responsible practices voluntarily. At the other end, many of these principles or standards can be turned into mandatory requirements with consequences for failing to comply.

Governance responses contemplated for AI can also be considered in the context of ADM systems, whether or not those systems use AI.

In determining its governance responses, Australia must consider what is best for our economy and society. One issue for Australia is the extent to which it needs to harmonise its governance frameworks with those used globally or by its major trading partners. As a relatively small, open economy, international harmonisation of Australia's governance framework will be important as it ultimately affects Australia's ability to take advantage of AI-enabled systems supplied on a global scale and foster the growth of AI in Australia.

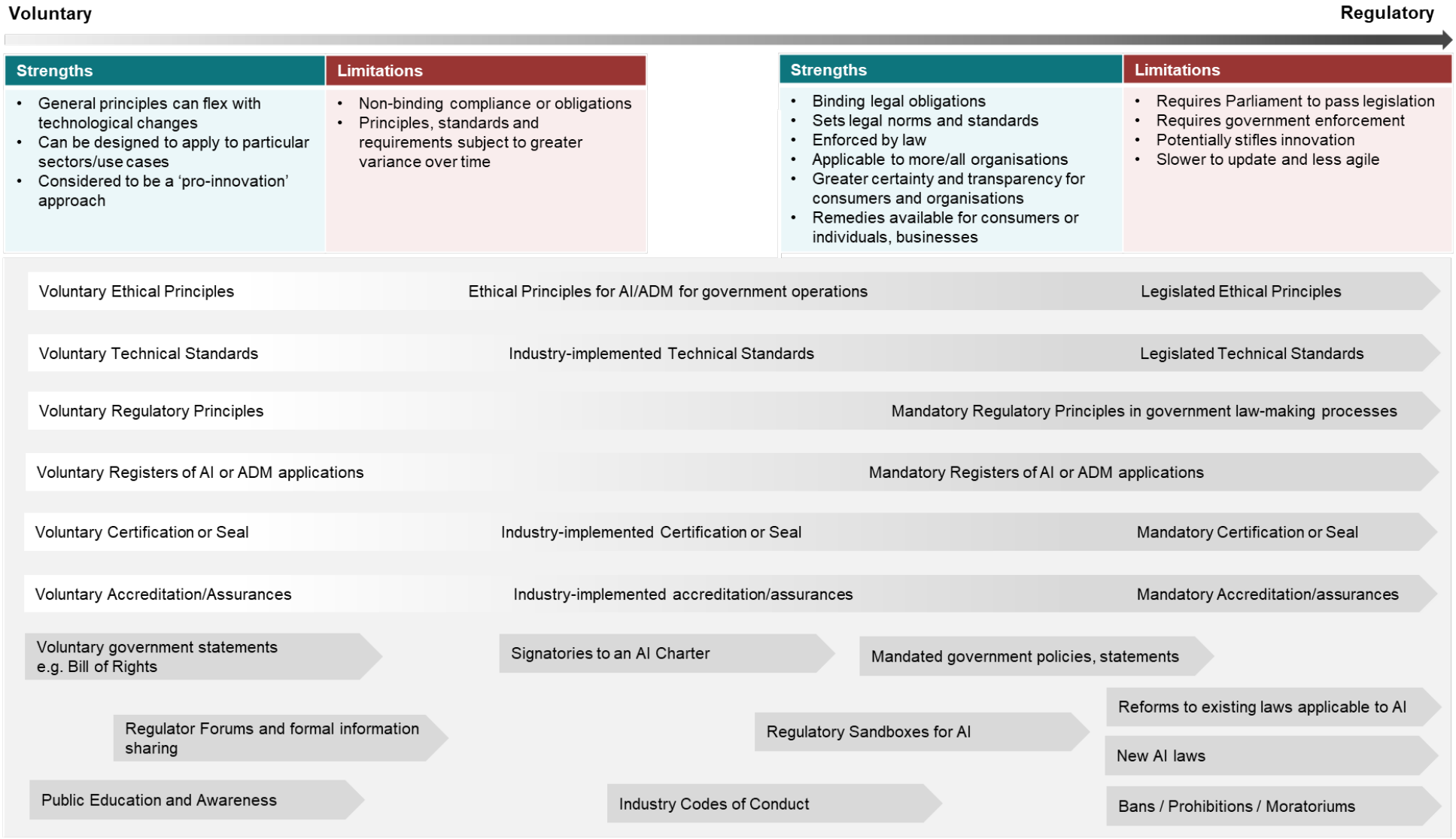
One of the areas of feedback being sought in this paper (see section 5) are the implications for Australia's domestic tech sector and our current trading and export activities with other countries if we took a more rigorous approach to ban certain high-risk activities.

Ultimately, it is proposed that governance measures adopted by Australia will be guided by the need to:

- ensure there are appropriate safeguards, especially for high-risk applications of AI and ADM
- provide greater certainty and make it easier for businesses to confidently invest in AI-enabled innovations and ADM activities and engage in these activities responsibly.

[92] The Productivity Commission, *5-year Productivity Inquiry: Australia's data and digital dividend*, Volume 4, p 90, 2023.

Figure 3: Options across the governance spectrum for AI



The options in Figure 3 are complementary and can be used together to achieve desired policy goals. They are broadly grouped as follows.

Regulations

These can be new AI-specific laws, reforms to existing sector-specific or general laws to protect the community, especially in high-risk settings where AI is being developed, applied or operated.

Regulations have the advantage of creating binding obligations that can be legally enforced and provide certainty, especially for smaller organisations. They can also provide remedies for businesses, consumers and individuals.

Industry self-regulation or co-regulation

This is where industry formulates its own rules through codes of conduct or voluntary schemes. Sometimes these codes can be accompanied with government legislative backing, for example, mandatory certifications or legislated standards.

Self-regulation options can often be implemented more quickly than government regulatory options. They are often more flexible and less burdensome for industry than government regulation.

Regulatory principles

These outline when and how policymakers should regulate. Principles can support greater regulatory coherence and alignment, as well as reduce the complexity of requirements.

For example, the UK is consulting on a set of cross-sectoral principles which regulators will be tasked with implementing within their regulatory remits.[93] This could include regulatory best practice principles such as technology-neutral and outcomes-based legislation rather than prescriptive legislation.

While outcomes-based legislation can provide flexibility for businesses, especially larger business, small businesses with fewer resources may prefer more prescriptive requirements as they provide greater certainty.

Regulator collaboration and engagement

Regulators play an important role by administering existing laws and reviewing and engaging with technical experts on the effectiveness of current laws. Greater collaboration and information sharing among regulators can reduce the compliance burden different regulators can place on the same regulated entities.

Examples of beneficial AI-related initiatives include the:

- Australian Actuaries Institute partnering with the Australian Human Rights Commission to develop guidance specific to AI and discrimination in insurance pricing and underwriting[94]
- ACCC, ACMA, OAIC and the Office of the eSafety Commissioner forming the Digital Platforms Regulators Forum to support collaboration, the sharing of information, and coordination on matters relating to digital platforms regulation.

Governance and advisory bodies and platforms

Bodies and platforms are being established to support AI governance outcomes, policies or initiatives. They have different roles such as aiding implementation or providing advice or information to support the community.

For example, Australia's Responsible AI Network will act as a gateway for Australian industries to uplift their responsible AI practices. Relevant developments in the US include:

[93] Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, UK Government, p 6, March 2023, accessed 20 May 2023.
[94] Actuaries Institute (AI) and Australian Human Rights Commission (AHRC), 'Guidance Resource: Artificial Intelligence and Discrimination in Insurance Pricing and Underwriting', *AI website,* December 2020, accessed 12 May 2023.

- a National AI Advisory Committee to advise the President and National AI Initiative Office on many AI-related issues, including accountability and legal rights[95]
- the National Institute of Standards and Technology's Trustworthy and Responsible AI Resource Centre providing a one-stop-shop for foundational content, technical documents and AI toolkits for AI actors to collaborate on trustworthy and responsible AI technologies.

The UK's Centre for Data Ethics and Innovation was formed 4 years ago as a governance expert body enabling the trustworthy use of data and AI.[96]

Enabling regulatory levers

Regulations can be designed to facilitate emerging technologies rather than hinder innovation.

For example, in February 2022 infrastructure and transport ministers across Australia agreed to an end-to-end regulatory framework for the commercial deployment of automated vehicles.[97]

Regulatory sandboxes like the financial services sandbox administered by ASIC have also helped allow a limited form of experimentation with some AI-powered technologies.[98] The UK and EU are also considering or putting in place an AI sandbox.[99]

Technical standards

Technical standards support technology interoperability, improve consistency for consumers and facilitate international trade.

Good standards are often designed by consensus of technical experts in industry-led organisations, but their development can take time.

While many standards for emerging technologies are voluntary, governments can choose to make them mandatory. This is the approach being taken by the European Union as part of the proposed AI Act.[100]

Assurance infrastructure and conformity processes or practices

These measures can test and verify that an AI system achieves or meets certain standards or quality requirements. This may extend to:

- identifying and accrediting data sources, given the potential risk to consumers of relying on AI-generated outputs
- building explainability into AI systems that could incorporate by-design considerations amongst other things to support greater transparency.[101]

These requirements can be implemented internally or via independent third parties. They can also be voluntary, industry-led or mandated through government laws, including technical standards.

---

[95] U.S. Department of Commerce, 'National Artificial Intelligence Advisory Committee (NAIAC)', *NIST website*, 21 April 2023, accessed 19 May 2023. It has also set up a subcommittee to consider matters related to the use of AI in law enforcement and advise the President

[96] Department for Science, Innovation and Technology, 'Centre for Data Ethics and Innovation', *Gov.UK website*, n.d., accessed 15 May 2023.

[97] National Transport Commission (NTC), *The regulatory framework for automated vehicles in Australia*, Australian Government, February 2022, accessed 15 May 2023.

[98] If certain things are satisfied, the sandbox allows businesses to test certain innovative models, subject to limits and conditions.

[99] European Commission (EC), 'First regulatory sandbox on Artificial Intelligence presented', *EC website*, 27 June 2022, accessed 17 May 2023; Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, UK Government, March 2023, accessed 20 May 2023.

[100] European Commission (EC), 'Regulatory framework proposal on artificial intelligence', *EC website*, 2022, accessed 14 April 2023.

[101] For example, the UK Information Commissioner's Office and the Alan Turing Institute have released guidance to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them. Information Commissioner's Office (ICO), 'Explaining decisions made with AI', *ICO website*, n.d., accessed 20 May 2023.

The US National Telecommunications and Information Administration (NTIA) is currently consulting on how best to ensure that AI systems work as claimed.[102]

NSW has developed an AI Assurance Framework and established an AI Advisory Committee to guide and oversee the use of AI in the NSW government. The framework allows the NSW government to assure their AI projects against its AI Ethics Framework as a way to build community trust.[103]

Policies, principles or statements guiding the operations of government

These can increase awareness of government expectations both internally (to ensure compliance) and externally (to improve practices and build public trust in how government is using AI).

The way that government implements new technologies can also influence private sector behaviour, such as through procurement and by modelling responsible AI practices.

Transparency and consumer information requirements

Initiatives such as publishing AI impact assessments provide the public with information about potential impacts of AI. They can also notify the public when AI applications are in use, similar to the objective of privacy policies.

For example, the City of Amsterdam hosts a searchable public AI register that provides information on the algorithmic systems that it uses. This includes documenting the decisions and assumptions made in the process of developing, implementing, managing and dismantling the algorithms.[104]

Bans, prohibitions and moratoriums

This is where governments prohibit an activity by law.

For example, facial recognition by governments is banned or severely limited in several US states and municipalities, while other states are banning ChatGPT in classrooms.[105] Similarly, some Australian jurisdictions (NSW, QLD, WA and Tasmanian) have also banned ChatGPT in schools.[106]

The draft EU AI Act proposes prohibiting social scoring and real-time biometric identification in certain circumstances unless exceptions apply.[107]

Public education and other supporting central functions

These are non-regulatory options that influence and encourage certain behaviour by increasing awareness and information to help achieve certain outcomes. For example, the UK's pro-innovation regulation framework will be supported by various functions, such as education and awareness and cross-sectoral risk assessments.[108]

By-design considerations

These are becoming increasingly popular as preventative mechanisms to ensure the design of appropriate AI or other digital systems. They include privacy by design, data protection by design

---

[102] National Telecommunication and Information Administration (NTIA), 'NTIA AI Accountability RFC', *Regultions.gov website*, 13 April 2023, accessed 14 April 2023.
[103] NSW Government, 'Artificial Intelligence', *Digital.NSW website*, n.d., accessed 14 April 2023.
[104] City of Amsterdam Algorithm Register Beta, 'What is the Algorithm Register?', *City of Amsterdam website*, 2020, accessed 14 April 2023.
[105] N T Lee and C Chin, 'Police surveillance and facial recognition: Why data privacy is imperative for communities of color', *Brookings website*, 12 April 2022, accessed 14 April 2023; Associated Press 'States Push Back Against Use of Facial Recognition by Police', *U.S. News website*, 5 May 2021, accessed 14 April 2023; A Johnson, 'ChatGPT In Schools: Here's Where It's Banned - And How It Could Potentially Help Students', *Forbes website*, 18 January 2023, accessed 18 April 2023.
[106] C Cassidy, 'Queensland public schools to join NSW in banning students from ChatGPT', *The Guardian website*, 23 January 2023, accessed 15 May 2023; A Davis, 'ChatGPT banned in WA public schools in time for start of school year', *ABC News website*, 30 January 2023, accessed 15 May 2023; M Whitfield, 'Can teachers spot a ChatGPT fake? Tasmania's education department says yes', *ABC News website*, 25 January 2023, accessed 15 May 2023.
[107] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council - Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts', noting these are still being negotiated.
[108] Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, UK Government, March 2023, accessed 20 May 2023.

(DPbD) and safety by design. They ensure AI systems are designed with privacy or safety considerations in mind from the outset.

For example, DPbD allows digital systems to automatically delete data once it is no longer needed for the specific business purpose. These design concepts can be voluntary or mandated through laws.

Risk management approach

A risk management approach could guide the implementation of any of these options. Such an approach can be voluntary or mandated through laws. Internationally there is a trend towards risk management guiding governance responses.

Some countries are using risk management principles to ban AI applications based on certain criteria or in select use cases. The EU's draft AI Act, Canada's mandatory directive on automated decision-making and the US NIST's AI Risk Frameworks are all underpinned by risk management principles.

As these risk management approaches grow in popularity, the merits and limitations of a risk management approach in Australia warrants exploration - either as a voluntary self-regulation tool or through government regulation.

A possible draft risk management approach for AI is outlined in Box 4. Feedback and comment are invited on this possible draft risk approach, which builds on the EU's proposed AI Act and Canada's directive. Further detail on possible elements is outlined in **Attachment C**.

Box 4: A possible draft risk management approach for managing AI risks (for feedback)

A risk management approach:

- caters to the context-specific risks of AI, so requirements can change depending on how the AI is deployed
- allows for less onerous obligations for lower risk AI uses
- allows AI to be used in high-risk settings where the risk and costs are justified and can be explained.

The first step to apply the risk management approach is for an organisation to consider the risk level of the AI application being considered. The second step to applying the framework is to determine which requirements apply, based on the assessed risk level. Under this approach, the risk management requirements for medium and high-risk applications of AI are commensurately more onerous than for low-risk applications of AI.

As this approach focuses on context-specific risks and potential impacts, the use cases provided are only for illustrative purposes.

		Risk management requirements / obligations					
Risk level	**Example/indicative use cases**	**Impact assessment**	**Notices**	**Human in loop**	**Explanation**	**Training**	**Monitoring and documentation**
Low risk Minor impacts that are limited, reversible or brief	• Use of AI in computer chess systems • Algorithm-based spam filters that identify and block unwanted or dangerous emails • AI-enabled recommendation engines to enable personalised online shopping recommendations based on users' browsing history, preferences and interests • AI-enabled applications that automate discrete business processes (e.g. processing business expenses) • AI-enabled chatbots that direct consumers to service options according to existing processes	Basic self-assessment	N/A	N/A	General explanation	Users must be trained	General internal monitoring and general documentation on functionality of the system

Medium risk High impacts that are ongoing and difficult to reverse	• AI-enabled application that preliminarily assesses a business loan applicant's creditworthiness • Use of generative AI in educational settings to assess the performance of teachers and students • Use of AI-enabled chatbots to direct citizens to essential or emergency services • AI-enabled applications in hiring and employee evaluation processes • Use of AI to generate patient records in care settings	Comprehensive and specific self-assessment	Plain language notice	Self-assess and implement appropriate and meaningful points of human involvement commensurate with the risk	Specific explanation of decision, AI output of application made available to users	Recurring training	Special internal frequent monitoring and specific documentation on design and functionality of the system
High risk Very high impacts that are systemic, irreversible or perpetual	• Use of AI-enabled robots for medical surgery • Use of AI in safety-related car components and in self-driving cars to make real-time decisions	Impact assessment peer reviewed by external experts	Publish system explanation	Must have meaningful human intervention at specific points and final decision made by human/s	Specific explanation of decision, AI output or application and made available publicly or to experts and regulators	Recurring training and a means to verify that training has been completed	External audit of the special internal frequent monitoring and specific documentation

5 How to get involved

We welcome your contributions as the Australian Government considers regulatory and governance responses to:

- mitigate the potential risks from AI and ADM
- increase public trust and confidence in their development and use.

Consultation questions are included below. Please submit your answers at https://consult.industry.gov.au/supporting-responsible-ai

Definitions

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

Potential gaps in approaches

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

Responses suitable for Australia

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

Target areas

6. Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?
7. How can the Australian Government further support responsible AI practices in its own agencies?
8. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.
9. 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?
    b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.
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?
11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

Implications and infrastructure

12. How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?
13. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

Risk-based approaches

14. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?
15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?
16. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?
17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?
18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?
19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?
20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:
    a. public or private organisations or both?
    b. developers or deployers or both?

Attachment A: Overview of current Australian Government initiatives/work relevant to AI

Topic	Agency	Short Description
Whole-of-economy	Attorney-General's Department	**Privacy Act Review:** The Attorney-General's Department has released the *Privacy Act Review report* which recommends reforms to ensure Australia's privacy settings empower consumers, protect their data and best serve the Australian economy in the digital age. The Australian Government is developing its response to the 116 proposals in the report following consultation with the public. The proposals include ones affecting the use of personal information in automated decision making, and the use of algorithms to target users.
		The Ministerial Roundtable on Copyright: at its inaugural meeting on 23 February 2023, the Roundtable brought together 30 organisations from a wide range of sectors with an interest in copyright**.** Chaired by the Attorney-General, participants at this meeting identified the uncertain implications of AI for copyright law – including in relation to text and data mining, database protection, and authorship of AI-created works – as an issue. This issue will be discussed further in another Roundtable to be hosted by the Attorney-General's Department later in 2023.
	IP Australia	**The AI Working Group of the Intellectual Property Policy Group:** is exploring issues at the intersection of AI and IP. These issues may include: • the role of IP in the development and adoption of AI systems • IP rights and the increasing use of AI in innovation and creativity • implications of AI on the rules of IP protection • AI and its impacts on consumers' interaction with IP such as trademarks. This group comprises representatives from relevant Australian Government agencies including the Attorney-General's Department, DFAT, DISR and the DTA. This group provides an opportunity to seek views from diverse stakeholders on how the current IP settings intersect with AI development and adoption in Australia and whether the government could consider any potential changes.
	ACCC, ACMA, eSafety Commissioner and OAIC	**Digital Platform Regulators Forum:** 2022–23 priorities of the forum between the ACCC, the ACMA, eSafety Commissioner and the OAIC include the impact of algorithms, increasing transparency, protecting users, and increased collaboration and capacity building. The Forum's Digital Technology Working Group is looking closely at generative AI, focusing on large language models. The purpose of this exercise is to gather information about this technology and build a shared understanding of its implications across each member's regulatory sphere. This work includes conducting workshops and desktop research as well as liaising closely with academic experts and other relevant Australian Government agencies.
	ACCC	**Digital Platform Services Inquiry:** The ACCC is conducting a 5-year inquiry with 6-monthly reports into markets for the supply of digital platform services, including search engine, social media, app store, online private messaging and electronic marketplace services. The inquiry's fifth interim report, published in November 2022, recommends regulatory reform to address matters identified in the ACCC's digital platform reports to date, including a new regulatory regime to apply to the largest platforms to address issues such as a lack of transparency in the use of algorithms, data advantages, and self-preferencing, including through the use of algorithms.
	Department of Home Affairs and Department of Infrastructure, Transport, Regional Development, Communications and the Arts (DITRDCA)	A **report by the former House of Representatives Select Committee on Social Media and Online Safety** makes two recommendations on the use of algorithms in digital platforms, including examination of the types and scale of harms caused as a result of algorithm use (recommendation 13) and the potential mechanism to require digital platforms to report on their use of algorithms (recommendation 14). Recommendation 13 also suggests the development of a roadmap for Government entities to build skills and expertise for the next generation of technological regulation. As noted in the Australian Government response to this report, the Department of Home Affairs and DITRDCA are progressing work to understand the operation of algorithms on digital platforms. They will jointly report back to the government by the first quarter of 2024 with options to build capability

Topic	Agency	Short Description
		around future algorithm research and expertise, and with advice on whether government regulation of algorithms is required and, if so, what options for regulation are available.
	DISR	**The AI Ethics Principles (2019) and pilot (2021):** the AI Ethics Framework aims to guide businesses and government to responsibly design, develop and implement AI. It has 8 voluntary AI Ethics Principles to ensure AI is safe, secure and reliable. The pilot resulted in the publication of various case studies sharing insights and best practices which were adopted by some of Australia's biggest businesses to operationalise the ethics principles.
	eSafety Commissioner	**Safety by Design:** Safety by Design puts user safety and rights at the centre of the design and development of online products and services. Rather than retrofitting safeguards after an issue has occurred, Safety by Design focuses on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur. This proactive and preventative approach focuses on embedding safety into the culture and leadership of an organisation. It emphasises accountability and aims to foster more positive, civil and rewarding online experiences for everyone. eSafety has worked with representatives from across the digital industry to produce a range of Safety by Design resources, including: • a set of principles that position user safety as a fundamental design consideration • interactive assessment tools for enterprise and start up technology companies • resources for investors and financial entities • engagement with the tertiary education sector to embed Safety by Design into curricula around the world.
		Tech trends and challenges position statements: The eSafety Commissioner publishes global public position statements on tech trends and challenges, which provide guidance and support for the public, whilst informing eSafety's regulatory posture. The papers consider the online safety implications of a technology, regulatory and technical updates, and guidance for industry to mitigate risks, including through adoption of Safety by Design measures. The position statement on recommender systems and algorithms provides an overview and eSafety's position on recommender systems, which prioritise content or make personalised content suggestions to users of online services. eSafety is currently developing a position statement on generative AI.
AI use in government	Department of Finance	**The initial Data and Digital Government Strategy** sets out the Australian Government's intent to harness analytical tools and techniques, including AI and machine learning to predict service needs, gain efficiencies in agency operations, support evidence-based decisions and improve user experience. The Strategy will also embed integrity and ethical behaviour in how Australian Government agencies use data and digital technologies through the adoption of a whole-of-government data ethics framework. The framework will have implications for the use of AI in the Government's operations. A final Strategy is scheduled to be released by the end of 2023.
	DTA	**Australian Government Architecture (AGA):** includes guidance to help public sector adoption of AI.
	Commonwealth Ombudsman	**Automated decision-making better practice guide (updated 2020):** provides guiding principles for the use of automated systems within Australian Government agencies, including to ensure these decisions are consistent with administrative laws and principles of fairness, accountability and transparency.
Sector / domain specific	Australian Communications and Media Authority (ACMA)	**The artificial intelligence in communications and media paper** explores the implementation of ethical principles in communications and media markets and the potential risks to consumers in interacting with automated customer service agents.

Topic	Agency	Short Description
	DITRDC	**New laws targeting online misinformation and disinformation** are being introduced to provide ACMA with new powers to combat online misinformation and disinformation. These powers will apply to misinformation and disinformation content on digital platforms generated by AI technology (such as bots and where dissemination occurs using automated means). A registered industry code or ACMA standard may include requirements for platforms to crack down on endemic bots spreading false information.
		End-to-end regulatory framework for automated vehicles: Infrastructure and Transport ministers across Australia have agreed to an end-to-end regulatory framework for the commercial deployment of automated vehicles.
	Therapeutic Goods Administration (TGA)	**Reforms to medical devices regulations and accompanying guidance (2021):** to clarify requirements for software and mobile apps used in medical contexts (known as software as a medical device, or SaMD).
	Department of Education	**Best-practice framework to guide schools in harnessing AI tools to support teaching and learning:** Education ministers agreed on 27 February 2023 to develop an evidence-based, best practice framework to guide schools in harnessing AI tools to support teaching and learning, and to establish a Taskforce to develop the framework. The framework will include the following four key objectives: • safe and ethical use of generative AI tools • best practice implementation of generative AI tools in the classroom to lift student outcomes • reducing workload burden and administration using generative AI tools • establishing education-specific standards and governance to meet the needs of Australian schools.
		OECD's High Performing Systems of Tomorrow Project: Australia is a participating country in phase 2 of this OECD project that includes consideration of the implications of AI for education systems.
		AI in higher education settings: The Tertiary Education Quality and Standards Authority, which regulates higher education providers, has provided a range of advice in relation to AI, including on academic integrity and the use of AI in higher education classrooms.
Specific AI programs and initiatives	DISR via CSIRO	**The National AI Centre is** funded by the Australian Government and coordinated by CSIRO. It recently established the Responsible AI Network (RAIN) to act as a gateway for Australian industries to uplift their practice of responsible AI by bringing together a national community of practice, guided by world leading expert partners, and enabling Australian businesses with best practice guidance, tools and learning modules.
		The **Next Generation AI and Emerging Tech Graduates programs** provide scholarships to post-graduate students to study and work with industry partners and address skills shortages in AI and emerging technologies.
	DISR	**The Responsible AI Adopt program:** will establish centres aimed at supporting Australian small to medium enterprises (SMEs) to adopt AI technologies responsibly to elevate and power their businesses to better compete in international and interstate markets.

Attachment B: European Union AI Act risk level

Classification of AI into risk levels under the European Commission's proposed EU AI Act (2021)[109]

AI Act risk level	AI types	Requirements
Unacceptable risk	• Practices that have a significant potential to manipulate persons through subliminal techniques • Practices that exploit the vulnerabilities of specific vulnerable groups (e.g. children, persons with disabilities) • AI-based social scoring done by public and private authorities	• Banned
High risk	• AI used for 'real-time' and 'post' remote biometric identification of people • AI used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity • AI that is used in the education sector and to determine further access to education • AI used for recruitment or evaluating job candidates, or for monitoring and evaluation of employees • AI used by public authorities that determines access to public assistance benefits and services • AI used to evaluate the creditworthiness of people (with the exception of AI systems put into service by small scale providers for their own use) • AI to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid • AI that assesses the risk of offending, reoffending, victimhood • AI used by law enforcement to detect people's emotional states (e.g. as polygraphs) • AI used by law enforcement to detect deep fakes • AI used by law enforcement in connection with a criminal offence (evaluating the reliability of evidence, predicting the occurrence or reoccurrence of an actual or potential criminal offence, profiling people, identifying patterns, assessing risk) • AI systems used by public authorities to assess a risk • AI used by public authorities to verify documents • AI that assists public authorities in examining applications for asylum, visa and residence permits and associated complaints	• Use high-quality training, validation and testing data • Establish documentation and design logging features • Ensure appropriate degree of transparency • Ensure robustness, accuracy and cybersecurity **Provider obligations** • Establish and implement quality management • Keep up-to-date technical documentation • Undergo conformity assessment and re-assessment (for modifications) • Affix CE marking and sign declaration of conformity • Register AI system in EU database • Conduct post-market monitoring • Collaborate with market surveillance authorities **User obligations** • Operate AI system in accordance with instructions of use • Ensure human oversight • Monitor for possible risks • Inform provider of any serious incidents or malfunctioning
Limited risk	• Human impersonation (i.e. chatbots)	• Notify humans that they are interacting with an AI system • Notify humans that emotional recognition or biometric categorisation system are applied to them • Apply labels to deep fakes
Minimal risk	• AI-enabled video and computer games • Spam filters	• No mandatory obligations • Encourage voluntary codes of conduct for low-risk AI systems

[109] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council - Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts', *The AI Act website*, 2021, accessed 17 April 2023.

Attachment C: Possible elements of a draft risk-based approach

A draft possible risk-based approach could comprise elements to support safe and robust practices by organisations developing and adopting AI in Australia to increase community trust and confidence.

Possible elements

Impact assessments

These are important measures to ensure organisations appropriately consider and mitigate potential risks. Publishing the final results of impact assessments provides greater transparency about how organisations are considering and managing the potential risks of AI. Impact assessments to be peer reviewed by external experts where the potential risks are high.

Notices

These are important for informing users where automation or AI is used in ways that materially affect them. Without this notification, individuals sometimes do not know that AI systems have been used. This may hamper their ability to seek reviews of decisions or lead to a lack of trust if and when they find out that AI is being used.

Human in the loop/oversight assessments

There may be circumstances when having humans in the loop or involved in reviewing or monitoring an AI systems' operations are important for minimising potential risks and supporting public trust and confidence. Assessments regarding when human oversight is appropriate may be based on considerations of potential risks as well as criteria regarding:

- the decision's complexity
- the level of discretion involved
- the extent of potential damage of a wrong decision
- how much specialist knowledge is required.

It is increasingly important to understand what human involvement is possible or desirable and to carefully design "meaningful" human involvement. For example, the CSIRO has suggested this design should:

- consider human competency and its limits
- identify the most suitable mechanisms (either human, technical or a combination of both) to deliver better control over dynamic systems like AI.

Human in the loop requirements may also not be appropriate for example where:

- the benefits of an application are dependent on efficiency at scale[110]
- there is increased speed and scale of automation[111]
- any potential impact on individuals is minor.

In these instances, it may not be feasible or desirable for a human to intervene in an automated process before individuals are affected. As such, involving a human in the AI process should be considered one component of a broader framework to reduce the potential risks associated with the use of AI and ADM.

Explanations

Explanations build on the concept of notices and transparency, which are important drivers for building greater public trust and confidence, as individuals are more likely to trust things they are aware of and understand. Explanations aim to provide sufficient clarity around the decision or outcomes, so that individuals affected by the decision can understand the factors that led to the result.

Training

Providing adequate employee training in the design, function and implementation of the AI so they can better understand the potential risks, how they can be mitigated and they can explain and oversee the AI's operation. The breadth of training should increase in proportion with the level of potential risk. Where there are one or more humans responsible for overseeing the AI, they must be competent, properly qualified and trained and have necessary resources to effectively supervise the AI.

---

[110] Bell, G., et al. (2023, March 24). *Rapid Response Information Report: Generative AI. p 11.*
[111] K Leins and A Kaspersen, 'Seven Myths of Using the Term "Human on the Loop": "Just What Do You Think You Are Doing, Dave?"', *Carnegie Council for ethics in International Affairs website*, 9 November 2021, accessed 26 May 2023.

Monitoring and documentation

Ongoing monitoring is important for ensuring AI systems operate as intended and that any adverse or unintended impacts such as unwanted bias are identified and rectified. The intensity of the monitoring will increase with risk (more frequent tests are needed where the potential risks are higher). Documentation in the design, function and implementation of AI also helps to develop a better understanding of the potential risks, how they can be mitigated and appropriate accountability for those involved and more senior decision-makers on the overall appropriateness (and impacts) of the AI product or service.