Safe and Responsible AI in Australia Discussion Paper

ADM+S Submission

Lead author: Kimberlee Weatherall

Contributing authors: Zofia Bednarz, Jose-Miguel Bello y Villarino, Jean Burgess, Loup Cellard, Tegan Cohen, Henry Fraser, Jake Goldenfein, Timothy Graham, Fiona Haines, Paul Henman, Nataliya Ilyushina, Jenny Kennedy, Jackie Leach Scully, Dennis Leeftink, Suvradip Maitra, Rita Matulionyte, Anthony McCosker, Robert Mullins, Kelsie Nabben, Christine Parker, Thao Phan, Flora Salim, Aaron Snoswell, Julian Thomas, Melanie Trezise, Libby Young, Jacky Zeng.

ARC Centre of Excellence for Automated Decision-Making and Society

4 August 2023

DOI: 10.25916/catx-q405

About ADM+S

The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) is a cross-disciplinary, national research centre which commenced operations in mid 2020. ADM+S has been established and supported by the Australian Research Council to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making (ADM).[1] Focus areas for ADM+S research are news and media, social services, health and transport. ADM+S brings together nine of Australia's leading universities, and more than 80 researchers across the humanities, social and technological sciences, together with an international network of partners and collaborators across industry, research institutions and civil society. More information about the ADM+S, our researchers and research projects can be found on our website: www.admscentre.org.au.

Our interest in supporting Responsible AI

ADM+S welcomes the opportunity to respond to the Department of Industry, Science and Resource's consultation on Safe and Responsible AI in Australia. Resolving the legal and regulatory challenges posed by artificial intelligence—and how regulatory systems can promote responsible, ethical, and inclusive AI and ADM for the benefit of all Australians—is one of the Centre's founding objectives.

This submission

This submission is the product of a collaborative process involving direct contributions from the above researchers from ADM+S, as led and consolidated by Professor Kimberlee Weatherall (University of Sydney Law School). ADM+S researchers come from many different institutions, disciplines and perspectives. It should not be assumed that every contributing author, or every member of the Centre subscribes to every comment or recommendation made below. The submission represents our best effort to consolidate research and thinking in a way that can be useful to the Department and the Commonwealth Government more generally.

[1] The ARC Centre of Excellence on Automated Decision-Making and Society is funded by the Australian Research Council (CE200100005)

Summary

The ADM+S is pleased to have this opportunity to engage with an important and complex question which confronts Australia: how should the Australian federal government take action - regulatory or otherwise - to promote artificial intelligence (and automated decision-making) that is safe and responsible? In our view, 'responsible' AI must also be inclusive, accountable, and genuinely beneficial - for Australia's people, society, economy, and environment. In this submission – which is the product of research, inputs, and debate across our multi-disciplinary ARC Centre of Excellence on Automated Decision-Making and Society - we address this question in the following way.

What is new here?

After discussing definitions, the submission seeks to distil what is (arguably) new and/or different about recent developments in automated decision-making (ADM) and AI technologies, and their legal, social, human and environmental impacts. We believe this as a critical first step that must be taken before thinking about what role tools such as a risk-based framework can play. Here we seek to highlight not only well-known harms and challenges brought about by these technologies (such as privacy risks, or unfair bias and discrimination), but also the new challenges and shifts that are emerging as a result of the rise of generative AI/foundation models and associated developments, including the broad take-up and rapid integration of generative AI, and its broad potential as a general purpose technology embedded in complex supply chains.

What is the impact on law, regulation, and policy?

In our view It is often problematic to target regulation at a particular. As a rule, regulatory efforts should be directed at categories of activities, behaviours, decisions or outcomes. This is consistent however with recognising that the impacts of AI/ADM are multiple and broad, and will demand a range of responses across the Australian government. In particular, we argue that these impacts require consideration of:

1. **How a range of laws are framed:** Australia will need to revise existing legal regimes: In Australia - in contrast to similar countries - some AI-generated harms lack any legal remedy. We also need to have a collective conversation about how to regulate - and whether to ban - certain new capacities. DISER is well-placed to lead the latter conversations in particular, which require a societal conversation about the capacities of emerging technologies that affect how we live, just as we've had past conversations about cloning, gene editing and nano-tech.
2. **Enforcement:** specific attention to AI-created enforcement challenges, particularly where enforcement is funnelled through under-resourced and over-worked regulators and mediation. We need more enforcement pathways and access for interest groups and collective actors (such as unions and advocacy groups), transparency and access to information and evidence, and consideration of how burdens of proof and responsibility should be allocated across complex AI supply chains stemming from data collection through to deployment.

3. ***Ex ante* risk mitigation to reduce individual and systemic harms*:*** introduction of requirements in the design and development phase aimed at drawing attention to and addressing potential risks before deployment that affects people, society and the environment.

The risk-based approach

ADM+S offers **qualified support for a risk-based approach.** The main potential benefits of a risk-based approach are (a) the ability to avoid or mitigate harms before they happen, at the design and development stage rather than waiting for *ex post* litigation; (b) promoting better (safer, more responsible) design; as well as incorporating (c) ongoing obligations on developers of systems to engage in monitoring and addressing risks.

However, the success of any risk-based framework in Australia will depend on the extent to which we address current gaps in both our *rules* (ie laws/legal frameworks) and our *enforcement capacities*. Other countries that are considering risk-based approaches are simultaneously, or have already, addressed these issues. In short: when we adopt a risk-based approach, we are requiring firms to identify, and mitigate, certain risks of harm. But for this to work, **there must be some kind of 'or what'?** For this to lead to genuine improvement in the technologies applied, there must be a risk of consequences - *liability for harms caused if organisations fail to take mitigating action.* That means laws prescribing the act that creates the risk of harm, and a credible threat that that law can be enforced.

In addition, we argue that:

- any risk assessment must take into account the sociotechnical context. Our submission highlights in particular questions of diversity and inclusion (especially in light of high levels of **digital exclusion** across Australia, especially as experienced by people in regional and remote areas), the impact of supply chains (including the actors who collect and clean data), and environmental concerns;
- In identifying what kinds of uses of technology are low, medium, or high risk, it will be critical to bring knowledge from a range of perspectives: both technical and non-technical. The need to ensure cutting edge knowledge is made more widely available is something we noted above; and
- whether though providing guidance, connecting researchers or, in larger organisations insisting on multidisciplinary and diverse teams, people from a wide range of backgrounds, including people affected by AI systems must be involved as we consider what precautions are needed around a proposed use of AI.

Questions of design for a risk-based approach

A first core question is **who decides** whether a system is low, medium, high (or very high) risk? Risk is multi-dimensional (it varies by type of impact/harm, severity and probability, and can shift over time) meaning that fixed categories may not work well, but the party best placed to assess the risk of a system (who could be the developer, or the deployer) may have incentives to underestimate risk. There may be mechanisms to manage this, including for example by requiring publication of risk assessments for at least some systems, and/or setting 'default' categories with the ability of entities to show that their system is lower risk than the default

would suggest.

ADM+S further makes a number of comments on elements of the risk-based approach set out in the Discussion Paper:

- Three 'categories' of risk may be insufficient, and that descriptions of the different categories could give rise to some anomalous results (such as where a risk is brief *and* severe, or where it is brief and 'reversible' (for example, as a loss of social benefits is 'reversible' as payments can be restored) but has lasting impacts (say because a person has become homeless in the meantime when they could not pay their rent);
- More guidance and deeper thinking will be required regarding the 'risks' that must be considered; some ideas are offered and we draw attention to our discussion of environmental risks (Consultation Question 2) and issues around digital exclusion in Australia (Consultation Question 14);
- In terms of the requirements, we draw attention to the absence of any reference to data quality considerations, suggest that further thinking is required regarding notices/transparency/explainability, and note that the appropriate role of human oversight ('human in the loop') is complex – indeed including human oversight can sometimes increase, or obscure, problems with an AI system.

What of Foundation Models in particular?

Much of the discussion throughout our submission is relevant to Foundation Models: in discussing what is new/different about AI; what new capacities require a societal-level discussion); and the need to connect government with cutting-edge research and ensure new research is incorporated into efforts to guide and educate to developers and deployers as well as the broader public.

There are concerns about the applicability of risk-based approaches in relation to foundation models, which the EU is presently grappling with. The submission discusses these developments, and ADM+S can offer further expertise as required and as regulatory positions consolidate internationally.

Finally, we note that foundation models raise genuine questions around the consolidation of power over the generation and transfer of knowledge. Steps may need to be taken to ensure research and pedagogical access for Australian researchers; it would be detrimental, for example, to the country's research efforts if researchers from certain countries where models are trained had preferential access for the purposes of research.

Non regulatory actions

We have a number of suggestions in relation to non-regulatory actions the government could take. In particular we focus on three in this submission, emphasising that the government should:

1. Invest in involving the Australian public in discussions about the direction of AI technology and its application: in order for the Australian public to trust AI technology and support its use, the current technocratic conversation needs to be broadened to be

more inclusive. DISER is well-placed to lead such efforts, using established methods for science consultation and participatory governance.

2. Invest in education at multiple levels across society and the economy, to reduce the knowledge gap between the AI specialists (who create AI programs), and AI end users (who are responsible at the coalface for the deployment of AI, impacting themselves or other parties) or subjects (who are impacted by decisions and/or actions using AI).
3. To better address the rapid development and rapid *deployment* of AI technologies including new models and methods, adopt mechanisms to better connect leading research with government and the broader set of technology users, including in particular those involved in assessing the risks of applying AI. These mechanisms could be based on mechanisms used to inform the current inquiry, including by activating Australia's learned academies more regularly.

In conclusion, throughout the submission, we have sought to highlight some relevant expertise in the Centre, although there is much more not mentioned here. We look forward to continuing the discussion, and are happy to provide more information or connect interested policymakers with expertise across any matter canvassed in this submission.

Table of Contents

Submission

Consultation Question 1: Definitions

Consultation Question

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

ADM+S Answer: The scope of regulation should be determined according to capacities, behaviours, and/or impacts - not technologies. We also see some risk that the definitions will become outdated (in ways we outline in our submission), while acknowledging the desirability of consistency in definitions across jurisdictions. A definition of 'foundation model' would also be helpful.

How will the definitions be used?

Defining the **technology** is not the same thing as defining the **scope** of any appropriate regulatory regime. We have concerns about defining the scope of regulation by reference to AI.

It is problematic to target regulation at a particular technology - especially one that presents in multiple forms, is rapidly developing, and is put to multiple uses in regular commercial and government activity. As a rule, regulatory efforts should be directed at categories of activities, behaviours, decisions or outcomes.[2] Regulation whose scope is determined by reference to definitions of technology risks inconsistency.

Creating a risk-based regime only for AI:

- Could be a *disincentive* to use systems defined as AI (to avoid regulation).
- Incentivizes firms to argue their systems 'aren't AI' to avoid scrutiny (as suggested by experience with the NSW AI Assurance Framework).
- Fails to address related and similar problems with non-AI systems,[3] or means those risks are governed by different laws - creating overlaps and more costs for business (what if a system has both 'intelligent' and 'non-intelligent' functions?).

These points are well illustrated by Australia's Robodebt scheme, and many similar automated schemes across the world, where the technology used was basic and many decades old, and definitely not AI – these systems would not be captured by an AI-specific risk framework. This example also illustrates that how a technology is used, for what purposes, and who is subjected to it, all matter as much as the definition of the technology itself.

ADM+S focuses on automated decision-making (ADM) because it ensures attention to a broader category of activities with significant impacts on people and society, some of which involve AI, and some of which do not.[4] Canada's risk assessment system for government-decision-making also applies to ADM.[5]

[2] Lyria Bennett Moses, 'How to Think about Law, Regulation and Technology – Problems with "Technology" as a Regulatory Target' (2013) 5(1) Law, Innovation and Technology 8.
[3] For example, rules-based systems that rely on statistical analysis are prone, like AI systems, to reflect biases and discrimination in underlying data: see, eg, 'Automating Society 2019' *AlgorithmWatch* (Web Page).
[4] We acknowledge that the rise of generative AI suggests that ADM does not comprehensively cover all AI-based activities that can impact human beings. We would, however, argue that when designing a regulatory regime, it is likely better than referring only to AI.
[5] Government of Canada, *Directive on Automated Decision Making* (April 2023) s 6.1.

Are these useful definitions?

Most definitions in the discussion paper are consistent with international standards and are similar to those being adopted by other leading jurisdictions. This has benefits for integration and understanding across borders, including for industry. We also note, however, four qualifications, regarding the use of these definitions to define the scope of any regulatory system.

The proposed definitions risk becoming outdated

There is some risk that the proposed definitions will become outdated, if they are not already. Some AI systems involve *AI agents* defining sub-objectives for a task they are trying to achieve,[6] and some AI systems modify aspects of their own objectives.[7] It is unclear whether these paradigms would be captured by the current language of 'human-defined objectives' (it is possible that an *overall* human-defined objective could be sufficient?).

The proposed definition of 'machine learning' is circular, and inaccurate

A more accurate and non-circular definition might read;

> 'Machine learning refers to AI systems that derive patterns from training data using algorithms or computational means.[8] These patterns are typically utilised to synthesise domain knowledge and can be applied to new data for prediction or decision-making purposes.'

A definition of 'foundation model' would be helpful

The paradigm of 'foundation models' represents a substantial shift in how AI technology is developed and deployed.[9] Instead of creating individual bespoke AI systems, AI developers are increasingly adapting and/or building on existing general purpose large-scale models for specific applications. This change in AI development has important ramifications - for instance, centralising biases and power in one technical and institutional location, or distributing and obfuscating the site of responsibility for down-stream AI system decisions.

For these reasons, defining 'foundation model' explicitly may help to clarify the discussion and questions. A suggested definition for foundation model, adapted from the Stanford Centre for Research on Foundation Models definition[10] could read:

> 'A foundation model is any model trained on broad data that can be adapted (eg, fine-tuned or integrated as one part of a larger system) to a wide range of downstream tasks. Foundation models can *focus* on one data-type (eg text only), or can be Multimodal Foundation Models (MfMs) - with the ability to process and learn from multiple data types (eg images and text).'

The definition of automated decision-making could be improved

We suggest a simpler formulation which explains the idea of automation and avoids confusion as to what may or may not be a 'technological' system. Drawing on the Canadian Directive on ADM, we recommend defining ADM as:

---

[6] See, eg, *AUTOGPT* (Web Page).
[7] See Stuart Russell, 'The History and Future of AI' (2021) 37(3) *Oxford Review of Economic Policy* 509.
[8] We add 'computational means' here because some relevant processes, such as 'bit quantisation', are not so much algorithms but techniques. See, eg, 'Quantization', *Hugging Face* (Web Page).
[9] See Rishi Bommasani et al, 'One the opportunity and risks of foundation models' (2021) arXiv:2108.07258.
[10] Ibid.

'A technology that assists or replaces the judgement of human decision-makers. Automated decision making includes systems that:

- Make a final decision
- Make interim assessments or decisions leading to a 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 decision-making systems range from those that automatically apply pre-defined rules to those that make predictions and decisions based on machine learning and other forms of artificial intelligence.'

Consultation Question 2: current regulatory settings and gaps

Consultation Question 2

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?

ADM+S Answer: This question is best answered by first understanding what has changed, or what is new, which we do immediately below. A picture of existing gaps follows from the changes we identify:

1. Australia will need to re-consider, and in some cases re-frame, a wide range of existing legal frameworks
2. Specific gaps around the application of laws to the process of training large models;
3. Where laws do exist, features of the technology and AI supply chains gives rise to enforcement challenges and gaps that will need revision; and
4. New capacities are emerging where DISER should initiate a broad societal conversation about what use is and isn't acceptable.

In addition, research, including at ADM+S is emerging regarding the **environmental** impacts of AI, and consideration of environmental impacts is being built into regulatory proposals overseas such as the EU AI Act. Environmental impacts must be a consideration when developing regulation, and a risk-based approach in Australia.

What is new[11] or different about AI and associated trends (data collection, digitisation, automation)?

We cannot analyse gaps in regulation without first identifying what is new or what has changed with the increased power and integration of AI, and how these changes may have impacts on people. ADM+S suggests the following list of both well-known and perhaps less obvious changes associated with AI that have implications across society and the economy, and across our legal and regulatory regimes.

1. **AI detects and extrapolates from patterns in existing data** to make predictions about new data – eg about the next word in a sequence, as in predictive text, or about the cost

[11] Aspects of both the technology and its socio-technical context have precursors; some current harms and problems are as much about long term trends in digitisation, data and data linkage, large scale analytics, and automation, as well as public and private sector drivers for efficiency and personalisation as they are about 'AI' as defined. Whether, and to what extent, AI (and in some cases, ADM) creates new problems or merely reveals existing gaps in the law is the subject of scholarly debate.

of a home insurance premium in a particular location. As has been repeatedly shown, existing human biases and structural inequities are embedded in the data used to train models, and models tend to perpetuate those biases and reinforce those inequities.

2. **Increased capacity for content generation and distribution:** especially with new generations of AI technology (including generative AI), individuals and organisations have increasing capacity to generate and distribute new content – including sophisticated visual and audiovisual content - easily and at scale. While offering many opportunities for creative expression and productivity, these developments may disrupt and/or transform the professions of people who make a living from content (such as copyright owners and creative industries). They may also change the type and increase the amount of harmful material available online (eg, child sexual abuse material (CSAM), as well as to present challenges around the authenticity and/or quality of information circulating in the news and political communication environment.[12] The ongoing blurring of the division between human-created and AI-generated content will in many cases make it more challenging to identify responsible actors and address real world harms.[13]

3. **Opaque,[14] unpredictable[15] technology[16]** creates challenges for predicting possible sources of error and potentials for harm, raising important questions about when it is acceptable to release technology that may not be fully predictable in its effects; how stakeholders can be considered to make informed decisions, when full information is lacking; who is to bear losses caused by unpredicted effects; and when principles such as the precautionary principle should be considered.[17]

4. **New challenges predicting harms, detecting breaches and allocating responsibility**: in the case of various forms of AI, particularly foundation models, developers may not have direct knowledge of the context of use, or a contractual or duty-based relationship with people affected by downstream applications. But they may be the only people in the value chain who are able to fix certain kinds of problems, like bias for example (to the extent that such problems are fixable at source, given the unpredictability and opacity of the models). The capacity for AI to act in unintended and unprogrammed ways may also create challenges for the allocation of responsibility.

5. **Technology that feels more human or is unable to be distinguished from humans**. This has been identified as a reason why the use of autonomous systems should always be

---

[12] Old cues that enabled people to judge the authoritative nature of information (such as poor-quality photoshopping) are disappearing.
[13] For example, when some CSAM is AI-generated, it will become more challenging to identify real children at risk.
[14] For example, unlike conventional IT/data-driven systems, AI outputs and predictions are variable, and low probability predictions may not repeat. One time out of one hundred, a health diagnosis, legal outcome, social service, or migration decision could be different from others, *with no obvious rationale* (other than; it's in the training data). This introduces a very different dimension to ADM systems from, eg, Robodebt or typical rule-based systems.
[15] Unpredictability may be increasing. For example, an important aspect to GPT's success is 'Reinforcement Learning via Human Feedback', which moderates model behaviour to 'align' with specific (and culturally-specific) ethical principles and values. This process embeds thousands of micro-decisions into the model in ways that become difficult for model developers to identify, diagnose and mitigate, or for independent observers to analyse.
[16] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3(1) *Big Data & Society*.
[17] This issue is certainly not new. We have had many technologies with the potential for unpredictable effects and impacts that have been developed and deployed in some way or another: nuclear technology; genetic editing; and cloning. Note that in various of these cases, there have been limits on the actors entitled to deploy the technology, and limits on use. For the most part, these technologies have not been made available to anyone on the planet with a computer. We note recent undertakings by technology companies to submit models for independent testing prior to release. See, eg, White House, *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (21 July 2023).

disclosed to their end users or subjects,[18] but the implications are broader, because even when people are aware they are interacting with AI rather than a person, systems that enable more 'human' styles of interaction may cause people to behave differently as a result of anthropomorphising:

a. Chatbot users may reveal more personal information than they would when interacting with a less 'human-seeming' technology (raising privacy issues and undermining consent-based models);

b. Users may trust systems more readily: requiring, for example, us to reform (i) regulatory frameworks controlling professional advice (eg, medical, financial, legal or other), and the lines we draw between 'information' and 'advice' (ii) other contexts involving interaction with vulnerable populations, such as education);

c. Users may become emotionally vulnerable or attached, raising serious potential for consumer harm. Chatbots based on NLP have been successful at modulating human emotion.[19] This requires careful management/regulation, for example in relation to automated therapies in mental health settings.[20] The capacity to iteratively experiment with auto-generated content that modulates human emotion can enable the creation of more effective targeted content, intended to modulate the emotions of a person or people, to benefit some other person or firm (eg, more effective targeted advertising or propaganda). Legal issues that remain unclear for now include, for example, the question when such technology would cross the line of unfair commercial practice under consumer protection law.

6. **New ways of acquiring and interacting with information**:[21] Depending on the interface, generative AI based on LLMs enables 'conversationally based' interactions: instead of 'Googling', we may prompt a knowledgeable model, with returned information at risk of ranking, manipulation and obfuscation, but provided without the cues people use to judge the quality of information: such as referencing, sources, and lists of alternative sources.[22] This could have undermine the effectiveness of legal regimes, such as consumer protection law, which rely heavily on providing information to people in order to achieve policy goals.

The upshot of these shifts

These changes highlight **gaps in existing laws** and **gaps in enforcement**, as well as introducing **new capacities and consequences** - addressing these gaps and the consequences of these new capacities requires a society-level conversation. The capacity of AI to cause harm on a scale and at a speed not previously possible is a strong argument in favour of **regulation ex ante:** requiring mitigation of risks at the design and development stage of AI. A risk-based

[18] See, Toby Walsh, 'Turing's red flag' (2016) 59 *Common.ACM* 34-7; Frank Pasquale, *New Laws of Robotics* (Belknap Press, 2020); Australian Human Rights Commission, *Human Rights and Technology* (Final report, March 2021) 60-7.
[19] See, eg, Samantha Delouya, 'Replika Users Say They Fell in Love with Their AI Chatbots, until a Software Update Made Them Seem Less Human', *Yahoo Finance* (23 March 2023); Laura Weidinger et al, 'Ethical and Social Risks of Harm from Language Models' (2021) arXiv:2112.04359.
[20] Eva Weber-Guskar, 'How to Feel about Emotionalized Artificial Intelligence? When Robot Pets, Holograms, and Chatbots Become Affective Partners' (2021) 23 *Ethic and Information Technology* 601-10.
[21] ADM+S will host a symposium on search: *The Web Search Revolution: The Past, Present, and Future of Web Search – Google, ChatGPT, Bing, and Beyond* on 17 August 2023.
[22] Compare the difference between the ways searches will be framed, and results presented, in 'search via traditional Google' and 'search via chatbot'.

approach is discussed below in relation to Consultation Questions 14-16.[23] However, when we adopt a risk-based approach, we are requiring firms to identify, and mitigate, certain risks of harm. But for this to work, **there must be some kind of 'or what'?** For this to lead to genuine improvement in the technologies applied, there must be a risk of consequences - *liability for harms caused if organisations fail to take mitigating action.* That means laws prescribing the act that creates the risk of harm, and a credible threat that that law can be enforced. In other words, the success of any risk-based framework in Australia will depend on addressing current gaps in both our rules (ie laws/legal frameworks) and our enforcement capacities. Other countries that are considering risk-based approaches are simultaneously addressing or have already addressed these issues.

Gaps in Australia's existing laws

When content generation is automated, and when people interact with information differently, the assumptions underlying many of our existing laws come into question.

This submission is not the place to discuss in detail reforms that will be required to existing legislation/laws. Nevertheless, to illustrate the kinds of gaps we see emerging, **Table 1** below makes preliminary comments regarding how the various shifts summarised above may 'change the context' in and against which various laws operate, requiring consideration for reform, updating, and/or clarification through guidelines over time (noting that many of these relate to generative AI). ADM+S researchers are already engaging with specific reform processes and will continue to do so.

Table 1: Indicative table illustrating some challenges for existing legal regimes.

Domain	**Challenges**
Harmful online content regulation	• Given increased capacity for content generation and distribution, *ex post* takedown systems/systems based on content flagging and focused on outputs are inadequate. Pre-emptive/positive obligations (like the Basic Online Safety Expectations[24]) become more important. • Given Generative AI, which regulations around online content depend on intent or awareness, and how does that apply to automated content generation or automated dissemination? Can an automated system bully or harass? Defame? Target a vulnerable person?
Consumer Protection	• How do new methods of content presentation affect what is 'misleading' or 'deceptive', and/or what is 'fair'? When is manipulation of human vulnerabilities and/or influencing human emotional states 'unfair'? • Do we need a positive duty requiring fair dealing with consumers? • How do consumers shop around/compare deals/assess claims made about products when prices, ads and product descriptions are automatically generated and change every time you look? • What should be required of online platforms in terms of maintaining a database of online advertising and the way it is adjusted and targeted using automation and providing access to regulators for monitoring and enforcement purposes?

[23] See below, *Consultation question 14: is a risk-based approach the right one?; Consultation question 15: addressing some limitations of a risk-based approach;* and *Consultation question 17: elements of a risk-based approach.*
[24] eSafety Commissioner, 'Basic Online Safety Expectations', *eSafety* (Web Page).

Administrative Law	• Can an automated system make a 'decision'? • Who is the 'decision maker' when an AI is involved? • How can automated administrative decisions be challenged and remedied at a systemic level?
Discrimination	• The challenge of applying Australia's anti-discrimination law to AI-generated harms was identified by the Australian Human Rights Commission (AHRC) as an issue and remains outstanding.[25] • Detecting discrimination is a significant challenge where content is ephemeral and personalised: is a positive duty to deploy only non-discriminatory systems required? • How to address AI's potential to discriminate against/disadvantage groups beyond those protected by Australia's anti-discrimination law (such as people already suffering socio-economic disadvantage)?
Copyright	• Who owns the original outputs generated by AI, especially by generative AI tools (such as ChatGPT)? • Who is liable if AI-generated outputs breach the law: the AI developer, AI user, or both to blame?
Data protection law	• Are conversations more revealing than other modes of information collection? • What are the privacy implications of personal or confidential information being included in prompts/context windows of LLM-based chatbots or search engines? • Advanced and widespread capabilities to simulate images and voices of individuals using generative AI systems pose privacy risks (among other things) by increasing the risk of misuse of one's likeness. • To what extent does privacy legislation impose restrictions on the use and disclosure of synthetic voice and image?[26] How would proposed reforms to the *Privacy Act 1988* (Cth), such as the fairness and reasonableness test,[27] operate in relation to such practices?
Cybersecurity	• Cybersecurity risks, via human factor manipulation (what will people tell a chatbot?) and covert (prompt injection risks)
Professional regulation (eg law, medicine, financial advice)	• Is personalised, chatbot information regulated as legal/medical/financial *advice* or *information*? Where's the line? Do definitions need to change if people are likely to *treat* outputs as advice because of how the information is presented? • How do we apply the fiduciary and tortious duties of advisors to circumstances where AI is being used to provide or support the information and advice given?
Political advertising and campaign laws	• In a context where it is easier to produce and rapidly disseminate deepfakes, do we need to strengthen our laws regarding misleading political communications? Watermarking or otherwise indicating AI generated content will likely be an important aspect of any legal response (see below). Is it time to also consider whether and how we prohibit political actors from

[25] Australian Human Rights Commission, *Human Rights and Technology* (Final report, March 2021).
[26] We note that this question has to some extent been engaged with in the course of the Privacy Act review, with certain changes to clarify the operation of the Act in relation to 'inferred' and 'generated' information proposed in the recent Privacy Act Review Report: Attorney General's Department, *Privacy Act Review* (Report, 2022).
[27] Ibid 110-21.

	producing and disseminating misleading and manipulative media content including but not limited to the context of election campaigns? ● Collective scrutiny is a challenge where such content is disseminated in personalised newsfeeds across digital platforms. Do we need to consider new record keeping and transparency requirements for digital political advertising,[28] as part of a combination of measures to address these challenges?
Negligence and fault based liability	● When harmful AI outputs result from multiple inputs from actors at various points in the value chain, how do ideas of fault and responsibility work? ● Are novel categories of duty of care required? Is the law likely to recognise duties applying to foundation model providers who don't have proximate relationships with users of AI or people affected by AI?[29] Might statutory duties for various actors in the AI value chain be appropriate? ● Are forensic challenges around identifying causes of harms, and proving breach of duty surmountable? ● Do we need to expand the concept of a 'manufacturer' for the purposes of product liability? At present, consumer guarantees seem likely to apply to downstream app providers, but not upstream foundation model providers.

Regulatory gaps at the training stage

Another set of issues relates to the **training** of AI models. There is considerable uncertainty regarding the position on data-scraping for training purposes - under Australian law and beyond.

AI requires significant inputs of data. Foundational models in particular involve the use of large and composite data sets – sometimes collected, curated and published by third party organisations, such as the Common Crawl text dataset used in ChatGPT, and the LAION image data sets used in Stable Diffusion and other text-to-image models.

Current Australian copyright laws restrict the use of copyright-protected content (text, images, videos, etc) for AI training purposes; as a matter of current law, AI developers should generally get permission and pay a licensing fee when they copy/reproduce copyright-protected content in AI training contexts.[30] Australia does not have a specific exception to cover the use of content as training data, and the existing copyright exceptions (fair dealing, or temporary reproduction exceptions) are unlikely to apply.

[28] A number of large digital platforms already maintain political advertising archives. However, the voluntary nature of these archives leaves the content, availability and accessibility of important public transparency resources to the discretion of platform companies. Numerous governments and governmental agencies in different jurisdictions have introduced proposals which may serve as models for reform. For example, the EU's proposed regulation on transparency and targeting of political advertising, in combination with the *Digital Services Act*, set out a comprehensive framework for record-keeping and maintenance of advertising repositories which imposes obligations on key actors in the advertising supply chain: European Commission, *Proposal for a Regulation of the European Parliament and of the Council on the Transparency and Targeting of Political Advertising*, COM/2021/731, 25 November 2021, arts. 6, 7, 10-12; *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)* [2022] OJ L 277/1.

[29] Arguably neither consumer law nor negligence imposes duties on them. In negligence it will be hard to show foreseeability, neighbourhood / proximity, control and other duty requirements.

[30] See Rita Matulionyte, 'Australian copyright law impedes the development of Artificial Intelligence: What are the options?' (2021) 52(4) *International Review for Intellectual Property and Competition Law -ICC* 417-443; for a discussion on whether the use of content in AI training could infringe moral rights of authors see Rita Matulionyte, 'Can AI infringe moral rights of authors and should we do anything about it? An Australian perspective', (2003) 15(1) *Law, Innovation and Technology* 124-147.

Comparable jurisdictions have broader copyright exceptions, such as fair use (US), or text and data mining exceptions (UK, EU, Japan,[31] Singapore[32]), which allow free use of content for AI training purposes at least in certain contexts (eg for non-commercial purposes), albeit with a small number of relevant legal actions pending in the US and UK.[33] If the Australian Government wants to encourage research into AI technologies, and the development of AI technologies locally, they will need to take into account the international practices in the field, and adjust local copyright laws accordingly.[34] To the extent that there is a need or desire to promote the interests of creators, it will be necessary to consider whether copyright is the right model to ensure creators' interests in this new environment.[35]

Similar issues exist with respect to the use of text, images, sound and other forms of data that are not protected by copyright. Australian privacy laws offer highly individualised and very limited protection against the aggregation of data to train and benchmark AI models. And there are no other explicit prohibitions on data scraping *per se*. Some State jurisdictions have offences for accessing computer systems without authority (similar to the provisions invoked against web scraping in the US Computer Fraud and Abuse Act), although no jurisprudence has emerged considering their application to scraping.

Open internet ideologies promoted for decades by large technology companies have been built into the legal status quo. Web scraping is generally either legally permissible even if in violation of contractual provisions (ie Terms of Service) and/or other laws,[36] or in Australia's case, apparently tolerated despite our failure to amend copyright law for many years. With the emergence of very large 'foundation models', and emerging frontiers of value creation associated with AI, this status quo may need to change. Lawsuits in other jurisdictions have been initiated on various bases, including breach of biometric privacy rules, unjust enrichment, and breach of rights to personality, amongst others.[37] These are highly contextual, but should prompt reflection in Australia as to whether more considered approaches to data scraping (as well as alternative models of data governance, such as contextual or sectoral data trusts) ought to be part of a regulatory strategy.

At the same time, data scraping and accessibility for the purposes of independent, public interest research may need additional safeguarding, especially when platforms are themselves implementing increasingly aggressive countermeasures to prevent user data from being scraped for machine learning purposes.

---

[31] Japan's Copyright Act was amended in 2019 to allow all users to analyse and understand copyrighted works for machine learning (Article 30-4); permit electronic incidental copies of works, recognising that this process is necessary to carry out machine learning activities but does not harm copyright owners; (Article 47-4) and allow the use of copyrighted works for data verification when conducting research, recognising that such use is important to researchers and is not detrimental to rights holders (Article 47-5).

[32] *Copyright Act 2021* (Singapore) art 243-4.

[33] For example, a class action lawsuit was filed in November 2022 in the US District Court over the GitHub AI software Copilot, which can speed up coding, suggesting multi-line code completions based on user prompts. Copilot was trained on publicly available code in GitHub repositories. A group of visual artists filed a lawsuit on 13 January 2023 in the US District Court in San Francisco against Stability AI, Midjourney and DeviantArt over their respective generative AI products based on Stable Diffusion that creates images based on a user's prompt.

[34] Sean Flynn et al, 'Legal reform to enhance global text and data mining research' (2022) 378(6623) *Science* 951-3.

[35] For a discussion of this issue, see Kimberlee Weatherall, 'IP and data, IP in data, IP as data', in Damian Clifford, Jeannie Paterson and Kwan Ho Lau (Eds.), *Data Rights and Private Law* (Hart, forthcoming 2023).

[36] See, eg, Han-Wei Liu, 'Two Decades of Laws and Practice Around Screen Scraping in the Common Law World and Its Open Banking Watershed Moment' (2020) 30 *Washington International Law Journal* 28.

[37] For example, unjust enrichment was asserted as both a cause of action and a remedy in a class action challenging the lawfulness of the operation of GitHub Copilot, filed in November 2022.

Enforcement

Another impact of the shifts set out above is that they both **expose existing enforcement gaps** in Australia's legal framework - and create challenges in terms of practical litigation.

Existing enforcement gaps

AI creates well-known risks for harms to individuals: breaches of their rights not to be subject to discrimination; breaches of privacy; and consumer protection infringements. Without reform to existing enforcement regimes, breaches of rights (including human rights) resulting from AI use are less likely to be the subject of any enforcement or remedy, either because:

- there is no remedy at law; or
- there is, but only via an indirect route requiring dedicated and sophisticated lawyers; or
- there is, but only via a regulator unable to investigate every potential breach brought to their attention; or
- Remedy waits until there is a major *ex post* investigation (such as a Royal Commission): victims wait years for recognition or remedy, by which time more lives have been unnecessarily blighted or opportunities lost.

One area where enforcement is particularly lacking is fundamental human rights. On the face of it, mechanisms for enforcement of human rights in Australia are either missing, or indirect. Australia has no comprehensive constitutional or legislated human rights instrument at a federal level (and only some States and the ACT have relevant legislation).[38] This creates a *potential* gap in Australians' rights to a remedy for human rights impacts of AI. However, it is wrong to say those rights are absent, and Australian lawyers and legal academics are identifying various causes of action in order to enforce rights more or less directly.[39]

In addition, where human rights *have* been legislated at a federal level - via anti-discrimination law and (to some extent) data protection (privacy) legislation, **enforcement is heavily constrained**[40] by requiring complainants to first approach the relevant regulator: the Office of the Australian Information Commissioner (OAIC)[41] or AHRC.[42] This can mean enforcement will

[38] The Australian Capital Territory (*Human Rights Act 2004* (ACT)), Victoria (*Charter of Human Rights and Responsibilities Act 2006 (Vic)*) and Queensland (*Human Rights Act 2019 (Qld)*) are the only states with a human rights statute.

[39] Examples of lawyers acting to enforce rights include the Robodebt class action and current class actions against Optus in relation to data loss; the recent Royal Commission into the Robodebt Scheme (Robodebt Royal Commission) suggested the potential for actions for misfeasance in public office. On the academic side, Dr Henry Fraser (QUT, ADM+S) has been examining the potential of the law of tort to provide remedies, as well as exploring human-centred ways to conceptualise fault and responsibility. See, eg, Henry Fraser, 'Legal issues around autonomous systems: Civil liability, fault and system safety'. In *IEEE International Symposium on Technology and Safety, 2022-11-10 - 2022-11-12*, Hong Kong, China; Henry Fraser, 'AI Safety doesn't make AI safe', *AI Learning Curve Blog* (Blog, 2023). The point is that direct routes to legal action are largely absent, and the need for the development of jurisprudence creates additional barriers to access to justice for most people experiencing human rights breaches caused by technology.

[40] Approaches at the State level vary. At the federal level and in States with no human rights legislation, breaches of rights other than those explicitly legislated do not have a legal remedy. In Victoria, individuals can complain to the Victorian Ombudsman about complaints about breaches of human rights by public sector organisations. In the ACT, the ACT Human Rights Commissioner can only investigate certain types of complaints and "does not investigate individual complaints about breaches of the Human Rights Act": ACT Human Rights Commission, 'Enforcing Human Rights' (Web Page).

[41] Note that there is no direct right to action under the Privacy Act (although one has been proposed by the recent Privacy Act review). Individuals seeking redress will have to file a complaint to OAIC who will generally require the individual to first file a complaint with the relevant organisation. The OAIC may also decline to investigate or further investigate a complaint if there is no reasonable likelihood of a conciliated outcome as per s 40A(4) of the *Privacy Act 1988* (Cth).

[42] Notably, the AHRC is not empowered to initiate own motion actions, or litigate claims of breaches of human rights in the public interest.

reflect priorities of regulators and officeholders and be constrained by their capacity to act in terms of mandates and resources. Complaints that are rejected - or delayed - by these actors have limited options.

Regulators including the OAIC and AHRC also focus on resolving individual complaints through mediation. This is pragmatic, but even if an individuals' immediate problem is resolved, this approach:

- prevents the development of a jurisprudence to enable others harmed to bring action;
- hinders our understanding of harmful systems, and leaves systemic issues unaddressed (as highlighted by the Robodebt Royal Commission);[43]
- where many individuals suffer small harms - as can happen when AI systems scale - then *individual complaints* may be rejected as immaterial or non-compensable, leaving the broader systemic harm unaddressed.

Questions of **proof** are also very challenging. For example, a feature of AI is its variability of output - eg, it is possible that a well-intentioned AI system may be accurate and reliable under testing, but respond in an inaccurate and harmful way just once (a low probability but high impact event), making it very difficult to reproduce and therefore for a complainant to prove the event occurred. Firms deploying AI systems will be aware of these limitations on the likelihood of legal liability, meaning their incentives to manage smaller, occasional or widely dispersed harms in particular are reduced.

An absence of direct remedies - and the resulting impacts on systemic issues and firm incentives - applies across the board in Australia's legal regime. Australia's commercial law also makes extensive use of (more or less) voluntary industry codes of conduct[44] with varying levels of regulator involvement and a range of mechanisms for complaint.[45] This distinguishes Australia from other jurisdictions considered in the Discussion Paper, which have more developed and sophisticated peripheral regimes. For instance, the EU AI Act proposal refers to Europe's GDPR regime, which affords more stringent protections against unauthorised data collection and use compared to the Australian context.

In summary

Australia needs systematically to consider how it is going to ensure enforcement of (both existing, and any new) standards for (automated and) AI systems, including both addressing systemic issues, and providing remedies for harm where appropriate.[46] This means:

- identifying what harms, which may be caused or exacerbated by AI, remain largely without remedy under Australian law; *and*
- assessing the current system of seeking to funnel most complaints through mediation, and most enforcement through underfunded and overworked regulators.

To address both the need for individual remedies, and systemic issues, Australia will need to retain an important role for regulators (both the ability to receive complaints from individuals,

---

[43] *Royal Commission into the Robodebt Scheme* (Final Report, July 2023) ('Robodebt Royal Commission').

[44] See, eg, the *Telecommunications Code of Practice; General Insurance Code of Practice; Banking Code of Practice;* and *Energy Retail Code of Practice.*

[45] Some industries have an ombudsman, eg, Telecommunications Ombudsman; Financial Services Ombudsman; energy ombudsman scheme (with separate ombudspersons in different geographical areas). Alleged breaches of the Banking Code of Practice are investigated by an independent body, the Banking Code Compliance Committee (BCCC).

[46] One approach could be to allow for recovery of damages in the event that proper assessment of the risk of harm, or proper mitigation is not undertaken, and one or more individuals are harmed. Proving, however, that the harm was caused by the failure (to undertake proper risk assessment or mitigation) could be challenging for an individual: there would need to be, at least a reversal of the onus of proof or rebuttable presumption, and perhaps in some cases a non-rebuttable presumption of causation for such a legal mechanism to be effective in enabling remedies for individuals.

and the power to undertake own motion investigations). Regulators do have a unique capacity to collate issues, monitor trends and identify priority areas for action or enforcement. But we will *also* need:

- a wider range of enforcement pathways: class actions; individual rights of action; and extended standing that could enable competitors and/or advocacy organisations to bring actions;[47]
- in relation to systemic issues, oversight mechanisms directed at both upstream design, and ongoing monitoring (discussed below).

Expanding enforcement this way does not imply a litigation free-for-all: litigation will always be hard, and expensive, and likely rare. But their *potential* existence can have an impact on incentives, changing the calculus for all firms in the market.

New capacities requiring a public conversation

Another important implication arising from the shifts we have identified is that new capacities and new directions in the relationships between people on the one hand, and public and private institutions on the other. A combination of large-scale data and data collection, automation, and AI systems are making possible analyses and uses not previously feasible, as well as more automation. A societal-level conversation about acceptability and limits is needed in all of these cases.

1. **Automated, remote mass surveillance** and facial/biometric recognition (by public and private sector actors): Concerns exist about inaccurate and biased data that disproportionately impact marginalised populations.[48] Even accurate technology which functions as intended raises ethical issues around: the prospect of living in a society with 'perfect surveillance'; normalisation of surveillance; and how to guard against invasive, unethical and undemocratic surveillance. In particular, facial recognition is seen as uniquely concerning compared to other methods of biometric surveillance: more information can be extracted from the face than from other forms of biometrics such as fingerprints, the face can be surveilled remotely without requiring active consent or physical interaction, and particularly broad legal lacunae exist in relation to facial recognition.[49]

2. Beyond facial recognition, emerging AI systems will exploit the mass **spatio-temporal data** from the proliferation of **Internet of Things**, enabling environmental and behavioural patterns to be observed at individual homes, buildings, and precincts at scale. While posing risks to privacy if data is collected and linked at an individual or granular level, if used responsibly, these data streams can be used by AI systems to improve the sustainable operations of our cities, towns, and farms: privacy-respecting sensor data streams take the form of aggregated time-series data (eg water flow, humidity level, traffic volume, energy use), gathered at spaced time intervals (eg every 15 minutes) without personally identifying information.

3. **'Social scoring**': linking data across domains for analysis and action (within the private and public sectors) and creating the potential for a person's history and behaviour in one field to impact their opportunities and activities in unrelated domains to an extent never before possible. This is an area that may become extremely tempting to both

[47] See, eg, section 487 of *Environment Protection and Biodiversity Conservation Act 1999* (Cth), which extends standing to environmental and conservation groups in judicial review proceedings under that legislation.
[48] Nicholas Davis, Lauren Perry and Edward Santow, Human Technology Institute, The University of Technology Sydney, *Facial Recognition Technology: Towards a Model Law* (Report, September 2022).
[49] See, eg, Evan Sellinger and Brenda Leong, 'The Ethics of Facial Recognition Technology' in Carissa Véliz (ed), *The Oxford Handbook of Digital Ethics* (Oxford University Press, 2021).

commercial and government actors and one that at the same time promises particularly acute social harm.[50]

4. **Invasion of mental privacy/autonomy:** contra much industry hype, we may not yet be at the stage of 'mind-reading robots' or 'AI-facilitated behavioural manipulation', but there are strong incentives across both the public and private sector to develop technology to identify, predict, and influence individuals' mental states and behavioural intentions: for benign reasons,[51] bad reasons,[52] and reasons that while not inherently 'good' or 'bad', might advantage some groups and actors in society over others.[53]

5. **The broader societal impacts of automation:** more (public and commercial) services are becoming automated, and given the imperative for **efficiency** across both public and private sectors, the impulse to automate as much as possible will be hard to resist. It is critical to understand and take account of the impact of such automation on people, especially (but not exclusively) vulnerable and/or isolated groups of people: people without constant access to reliable and affordable internet (as identified in ADM+S' Digital Inclusion work); [54] people unable, because of disability or for other reasons, to engage with digital modalities; and people with limited social contacts for whom interactions with service providers represent necessary human contact. While this might read as a policy issue rather than a regulatory/governance issue, internationally there are laws that require service providers to provide non-automated alternatives.[55]

We argue below[56] that a broader, *public* conversation is needed about AI generally – one that actively engages communities and the not-for-profit sector as well as industry stakeholders – and here we argue that these specific new capacities we identify require such a consultation. DISER, as the department most connected to science and industry is well-placed to lead a public discussion about these emerging capacities. As a nationally funded and internationally connected, multidisciplinary research centre with partnerships spanning key industry, government and community sectors, the ADM+S is ideally positioned to help broker such a conversation.

---

[50] 'Social scoring' tends to evoke images of the (reported) Chinese Social Credit system, but there is broad potential for a data about a wide range of a person's activities to impact on educational opportunities (through scholarship or admission analyses), job opportunities (via CV-sorting), or even the ability to participate in society via access to semi-public spaces. Concerns about social exclusion extending beyond single venues are raised by systems like Auror which potentially record anti-social behaviour and share that information across stores, as investigated by Crikey: Cam Wilson, 'Crime Tech Used in Woolworths, Coles, Bunnings Raises Concerns', *Crikey* (Web Page, 22 February 2023), or facial recognition at stadiums, recently investigated by Choice: Jarni Blakkarly, 'Facial Recognition in Use at Major Australian Stadiums | CHOICE', *Choice* (Web Page, 5 July 2023).

[51] For example, to develop robots that can respond automatically and directly to the thoughts of people with disabilities.

[52] Such as extremist recruitment.

[53] Technology designed to predict - and nudge or change - purchase intentions. Advertising has long sought to influence purchase intentions, but any improvements in such technology shift the balance in information and power between sellers and buyers in ways that undermine rationality of consumers and the operation of open competitive markets. See, eg, Tegan Cohen, 'Regulating Manipulative Artificial Intelligence' (2023) 20(1) *SCRIPTed* 203.

[54] See below, *Risk assessment must situate responsible AI in its social contexts.*

[55] For example, article 22 of the *GDPR* grants individuals a right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning that individual or similarly significantly affects the individual: *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)* [2016] OJ L 119/1, art. 22.

[56] See below, *Consultation question 14: is a risk-based approach the right one?;* and *Consultation question 15: addressing some limitations of a risk-based approach.*

The gaps around environmental impacts

To foster the development of Safe and Responsible AI, strong and active government engagement with citizen-facing advocacy groups, professional bodies, industry groups and worker groups impacted by AI and or developing AI technologies is strongly encouraged. There is growing attention and concern from both the public and tech industry professionals to the direct and indirect environmental impacts of the training and use of machine learning for AI.[57] As a result, leading approaches to safe and responsible AI now acknowledge that regard must be had to environmental impacts in creating governance frameworks for AI. For example the OECD's principles for responsible AI now include recognition that both positive and negative impacts of AI on the environment need to be considered.[58] The current draft of the EU AI Act, approved by the European parliament, also foregrounds being environmental protection friendly as a fundamental value for AI regulation, and requires environmental impact assessment and environmental reporting for high risk uses of AI and large language models and other generative AI (discussed further below).[59]

The environmental impacts of AI are similar *in kind* to the environmental impacts created by the use of computing and data centres in general (for cloud storage, intensive computation, data analysis, software, mass digital media streaming and digital platform operations).[60] Any scaling up of machine learning (ML) in terms of size and complexity of models developed, trained and adopted, and the frequency of use and speed of development significantly heighten the demand for and use of computing for the purposes of AI.[61] Moreover, these impacts will become more critical with increasing global heating, more acute pressure over the extraction and supply of silicon and critical minerals, and growing expectations from consumers, the public and trade partners that Australia move to a fully circular green economy.

The environmental impacts of AI training and use include:

- Energy consumption - including greater demand and competition for energy from both fossil fuels and renewable energy. In a survey of the carbon emissions of 95 ML models across time and performing different tasks in natural language processing and computer vision, researchers have found that the majority of their models (61) used high-carbon energy sources such as coal and natural gas as their primary energy source, whereas less than a quarter of the models (34) used low-carbon energy sources like hydroelectricity and nuclear energy. They also show that the main sources of variance in the amount of emissions associated to training machine learning models is due to the carbon intensity of the primary energy source and the training time, with the power consumption of the hardware (eg GPUs) having less influence;[62]

---

[57] Roel Dobbe and Meredith Whittaker, 'AI and Climate Change: How they're connected, and what we can do about it' (2019) *AI Now Institute*, Medium; Roy Schwartz et al, 'Green AI.' (2020) 12 *Communications of the ACM 63* 54–63; AlgorithmWatch, 'Digging Deeper: AI's Environmental Report Card. Does Artificial Intelligence Consume More Resources than it Conserves?' (2023) *SustAIn magazine, Issue #2*.

[58] Kaith Streier et al, 'Can AI help save the planet?' *OECD.AI Policy Observatory*, (Web Page, 17 November 2022); OECD, *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint* (Report, 2022).

[59] European Parliament, 'EU AI Act: first regulation on artificial intelligence' (Media Release, 14 June 2023). This is partly adopted recommendations by ADM+S partner organisation AlgorithmWatch: see, AlgorithmWatch, *Ensure minimum transparency on the ecological sustainability parameters for all AI systems in the AI Act*, (Issue Paper, April 2022).

[60] Jesse Dodge et al, 'Measuring the Carbon Intensity of AI in Cloud Instances.' In 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 1877–94.

[61] Alexandra Sasha Luccioni, 'The mounting human and environmental costs of generative AI', *Ars Technica* (Op-ed, 12 April 2023).

[62] Alexandra Sascha Luccioni and Alex Hernandez-Garcia 'Counting carbon: A survey of factors influencing the emissions of machine learning' (2023) arXiv:2302.08476.

- Water use - for cooling computing facilities;[63]
- Raw material use - the impact of building, maintaining and using the material infrastructures associated with computing equipment (ie GPUs) - including the extraction and refining of rare metals;[64]
- Land use - including siting and building the facilities, creating new energy infrastructures including solar arrays, etc;[65]
- Undersea cables - the placement and maintenance of cables, including undersea cables and so on.[66]

Because of the growing understanding of the various environmental impacts of AI across its whole supply chain, professional bodies and NGOs are advocating for life cycle analyses of AI equipment and software that would consider the extraction of raw materials to produce them, their manufacturing, transport, use and end of life.[67] For example, through their SustAIn project, the Berlin-based NGO Algorithm Watch (an ADM+S partner in the ADM+S Centre) is working with academic researchers and industry to conduct whole life cycle analysis of the environmental impact of AI, including not only development, training, inference, application, but everything from critical minerals mining to hardware manufacture to carbon emissions of data centres to end of life e-waste.[68] The Green Software Foundation — an industry consortium led by tech companies and world-wide consulting firms — is working toward the creation and implementation of a software carbon intensity specification describing how to calculate the carbon intensity of a software application, including AI models.[69]

Current Australian law places few restrictions on AI and App developers to take account of the ecological impact of their activities. It is likely that the growing environmental impact of AI will create the need for both review of existing legal and policy frameworks (horizontal approaches) and domain specific regulation:

- Horizontal: the uptake of AI, including creation of new data centres will lead to the need to update and review existing Australian legal and policy frameworks that apply to issues such as the siting of facilities such as data centres and undersea cables, product stewardship and e-waste, energy grids including the creation and use of renewable energy facilities.
- Domain specific: New frameworks for governing AI in Australia could follow the example of the current proposed EU AI Act (as of 14 June 2023) which includes in its purposes and principles to ensure the uptake of AI is consistent with a high level of protection for the environment (Recital 1, Article 4a).[70] The proposed requirements in the draft Act include risk assessment of potential environmental harm (Article 9.2a) and an obligation on AI developers to provide logging capability within AI systems to record energy consumption, measure resource use and environmental impact (Article

[63] Pengfei Li et al, 'Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models' (2023) arXiv: 2304.03271; Mel Hogan, 'Data flows and water woes: The Utah Data Center' (2015) 2(2) *Big Data & Society*.
[64] Ludovico Rella, 'Close to the metal: Towards a material political economy of the epistemology of computation' (2023) *Social Studies of Science*.
[65] Mel Hogan and A Vonderau, 'The nature of data centers' (2019) *Culture Machine*.
[66] Nicole Starosielski, *The Undersea Network* (Duke University Press, 2015).
[67] Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz, 'Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions' (2022) 14 (9) *Sustainability* 5172; Aimee van Wynsberghe, 'Sustainable AI: AI for Sustainability and the Sustainability of AI' (2021) 1(3) *AI and Ethics* 213–18; Lynn H. Kaack et al, 'Aligning Artificial Intelligence with Climate Change Mitigation.' (2022) 12(6) *Nature Climate Change* 518–27.
[68] 'SustAIn: The Sustainability Index for Artificial Intelligence', *AlgorithmWatch* (Web Page).
[69] *Green Software Foundation* (Web Page).
[70] European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the 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 (COM(2021) 206 final)* ('proposed EU AI Act'). Recital 1, Article 4a. See also initial proposal text.

12.2a). Providers of foundation models would also be required to draw on applicable standards to reduce energy, resource use and waste and to increase overall efficiency (Article 28b). The draft EU AI Act also explicitly encourages the development of codes of conduct (under Article 69(1)) including by individual providers of AI systems, any interested stakeholders and their representative organisations (Article 69 (3)).

There is a critical need for greater research including industry, civil society and academic researchers to better understand and govern the ecological impacts of AI in light of environmental justice perspectives.[71] ADM+S researchers are working on the ecological implications of AI, and how it is and could be governed. Chief Investigator Prof Christine Parker, Affiliate Hon Prof Fiona Haines and Research Fellow Dr Loup Cellard are working on the environmental governance of AI and the governance of data centres and undersea cables.[72] Chief Investigator Prof Sarah Pink was recently awarded an Australian Laureate Fellowship to examine the twin digital and green transitions[73] which will further her work and the work of colleagues at the Monash University ASDM+S group on sustainable energy futures.[74] The ADM+S will be curating a series of workshops throughout the second half of 2023 to further dialogue and understanding on mapping the environmental impacts of AI and appropriate responses convened by Dr Melissa Gregg.[75]

[71] Bogdana Rakova and Roel Dobbe, 'Algorithms as Social-Ecological-Technological Systems: an Environmental Justice Lens on Algorithmic Audits' (Speech, ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, 2023).
[72] Christine Parker et al, the ARC Centre of Excellence for Automated Decision Making and Society, Submission to Just Transitions in Australia, *Moving Towards Low Carbon Lives Across Policy, Industry and Practice*, (January 2022); Loup Cellard and Christine Parker, 'Will digital tech solve the climate crisis?' (Panel, ADM+S Symposium, 'Automated Societies: What do we need to know?', RMIT, Melbourne, 20-22 July); Loup Cellard and Clement Marquet 'Frictions de l'interconnexion globale : les câbles sous-marins de télécommunication face à la protection de l'environnement' *Revue d'Anthropologie des Connaissances* (forthcoming); Simon Coghlan and Christine Parker, 'Harm to Nonhuman Animals from AI: a Systematic Account and Framework' (2023) 36 *Philos. Technol* 25 see also, Aitor Jiménez, 'The crimes of digital capitalism' (2024) New York University Press (forthcoming).
[73] Loren Dela Cruz, 'Prof Sarah Pink awarded 2023 Australian Laureate Fellowship', *ADM+S*, (Web Page, 3 July 2023).
[74] Yolande Strengers et al, Monash University, *Digital Energy Futures: Future Living Energy Scenarios 2030/2050* (Report, 2023); Kari Dahlgren et al, Monash University, *Digital Energy Futures: Review of Industry Trends, Visions and Scenarios for the Home* (Report, 2020).
[75] 'Mapping the environmental costs of AI: A series curated by Melissa Gregg for ADM+S', *ADM+S* (Web Page).

Consultation question 5: Australia in relation to international developments

Consultation Question 5

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?

ADM+S Answer: Australia can benefit from, and adapt, approaches that are developing in other countries - but must be mindful of key differences in Australia's legal framework which mean that those approaches will fail here without addressing those differences. Australia is well-positioned to cooperate - and should cooperate - with like-minded countries in developing international frameworks for the regulation of AI.

In doing so, Australia should be conscious to:

- not leaving Australians behind in rights and remedies,
- consider domains and issues where there is a role for Australian governments in protecting and promoting uniquely Australian voices and interests;
- continue to develop/invest in Australia's domestic capacity to develop and innovate in AI research and technologies, responsible AI practices and the regulatory frameworks.

In considering whether and how other countries' approaches are relevant, adaptable and desirable for Australia, it is important to consider:

1. What might be similar or different about Australia that might make initiatives overseas more or less relevant to Australia, or raise different issues
2. Whether there are ways that Australia can and should benefit from or integrate with systems overseas.

On the first point, developing EU or Canada-style risk-based approaches require developers and deployers of AI to identify risks of negative outcomes, and avoid or mitigate them. One **difference** between Australia and other jurisdictions developing risk-based approaches that inform the Discussion Paper is the lack or potential lack (in most Australian jurisdictions) of legal mechanisms to challenge harms caused by AI outlined above.[76] As we note further below, adopting only the risk-based approach from overseas without addressing these broader differences will undermine the effectiveness of the system in Australia.

There are other important differences. For example, the EU AI Act and its risk-based approach depends on that region's conformity and assessment infrastructure. It has been designed with the European single market in mind, and promotes the evolution of private risk-assessment certification and assurance in line with the comprehensive network of standards that exists in that jurisdiction. In other words, the EU risk-based approach is effectively a product safety regime certified through networks of private actors ('notified bodies').[77] Australia does not have the same conformity ecosystem nor does it orient its product safety regime around trade and market harmonisation. The EU Risk-Based approach is intended to comprehensively guide the formation of a certification and conformity market and ecosystem. It is unclear whether Australia's approach is intended to have the same effect.[78]

We note also that both **Japan** and **India** are not much discussed in the Discussion Paper despite their economic importance to Australia. **Appendix 1** includes some further information

[76] See *Enforcement.*
[77] Proposed EU AI Act (n 67), art 33.
[78] For more commentary on standards and assurance, see submission in response to the Discussion Paper by Fraser et al.

on those jurisdictions; we have expertise in the Centre should the Department want more information.

Can Australia benefit from integration with legal, regulatory and governance developments overseas?

As to the second question: while Australia is a small market, it still plays a key role in the development of large, international models. Australian firms are already building on or deploying technology built on foundation models developed elsewhere; Australian consumers and residents are already subject to AI-driven products. ADM+S project *Testbed Australia* (led by Dr Thao Phan) demonstrates too that Australia is seen as a location for global firms to test and refine new technologies, including AI and automation.[79] Australia's population size, its diversity, and relative wealth make it an ideal proxy for larger Western markets.

Its position as a middle-power has also seen it play a role in global processes such as standards setting and regulatory modernisation. For instance, in the case of commercial drone delivery, Australia's Civil Aviation Safety Authority (CASA) were proactive in creating new risk assessment benchmarks for beyond visual line of sight (BVLOS) operations. The successful meeting of these benchmarks by firms such as Alphabet's Wing in their local trials in Canberra and Logan effectively created a proof of concept for their autonomous drone delivery system, a move that opened the doors for operations in their intended primary markets in Europe and North America.

Australian moves to regulate AI will need to coordinate, or be aware of, and even take advantage or join in regulatory moves overseas. We recommend:

1. **Strong participation in international cooperative mechanisms to manage risks arising from the largest models and actors**, whether at a treaty level or technical standard-setting, to address issues regarding larger models at a global level. Australia already participates in technical standard-setting efforts and global discussions regarding the development of common principles and cooperation in their implementation. There is an opportunity to work with governments in a similar position - across countries like Canada, New Zealand, Singapore, Japan - with a strong interest in ensuring an open digital economy but coupled with genuine protection for competition and individuals. Australia's access to a sophisticated workforce and expertise, and large trade relationship with China, values aligned with the EU and Canada, good relationships with the US and UK and regional relationships in the Pacific means it is well placed to play an important role in such discussions.

2. **Not leaving Australians behind in rights and remedies**: Australians are entitled to expect that they will receive similar levels of protection for their fundamental rights, health and opportunities as are enjoyed by people in countries with a similar high level commitment to democracy and human rights. That that expectation is not being met, whether across consumer protection law, or privacy, or human rights.[80]

3. Consideration of domains/issues where there is a role for Australian governments in protecting and promoting **uniquely Australian voices and interests**. **Education**-related AI systems trained on US data may not be appropriate for Australian schools: the makeup of our student body is different, as is the socio-technical context. Australian voices and perspectives in **news and entertainment media** would be another area for priority consideration, as would the question of how Australia's **First Nations people** will exercise their culture and sovereignties in this environment. Existing Indigenous

[79] See 'Tested Australia', *ADM+S, Research Projects* (Web Page).
[80] See *Consultation Question 2: current regulatory settings and gaps.*

'protocols' for AI governance developed globally,[81] and in Australia,[82] can serve as a starting point for a 'productive conversation with Indigenous communities about how to enter into collaborative technology development efforts' which, inter alia, respect context, cultural knowledge and Indigenous data sovereignty. Additional support for entities working in these spaces, and/or additional regulations or procurement requirements to require Australian content or training on Australian data for systems in these contexts may be needed.

4. Continue to **develop/invest in Australia's domestic capacity to develop and innovate** in AI research and technologies, responsible AI practices and the regulatory frameworks. Especially in light of point (3), it will be critically important that Australia does not simply import AI technologies and systems unmodified.

Consultation Question 3: Non-Regulatory Actions to support responsible AI practices in Australia

Consultation Question 3

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?

ADM+S Answer: among a wide range of non-regulatory actions the Australian government could take in this space, we suggest a focus on three in particular:

1. Invest in involving the Australian public in discussions about the direction of AI technology and its application: in order for the Australian public to trust AI technology and support its use, the current technocratic conversation needs to be broadened to be more inclusive. DISER is well-placed to lead such efforts, using established methods for science consultation and participatory governance.
2. Invest in education at multiple levels across society and the economy, to reduce the knowledge gap between the AI specialists (who create AI programs), and AI end users (who are responsible at the coalface for the deployment of AI, impacting themselves or other parties) or subjects (who are impacted by decisions and/or actions using AI).
3. To better address the rapid development and rapid *deployment* of AI technologies including new models and methods, adopt mechanisms to better connect leading research with government and the broader set of technology users, including in particular those involved in assessing the risks of applying AI. These mechanisms could be based on mechanisms used to inform the current inquiry, including by activating Australia's learned academies more regularly.

'No decisions about us without us' - invest in involving the Australian public in decisions

Conversations about AI and its future directions have been dominated, to date, by a technocratic discussion among experts from government, industry, and academia: which is not

[81] Jason Edward Lewis et al, 'Indigenous Protocol and Artificial Intelligence Position Paper' (Honolulu, Hawai'i: The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR), 2020).
[82] Angie Abdilla et al, 'Out of the Black Box: Indigenous protocols for AI', (Web Page).

surprising, because the technology can seem complex, the policy trade-offs around its use *are* complex, and both technologies and their deployment are changing.

But it is hard for communities to trust data-driven technologies that appear to be changing extraordinarily rapidly, and which they are told bring benefits, but also seem designed to find out everything about them, to get inside their head, to direct their behaviour, to make decisions about their life without the comfort of a human face.

To date, limited efforts have been made to involve Australians in the conversation about AI. Involving people in conversations about AI and ADM demystifies technologies and can steer their use to socially beneficial outcomes.[83] In particular, involving communities most at risk of harm from the use of AI and ADM enacts the principle of 'nothing about us without us', and can at the same time help de-risk the deployment of AI. This need not be a solely public process: the Ada Lovelace Institute recently studied the use of such processes in commercial AI labs.[84]

Doing participation raises complex questions: what is needed for representative participation to work?; what communication and information is needed to support participatory processes?; and, what resources are needed to sustain continuous participation processes that, like technologies, are never 'done'? These questions are already being taken on in many contexts: in participatory processes in government, which have been on the mind of the public service here for over a decade,[85] for which there are recent successes in Aotearoa New Zealand,[86] and about which the OECD recently produced guidelines;[87] in the work of Australian organisations like the Sydney Policy Lab and the new Democracy Foundation, researching and applying methods to engage citizens and build relationships in public decision-making; and in government-funded projects to use data differently, led by or with communities, such as Just Reinvest and the National Disability Data Asset.

There are a number of models available for public consultation on science. There is expertise available in Australia both within and beyond ADM+S. ADM+S Partners such as Consumers Health Forum of Australia can be engaged to involve citizens in applications of AI in healthcare. Consumer advocacy organisations such as the Australian Communications Consumer Action Network (ACCAN), CHOICE, and others are well placed to bring consumer rights and perspectives to the table in a range of specific domains. ADM+S Investigators have used participatory methods in partnership with a network of not-for-profit organisations to develop a tailored data capability framework for the sector, including data and AI governance best practice.[88]

In particular, we have identified above new capabilities which warrant a public discussion.[89] To this we could add: the challenges around the emergence of very large 'foundation' models, and emerging frontiers of value creation associated with AI create an opportunity for a democratic discussion about data governance that established some guidelines for what kind of data should be available for ingestion by AI models in what circumstances, ie for what purposes / and by whom.

---

[83] This is a core function within democracy, see John Keane, *The Shortest History of Democracy* (Collingwood: Black Inc, 2022) 8.
[84] Lara 'Groves, 'Going public: the role of public participation approaches in commercial AI labs' (2023) *ACM Conference on Fairness, Accountability, and Transparency*.
[85] Brenton Holmes, *Citizens' engagement in policy-making and the design of public services* (Parliamentary Library, 2011).
[86] Emma Blomkamp, 'Systemic design practice for participatory policymaking' (2022) 5(1) *Policy Design and Practice* 12-31.
[87] OECD, *OECD Guidelines for Citizen Participation Processes* (Guideline, 2022).
[88] Anthony McCosker et al, 'Developing data capability with non-profit organisations using participatory methods' (2022) *9*(1) *Big Data & Society*.
[89] See *New capacities requiring a public conversation.*

Education

It will be essential for Australia to develop multidisciplinary educational programs for the responsible development and use of AI and automation more generally – and these programs need to span the full range from formal higher education to SME-focused and community programs involving activities such as peer mentoring.[90] This program of work needs to extend beyond STEM education aimed at increasing the numbers of students studying relevant areas of computer science. Investing in Australia's domestic AI capability is important, but focusing only on this aspect of education risks exacerbating the ethical risks of AI, as well as leaving knowledge gaps between the AI specialists (who create AI systems), AI end users (who are responsible at the coalface for the deployment of AI, impacting themselves or other parties), and consumers or subjects (who are impacted by decisions and/or actions using AI).[91]

Addressing these gaps will support the responsible and ethical use of AI, by educating developers on the social and ethical issues associated with 'responsible AI', educating users on potential negative aspects of AI on the one hand (for example, improving recognition of bias and inappropriate uses), and generating awareness of new opportunities for responsible automation that serve the needs of diverse communities.

Ideally, AI education efforts would be targeted at multiple levels, to reach all demographics and circumstances where AI may be used. Education needs to occur in formal settings in the education sector, but also:

1. In workplaces, through training for employees who use AI in the course of their work.
2. Via professional organisations as part of continuing professional development (medical and legal professionals; health professionals, financial advisers etc), and via training, certification and accreditation for AI auditors;[92]
3. Efforts to improve general public awareness of AI and ADM, and the ability of the non-expert public to engage in meaningful debate about regulatory and ethical issues. This would require a targeted strategy of information and public engagement, ensuring the inclusion of marginalised groups that are likely to be affected by ADM processes but less likely to have access to necessary resources and support. This could include non-traditional communication and research outputs such as NYU's Responsible AI Comics,[93] and support mechanisms for potential subjects of AI decision-making (ie members of the public).

Education for AI end users needs to be more than passive poster campaigns promoting 'awareness', or guidelines for use, but must aim for genuine AI literacy through face-to-face or online training (even certification/ micro-qualifications), including peer mentoring programs such as those that have already been successfully used in other areas of digital literacy and inclusion.[94]

[90] Some inspiration can be drawn from Japan's Social Principles on AI, which includes 'Education/Literacy'
[91] An analogy can be made with medicine — while medical specialists require high-level training, the ability to administer medical first aid is the responsibility of a much broader demographic, which is extensively supported by practical training in schools and workplaces.
[92] Inioluwa Deborah Raji, Sasha Costanza-Chock and Joy Buolamwini, 'Change from the Outside: Towards Credible Third Party Audits of AI Systems' in *Missing Links in AI Governance* (UNESCO Digital Library, 2023) 5.
[93] *The We are AI Comic Series* (Web Page).
[94] Michael Dezuanni, Amber Marshall, Amy Cross, Jean Burgess and Peta Mitchell, 'Digital Mentoring in Australian Communities' (Australia Post and QUT Digital Media Research Centre, 2019)

Connect government and risk assessment with researchers in emerging areas

Research into emerging technologies and risks will be essential to Australia's future engagement with AI and ensuring responsible development and use in Australia. In addition, whatever regulatory stance is adopted, given the fast-moving nature of this field, the Australian government will need to better **connect and engage** with research into emerging developments and risks. This could be through regular research-focused roundtables and/or reports drawing on the full range of relevant disciplines, perhaps coordinated with the support of Australia's Learned Academies.

In the process of preparing this submission ADM+S researchers across disciplines and nodes contributed views on emerging areas of technological development and potential opportunities and risks. In these areas, there is no technical consensus as such; research is ongoing. Areas that were raised include the following.

The impact of increasing AI-generated content online

The amount of content online generated (wholly or partially) by AI is increasing. This content will be intentionally or unintentionally captured in scraping and data collection activities. This has implications for many forms of social science research.[95] Such content will likely also be used to train future models.[96] Research is emerging on the impact of this phenomenon. For example, some research suggests that AI systems trained on AI-generated data exhibit less robust real-world performance than systems that are trained on human data;[97] it could also amplify and reinforce existing biases. This has important implications for the safety, reliability, transparency, and predictability of AI systems efforts to date to distinguish AI-generated content from human-generated content have not been successful; well-established mathematical results known as the Neural Network universal approximation theorems imply that there will never be a generally reliable technological means to draw that distinction.[98] Research into other proposals such as mandated watermarking or labelling[99] are ongoing, but controversial for a range of reasons: difficulties in determining when watermarking ought to be required;[100] potential privacy and surveillance implications;[101] competition implications;[102] as

[95] Social science researchers use data scraping techniques to investigate public opinions and trends, which becomes harder to the extent that content is artificially generated. This has implications for the capacity of such research to provide the evidence base for policy. In addition, mechanisms and tools for annotating data which leverage human judgment, have increasingly become more reliant on Generative AI and synthetic data: for example the widely-used Amazon SageMaker now enabling synthetic data embedding, see, <https://www.youtube.com/watch?v=9q3_GSz_VCM>. This means 'ground truth data' for training and validating models might include real data, real-world-like data, and novel and unseen variations. This enables deep learning models to be more robust to novel scenarios. Similarly, Mechanical Turk workers that are 'hired' to perform human tasks or provide human-level judgments, or 'labels' for training data, are becoming more reliant of generative AI, see, 'Mechanical Turk workers are using AI to automate being human', *Tech Crunch* (Web Page).

[96] For instance, OpenAI's current generation of LLMs (GPT-3.5 and GPT-4) are known to have consumed data generated by previous iterations of these same systems (GPT-1 and GPT-2) in their training processes: anecdotal observations by ADM+S researchers, as well as conversations with OpenAI employees.

[97] Arnav Gudibande et al, 'The false promise of imitating proprietary LLMs' (2023) *arXiv:2305.15717*.

[98] See 'Universal approximation theorem', *Wikipedia* (Web Page).

[99] See Valérie Pisano 'How can we tell whether content is made by AI or a human? Label it', *Maclean's* (Web Page, 2023).

[100] Consider how would every two sentence or two word Gmail auto-reply be flagged as "AI-generated" - beyond including the words "AI-generated"?

[101] Cryptographically signed content may contain identifiers of devices carried by natural persons: For instance, as proposed by the industry *Coalition for Content Provenance and Authenticity* (C2PA).

[102] While OpenAI, Google and other companies could be compelled to embed watermarking, it's less clear whether this could be applied to the range of open source models and fine-tuned derivatives.

well as implications for people with disabilities or from different language backgrounds who may use AI-driven systems to engage on an equal footing in society.

Surrogate models and synthetic data

AI developers and data scientists in many fields appear to be embracing synthetic data (artificial data generated to mimic real-world data) and/or surrogate models (models used to substitute for real-world systems or users, for instance by generating surrogate data). The nature of 'synthetic' data can vary greatly - from augmenting real-world data (eg by flipping or rotating images, or changing the gender of words in text), to creating black-box surrogate models such as Generative Adversarial Networks that can mimic the distribution of data in a given data set. For example, in medical computer vision AI research it is now common for researchers to use synthetic data to augment real world training sets.[103] Similarly, simulated data are used heavily in the development and testing of autonomous vehicles due to the complexity and expense of collecting real-world driving data for numerous road, vehicle, weather, (etc.) configurations.[104] There will be strong incentives to use synthetic data to avoid the ethical and privacy concerns of using data about real people, but research will be needed to properly understand the potential benefits and/or risks of surrogate models and synthetic data. It is important that research continues on any limitations and risks of using synthetic data - and that any insights emerging from that research are widely communicated to professional communities who are likely to have incentives to use it.

As these two examples illustrate, technical expertise will be needed to explore, and explain to broader audiences in an ongoing way: we do not know all the risks (or opportunities) that developments around AI represent: but at the same time, the commercial imperatives for adoption are strong. The point, then, of this brief discussion of emerging technologies is threefold:

- With much cutting-edge model and method development occurring in commercial entities, the gulf or delay is shortening from technical development to *implementation* affecting people, society, the economy and environment;
- Ongoing research - including multidisciplinary research - will be needed into emerging technologies, the risks they may carry as well as the opportunities;
- Insights from this research must be connected more rapidly into government and explained to the broader user base. This speaks to the need to improve the connection between independent research and government.[105]

Consultation Question 4: coordination of AI governance across government

Consultation Question 4

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?

[103] See, eg, Sema Candemir et al, 'Training strategies for radiology deep learning models in data-limited scenarios.' (2021) 3.6 *Radiology: Artificial Intelligence* ; and Alvaro Fernandez-Quilez. "Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability." (2023) 3.1 *AI and Ethics* 257-265.
[104] See, eg, NVIDIA's DRIVE Simulator which is widely used in industry, or the open-source CARLA simulator.
[105] See *Connect government and risk assessment with researchers in emerging areas*.

ADM+S Answer: We recommend a central function, across government, with technical expertise and the ability to connect to further technical expertise outside government, and the power to drive practical change within and beyond government.

We have argued that the development of AI has broad impacts[106] demanding:

- Reconsideration of assumptions across existing legal regimes, including Australia's laws around the use of data for training;
- Consideration of Australia's framework for the enforcement of laws relating to human rights and consumer protection;
- Societal level discussion of new capacities, in order to make decisions as an Australian community regarding which uses of automation and AI are acceptable, and where any limits lie;
- Consideration of environmental impacts;
- Connection between cutting edge research (across technical and non-technical disciplines) and government.

Addressing all of these questions will, we think, require some central function across government with technical expertise, the ability to connect to further technical expertise outside government, and the power to drive practical change within and beyond government. This is a tall order. This function will need to:

- Provide leadership on issues that cross specific domains - potentially including, for example, questions around transparency, accountability across AI supply chains, and basic requirements such as data quality, documentation;
- Provide leadership on a long term strategy and governance approach;
- Provide leadership in the non-regulatory initiatives we have identified above: public conversations, education efforts, and developing connections with the independent research and commercial research communities;
- Provide guidance to domain regulators on technical questions.

We would advise against reliance solely on an independent statutory officer (in the style of an Ombudsman) or advisory body. Such actors can tend to create an appearance of 'doing something', but are generally either reactive (their role is to respond to complaints for example) or they have no real capacity to change things (their only role being to make recommendations, without powers or implementation or monitoring). Such entities will have ongoing roles, including by collating information and learnings across government silos, helping to identify patterns to inform future legislative reform relating to AI. But to drive change within the public sector, we would suggest that what is needed is a body attached to a powerful ministry, with a broad mandate to supervise the adoption and use of ADM systems across the public sector.

[106] See *Consultation Question 2: current regulatory settings and gaps*; *Consultation Question 3: Non-Regulatory Actions to support responsible AI practices in Australia*; and *Consultation question 5: Australia in relation to international developments.*

Consultation Question 6: distinction between public and private sector AI use

Consultation Question 6

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

ADM+S Answer: yes, a different approach is required: different legal obligations apply to public and private sector decision-making. Public sector decision-making must meet the requirements of administrative law, which imposes standards of transparency, the obligation to explain (in some cases) and obligations to observe natural justice. The higher standards applicable to public sector decision-making, among other factors, make use of AI in the public sector a good place to develop and test governance mechanisms.

Is a different approach required for the public sector?

The ways in which the regulation of automated tools and AI interacts with the requirements of public and private law is the subject of ongoing research.[107] Nevertheless, we can say with certainty that use of AI by the public sector is different, for a number of reasons.

First, government has **power** the private sector does not have, ie government decisions may be unilaterally binding on individuals and businesses, who do not have a choice to opt out from government influence, or choose a different service provider, as they (may) do in the private sector. For these reasons, and as a matter of democratic legitimacy, public sector decision-making is held to higher standards. The exercise of public power must be **authorised** and **legitimate**; it must meet the requirements of **administrative law**.[108] There are accordingly different legal standards for decision-making and action, as well as transparency and contestability, in the public and private sector. Governments are obliged to promote the well-being and rights of *all* residents and citizens, and to pay attention to societal inequality and social justice.[109]

This does not mean that private sector decision-making should escape scrutiny. Even beyond the special scenarios of monopolies, or essential and/or outsourced services, we may need to think about the standards we wish to apply to private decision-making in the face of known risks (such as bias or unfairness) arising from the use of AI. Private actors are bound by different sets of legal principles found in corporations law, consumer protections, privacy etc. Still, at present, and subject to anti-discrimination law, we do not require, for example, private actors to give people an explanation of commercial decisions (say, to refuse commercial service to someone) in the way we would expect of a public sector entity.

There is also an intermediate set of actors: private sector entities holding (regulated) monopolies, delivering essential services, or providing services to/on behalf of the

[107] Potential implications, for example, include the need to resolve differently what is sometimes a choice between greater accuracy and greater explainability. Another implication is that having to test models for explainability purposes creates operational and environmental costs - costs that may be more heavily imposed on the public than the private sector.

[108] For a detailed discussion, see NSW Ombudsman, *The New Machinery of Government: Using Machine Technology in Administrative Decision-Making* (NSW Government, 29 November 2021).

[109] AI systems threaten to increase inequality across society. It is more challenging to argue that (most) private sector entities have obligations to address broader societal equality beyond the obligations under anti-discrimination law. We do note however, that to the extent that private sector entities deliver essential services, they will have broader obligations, reflected in regulatory regimes relating to telecommunications, health, energy, water and more.

government.[110] In general, such entities are already (and should be) held to higher standards than other parts of the private sector, and when developing rules for the public sector, consideration must be given to which rules need to be extended to these entities.

Public sector use is a good place to start

The higher standards that apply to the public sector - together with other factors - make public sector use of AI a good place to start when developing requirements for responsible/safe AI. Governments could readily take action via a combination of legislation, binding guidelines and/or internal policies to ensure that all systems designed and built, procured, and used by public sector actors adhere to the highest standards across issues of respect for human rights, privacy, transparency, safety, cybersecurity and robustness, and are developed in an inclusive way and by actors and companies with appropriate expertise.[111]

Focusing on the public sector (including private sector entities delivering essential services or acting on behalf of government):

- moderates the risk of restricting innovation;
- ensures that the activity that cannot be avoided by citizens (eg, taxation, public services) takes a more precautionary approach; and
- Is consistent with Australia's obligations under international law to protect and promote human rights;[112]
- Can lead the way: the public sector is a significant and influential procurer of AI services and systems in the Australian market; by setting appropriate in standards for procurement for its own services, the public sector can invest in and promote the development and adoption of private sector standards that will have a broader impact on the market as a whole;
- Monitoring and enforcement are possible without judicial intervention (if contracting is properly managed) and rules can be quickly adapted according to the lived experience of the public sector and the residents and citizens it serves; and

Carved-out regimes in certain sectors of public activity could be established, creating 'de facto' regulatory sandboxes for experimentation.

[110] See, eg, *O'Brien v Secretary, Department Communities and Justice* [2022] NSWCATAD 100, considered in a case study #1 examined by the NSW Information and Privacy Commissioner, in which the use of a third-party contractor impacted a GIPA request including information on an algorithm (the request was initially denied because the information was not held by the agency but by the contractor). The IPC commented that: 'The case raises urgent questions about the access of information held by both third-party contractors and government agencies.'
[111] See, eg, Australian Computer Society, *'Frameworks and Controls for Data Sharing'* (February 2023), which highlights the need not only for technology to meet certain standards, but for providers who are handling data (and especially linking data) to have expertise and training appropriate to the sensitivities and impacts of the data and uses.
[112] José-Miguel Bello y Villarino and Ramona Vijeyarasa, 'International Human Rights, Artificial Intelligence, and the Challenge for the Pondering State: Time to Regulate?' (2022) 40(1) *Nordic Journal of Human Rights* 194.

Consultation question 7: Supporting responsible AI within government agencies

Consultation Question 7

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

ADM+S Answer: In our view, the most critical gap around public sector use of AI at present is inadequate accountability, and insufficient supervision or mechanisms in place to ensure the implementation of existing principles designed to govern government use of AI and ADM. We further recommend mechanisms to surface use and enable the sharing of expertise within and across government departments and agencies - such as a register of systems.

While government institutions have proposed principles that should guide government use of AI and ADM, there is at present inadequate accountability, and insufficient supervision or mechanisms in place to ensure their implementation in practice. For example, various Commonwealth level policy documents (also mentioned in this consultation paper) require that ADM systems are legal, safe, and of high quality. It is not clear however that any mechanism is in place to ensure this standard is met (from either a legal or technical perspective). Thus, an important step is to act to ensure **compliance** with framework and accountability for non-compliance, as well as ongoing monitoring of the use of the system.

There is a long history of regulatory institutions, laws, policies, procedures and processes to seek to provide administrative justice/procedural fairness and appropriate transparency in government decision making. These processes provide a strong basis for ensuring AI use in government decision making is safe and responsible. Specific gaps that could be addressed include the following:

1. **There is a need for greater legal and regulatory clarity in automated decision making of administrative decisions:** For over 30 years, the *Social Security Act* has recognised that computers can be delegates of the Secretary in making decisions under the Act.[113] However, *ATO vs Pintarich*[114] complicated the legal status of automated decision making, even when legislation says computer decisions can be used. These matters are also well canvassed in the Commissioner Holmes' Report from the Robodebt Royal Commission, Chapter 17.[115]

2. **We need avenues to challenge the administrative decision-making process to address systemic issues**: The administrative review system is built on a process for reviewing *individual* administrative decisions, but automation creates systemic processes that can create systemic erroneous decisions – as the Robodebt scheme demonstrates. New mechanisms in which to ensure systemic safety and responsibility is needed, including by ensuring access to system design in appropriate circumstances.

3. **We need to develop new oversight and review mechanisms to assess if use of 'upstream' AI/ADM is unfair or breaches discrimination laws:** There are a growing range of AI and ADM processes that are invisible to our administrative review and

[113] For example, the *Social Security (Administration) Act 1999* Section 6A states in part: "A decision made by the operation of a computer program under an arrangement made under subsection (1) is taken to be a decision made by the Secretary." There are now many other laws with similar provisions.
[114] *Pintarich v Deputy Commissioner of Taxation* (2018) ATC 20-657.
[115] Robodebt Royal Commission (n 43) ch 17.

regulatory and oversight review processes, because they involve (for example) background risk assessment rather than the making of a decision.[116] The Netherlands case of SyRi provides a contrasting example, whereby such tools were found to be biased only through the use of EU's Human Rights Act.[117]

4. **Public registers:** In our research experience, leadership within government may have only a limited sense of the extent of automation and AI use across government services, and there may be little shared awareness across departments and agencies. For this reason, mechanisms for surfacing use and sharing expertise and experience should be considered, such as registers of AI. Guidance could be sought from other jurisdictions that have been developing such registers of AI use.[118]

Risk assessment is discussed further below.[119]

Consultation question 9: transparency

Consultation Question 9

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.

ADM+S Answer: Transparency will be valuable for all those affected by the operations and outcomes of an AI, from non-technically trained users to expert investigators. However, the regulation for transparency requires an actionable, measurable standard, capable of graduated modulation depending on a range of critical contextual factors: the circumstances of an AI's operations (including users' interests in privacy and security); the level and nature of risk entailed; and the kind of people or organisations for whom transparency is intended.

- Transparency is a value that should be mandated for AIs and ADMs across the public and private sectors.
- Regulators will need to specify the content, volume and form of the communication required to satisfy a transparency obligation in accordance with the contextual factors.

Transparency in the context of AI is a term used by different actors, governments and regulators to mean different things.[120] There is little consensus, both in Australia and internationally, on what transparency in relation to AI means and how it relates to other concepts, especially AI explainability. For example, Australia's AI Ethics Framework[121] and DTA

[116] For example, risk assessments of tax or welfare fraud, risk profiling of VISA applications or profiling of long-term unemployment of jobseeker all provide insights for government action that may lead to administrative decisions. Yet, because they are not directly involved in the decision making they cannot be reviewed for possible bias, discrimination or accuracy, through traditional administrative review processes.
[117] Marvin Van Bekkum and Frederik Zuiderveen Borgesius 'Digital welfare fraud detection and the Dutch SyRI judgment' (2021) *23*(4), *European Journal of Social Security* 323-340.
[118] See, eg, Meeri Haataja, Linda van de Fliert and Pasi Rautio, *Public AI Registers Realising AI transparency and civic participation in government use of AI* (Whitepaper, September 2020).
[119] See below, *Consultation question 14: is a risk-based approach the right one?; Consultation question 15: addressing some limitations of a risk-based approach;* and *Consultation question 17: elements of a risk-based approach.*
[120] On the nebulous nature of transparency see, eg, Luke Munn, 'The Uselessness of AI Ethics' (2023) 3(3) *AI and Ethics* 869.
[121] Department of Industry, Science and Resources, *Australia's Artificial Intelligence Ethics Framework* (7 November 2019).

Guidance[122] refer to 'transparency and explainability' (albeit with more focus on the former), NSW AI Assurance Framework mentions 'transparency' only and closely ties it with contestability,[123] while the Commonwealth Ombudsman requires AI to be 'understandable'. Internationally, the US NIST AI RMF discusses the notions of 'accountable and transparent' noting that accountability presupposes transparency,[124] while both Canada and the EU have linked transparency to, among other things, an understanding of capabilities, limitations and potential impacts.[125]

Unpacking the concept requires clarity as to **who** is required to be transparent**,** as well as for **whom.** We need also to be clear about **what** information needs to be communicated, and **how**. In addition, transparency cannot be an end in itself: transparency without any capacity to act on the information disclosed is unlikely to foster trust and confidence. Therefore we also need to be clear about the purpose **of transparency.** For instance, the AI transparency we demand for the purpose of investigating a serious accident caused by an autonomous vehicle is likely to look quite different from the transparency required by the general public considering the use of a robotic vacuum cleaner.

Researchers and commentators currently agree on several key points:

1. Current AI and ADM systems are often insufficiently transparent;[126]
2. Greater transparency will be required to address the enforceability gaps identified in this submission;
3. Transparency mandates need to be well-defined and framed for a different purposes and situations. If we simply require 'transparency' then the party of whom it is required may interpret the concept to their own advantage, at the expense of broader policy objectives.[127]
4. The technologies of AI are developing rapidly, which means that some of our expectations of transparency and what they entail in system design and deployment are likely to change.

It is also important to note that while transparency and trustworthiness should always be an ethical design objective, public trust and confidence in AI systems are not always desirable. Some recent work suggests that explainability mechanisms in AI systems may give rise to over-confidence in automated systems, leading to a public misapprehension as to their infallibility.[128] This points to the fact that transparent systems are not a solution in themselves to the problem of responsible AI; we note elsewhere in this submission the need for further work on enhancing public literacy in the use of AI and ADMs.

Defining transparency

We understand transparency, like the related objective of explainability, to be an ethical value concerning the communicative aspects of AI: it refers to the quality of information that is

[122] Digital Transformation Agency (DTA), *Guidance for Adoption of AI in the Public Sector* (Guidance, 2023).
[123] NSW Government, *NSW AI Assurance Framework* (Guidance, March 2022).
[124] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (January, 2023) 15.
[125] Innovation, Science and Economic Development Canada, *The Artificial Intelligence and Data Act (AIDA) – Companion document* (Guidance, March 2023); Proposed EU AI Act (n 67) art 4a.
[126] See, for example, NSW Ombudsman, *The New Machinery of Government: Using Machine Technology in Administrative Decision-Making* (NSW Government, 29 November 2021).
[127] An example on how transparency could be ensured with relation to facial recognition technologies in law enforcement sector see Rita Matulionyte, 'Increasing Transparency around Facial Recognition Technologies in Law Enforcement: Towards a Model Framework' (March 20, 2023).
[128] H. Kaur et al, 'Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning' (Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020).

stored and communicated about an AI system's operations to those who use the system, or have an interest in what it does. Such communication may be generated by the system itself or may come its designers, developers or operators. While it is generally agreed that such communications must be honest and meaningful for their recipients, it is also well understood that no single standard of transparency is likely to capture the complexity and diversity of people's interactions with AI, or meet the needs of all users or stakeholders. Therefore, a differentiated and graduated set of requirements for transparency will be required, taking into account the critical contextual issues we have already noted as to the intended purposes of disclosure and the intended users of the information disclosed. Nor is transparency only about the kind and volume of the information communicated about an AI system and its actions: an appropriate context and format for communication is also necessary, whether that takes the familiar form of a product manual for consumer use, a visual depiction, or a detailed model card intended for a technical audience.

Transparency for what purpose?

Transparency can serve many purposes, such as increasing trust in AI, ensuring contestability of decisions produced with the help of AI, enabling external scrutiny and quality assurance of AI, or demonstrating accountability by responsible stakeholders. If these purposes are clear and achievable, then we can determine whether any given set of transparency measures contribute to their achievement.

It is important to ensure that legitimate information needs of all stakeholders can be satisfied, ie that suitable and relevant information is made available to various stakeholders. At the same time, it would be unreasonable (and financially and environmentally costly) to expect that *all* stakeholders are always provided relevant information about *all* AI technologies, especially in low-risk scenarios.

We therefore support a specified and graduated range of transparency requirements, modulated according to the critical contextual factors including the level of risk should differ depending on the risk/impact level of the technology, ie the higher risk/impact of technology, the more transparency is required. This approach has been adopted in the draft EU AI Act which requires minimum transparency information about low-risk AI systems (letting users know about the system) and sets high transparency duties for high-risk systems (registering in the EU high-risk AI system register where detailed information is to be provided).[129] The information provided should be related to the general and specific risks of the system, ie to enable stakeholders to assess whether risks have been properly identified and mitigated.

The actor involved is also important: as noted, governments should usually be held to higher standards of transparency, as the exercise of public power must be legitimate.[130]

Transparency for whom?

Before defining what information needs to be provided and how, the Government needs to clearly identify which stakeholders a specific transparency measure will target, since stakeholders require different types of information corresponding to their training and levels of understanding and their goals/tasks. Potential stakeholders include:

- the general public, including non-expert and expert members;

[129] See proposed EU AI Act, arts 13, 52 and Annex 8. See also Central Digital and Data Office, *UK Algorithmic Transparency Recording Standard* (Guidance, 5 January 2023) which sets transparency requirements only with relation to systems that meet certain risk/impact threshold.
[130] See *Consultation Question 6: distinction between public and private sector AI use*.

- individuals directly affected by an AI or ADM output;
- courts and litigants during litigation;[131]
- regulatory authorities;
- third party auditors/certifiers; and
- independent experts, including civil society organisations and university-based researchers.

Transparency of what?

In terms of type of information that various stakeholders might need, it could be, among other things:

- details on the capabilities, limitations, safety and risks of the model;
- details on the methodology used to train the model;
- details of how an AI system is relied upon and integrated within a broader decision-making framework;
- details on the datasets used in pre-training, fine-tuning, reinforcement learning;
- disclosure of information about who is buying and deploying AI systems in what contexts;
- informing individuals on when they are interacting with an AI agent embedding metadata and labelling AI generated material; and
- explanation of how and why an AI system has produced particular outputs, decisions or outcomes for particular individuals or classes of people.

Transparency how?

Information can be provided via different pathways, for example:

- Proactively: the AI developer or deployer proactively publishes certain information whether generally (eg, on websites, via product manuals, data or model cards or technical reports) or to individuals (eg accompanying a decision); and/or
- Reactively: information is disclosed upon request.
- Passively: There is an absence of barriers and obstacles to systematic public observation of a system's operations such as via audits of its outputs.

Information can also be presented in different, and more or less accessible ways, such as in dashboard or visual form.

From a technical perspective, there are different technical ways to provide **access** to information, such as:

- Descriptive access: access to documentation and data logs
- Query access: the ability to specify arbitrary model inputs and observe the computed model outputs.
- Debugging access: the same level of access that a model developer would have during development or to identify bugs.[132]

[131] *ACCC v Trivago N.V.* [2020] FCA 16.
[132] Henry Fraser, Aaron J. Snoswell, and Rhyle Simcock 'AI Opacity and Explainability in Tort Litigation' (ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea , June 21–24 2022).

Table 1: Levels of AI system access, and the corresponding affordances they offer a plaintiff

	Description of architecture, pseudo or actual model source-code without corresponding checkpoint files	Specify model inputs and observe model outputs	Incrementally step-through model processing, observe intermediate calculations
Debugging Access	✓	✓	✓
Query Access		✓	
Descriptive Access	✓		

Source: Fraser et al (2022)

Monitoring dynamic systems is a challenge. AI systems interact with and change the world in complex and recursive ways. Driver navigation systems, for example, change traffic patterns and flows; while recommender systems nudge end user choices, which can then help shape future recommendations. This demands sophisticated ongoing reporting with a view to understanding how the reward / reinforcement structures embedded in different systems are affecting and being affected by the world they act in, such as that proposed by Gilbert et al (2022).[133]

From a regulatory perspective, transparency of different kinds can be mandated:

- in domain-specific legislation;[134]
- in horizontal legislation where it is appropriate to impose documentation standards on systems;[135]
- via federal and state freedom of information legislation and procurement standards to establish what information government agencies should disclose about AI/ADM systems they use, and how transparency should be achieved. The aim should be to minimise the negative effects of trade secrets on transparency and also ensure that sufficient information can be accessed by both government and the public, even where held by a private contractor;[136] and/or
- via a public register of high-risk ADM systems, at least those used by government agencies (federal or state).[137]

The need to address barriers to transparency

In some situations, information desired by stakeholders cannot be accessed or disclosed due to privacy, confidentiality, security or other reasons. Commercial confidential information (trade secrets) might become an especially important barrier in ensuring transparency around

[133] Thomas Krendl Gilbert et al, 'Reward Reports for Reinforcement Learning' (2023) arXiv:2204.1081720.
[134] See for example the EU Digital Services Act sets specific transparency regulations with relation to specific online service providers or online services (intermediary services, automated recommender systems, online advertising), see Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), *OJ L 277, 27.10.2022, p. 1–102*. In Australia, sector/domain specific transparency duties could be implemented for example via laws on data protection, financial services, medical device legislation and more. Specific transparency rules are likely to be needed for specific government sectors, such as law enforcement or judiciary (administration of justice) where the use of AI might lead to significant impacts on human rights. For example, Interpol/ WEF Guidelines provide detailed transparency requirements for the use of facial recognition technologies in law enforcement context.
[135] An example being the proposed EU AI Act (n 70).
[136] See Rita Matulionyte, 'Government Automation, Transparency and Trade Secrets' (under review). Can be requested from rita.matulionyte@mq.edu.au.
[137] For precedents see the UK Algorithmic Transparency Standard; Amsterdam Algorithm Register, or the proposed EU Register for high-risk AI.

AI. In *O'Brien*,[138] trade secrets was a key reason for denying a request under freedom of information legislation for information about an algorithm that denied social housing benefits to the requesting individual.

While trade secret protection cannot and should not be entirely removed, there are ways that excessive or overreaching trade secret claims by AI developers and vendors could be addressed. For instance, laws or procurement contracts can require certain *essential* algorithmic information to be disclosed to certain stakeholders (eg regulatory authorities, independent experts) or to the public. Confidentiality interests should be balanced against other public interests, and cannot always prevail.

Transparency requires new research methods and infrastructure

The observation and evaluation of AI systems by independent researchers is an important aspect of transparency. This requires new methods, tools and infrastructures for gathering data and analysing algorithms, content and interactions. AI Audits, data donations and test environments are some of the key approaches being developed which will benefit from investment as national research infrastructure with applications across sectors and disciplines[139].

AI audits are a method of repeatedly querying an algorithm and observing its output in order to draw conclusions about the algorithm's opaque inner workings and possible external impact[140]. Audits are currently required by the EU's Digital Services Act, and are either required or being considered by some US jurisdictions. Audits may involve planning oversight, continuous monitoring, and retrospective analysis of system failures and require standardised audit trails as well as audit tools, methodologies, and resources – many of which are only emerging. AI audits have recently been taken up by civil society organisations such as the US-based NGO For Humanity[141] which has proposed an Independent Audit of AI Systems (IAAIS) and The Mozilla Foundation (2022) which has set up the Open Source Audit Tooling (OAT) Project.[142] Some AI audits are broader than purely technical audits and require an analysis of the surrounding socio-technical environment of an AI system. For instance, led by Professor Paul Henman, the ADM+S is developing a practical trauma-informed audit framework intended to enable social service providers, including the government, to assess whether AI systems used in social service delivery may cause or perpetuate trauma for service users.[143] The framework is designed to prompt critical reflection amongst service providers on whether decisions, practices or processes in the design, development and deployment of an AI system are consistent with well-established principles of trauma-informed care.

Data donations: Data donations sourced directly from users of online platforms is a new approach to enhancing the observability of digital services. This involves developing browser extensions, plugins or apps for collecting certain kinds of data or activities. ADM+S data donation projects include the Australian Search Experience and the Facebook Ad Observatory. National-scale research infrastructure such as the proposed Australian Social Data

---

[138] *O'Brien v Secretary, Department Communities and Justice* [2022] NSWCATAD 100.
[139] See Burgess et al, 'Towards large-scale research infrastructure for digital platform observability' (2023) *Computational Communication Research,* (submitted for publication).
[140] Danaë Metaxa et al, 'Auditing Algorithms: Understanding Algorithmic Systems from the Outside In' (2021) 14(4) *Foundations and Trends in Human–Computer Interaction* 272-344.
[141] 'Auditing AI and autonomous systems: building an #infrastructureoftrust', *NGO For Humanity* (Web Page).
[142] 'Mozilla Foundation Open Source Audit Tooling (OAT) Project', *Mozilla* (Web Page).
[143] 'Trauma-informed AI: Developing and testing a practical AI audit framework for use in social services', *ADM+S, Research Projects* (Web Page).

Observatory (ASDO),[144] promise to provide access to large-scale social, economic and cultural data generated by user interactions with a range of AI systems.

Test Environments: Test environments, also known as "cyber ranges" are emulations of technical systems often used in cyber security for testing issues in a contained environment. This requires a realistic simulation of a digital platform to evaluate the impact of attacks and the effectiveness of various defences. It is now possible to use generative AI to test different types of content and their impact within a test environment. Researchers need access to these types of tools at scale and the ASDO proposal includes a test environment for predicting, testing and analysing risks within AI and digital platforms.[145]

Consultation question 10: Prohibitions

Consultation Question

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?

ADM+S Answer: We have identified elsewhere in this submission new capacities arising from the combination of wide availability of data, automation and AI. We submit that a public conversation about those capacities - and limits on their use - is warranted. The conversation on the limits of acceptable technology use cannot be a purely technocratic conversation as tends to occur through consultation processes like the present one. A wide cross-section of the community, including impacted groups, must be enabled to participate.

Consultation question 13: Conformity and Assurance

Consultation Question

13. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

ADM+S Answer: we refer to a separate submission authored by a subset of ADM+S researchers (submission by Fraser et al) who have been considering conformity and assurance regimes in more detail and have discussed those regimes with colleagues in the EU.

As we have noted above,[146] the EU AI Act and its risk-based approach depends on a conformity and assessment infrastructure. It has been designed with the European single market in mind, and promotes the evolution of private risk-assessment certification and assurance in line with the comprehensive network of standards that exists in that jurisdiction. Australia does not have the same conformity ecosystem nor does it orient its product safety regime around trade and market harmonisation. A subset of ADM+S researchers (Fraser et al) have separately made

[144] *Australian Social Data Observatory* (Web Page).
[145] Australian Social Data Observatory, 'Test environments for social data and digital platforms' (Web Page).
[146] See *Consultation question 5: Australia in relation to international developments*.

a submission focused on these issues. Those researchers have been considering conformity and assurance regimes in more detail and have discussed those regimes with colleagues in the EU.

Consultation question 14: is a risk-based approach the right one?

Consultation Question

14. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

ADM+S Answer: ADM+S offers qualified support for a risk-based approach. In particular we argue:

- Any risk assessment must take into account key elements of the sociotechnical context: questions of diversity and inclusion, and the impact of supply chains (including the actors who collect and clean data).
- Requiring firms to undertake risk assessment and implement mitigations is an incomplete answer unless accompanied by shifts in the legal framework that ensure there are consequences when the process fails (as exist in other jurisdictions that are considering or have adopted a risk-based approach).

We also explain in our submission that there is an alternative approach that already exists in some public law mechanisms, which is a structured balancing test of public interest considerations. Such a process - familiar in the public sector - could be adopted/adapted in relation to the use of ADM/AI in public sector contexts.

Risk assessment must situate responsible AI in its social contexts

A risk-based approach is being developed in a number of countries (although not all). But every country is different. A risk-based approach for Australia must be informed by the social and technical context in which the technology will operate, including matters specific to the Australian context.

Any risk assessment must consider the challenge of digital exclusion in Australia as well as broader questions of inclusion and diversity

ADM+S research indicates that 23.6% of Australians are digitally 'excluded' or 'highly excluded' - meaning they have challenges of access, capacity to pay, and/or lack digital literacy.[147] First Nations Australians experience digital exclusion at higher rates: there is a 7.5 point gap in digital inclusion between First Nations and non-First Nations people in Australia - and a larger gap in remote (21.6 points) and very remote regions (23.5 points).[148]

Digitally excluded Australians experience challenges with access to reliable and affordable internet, making it hard to participate in society and schooling where that which increasingly depends on internet access. Downstream impacts include under-representation in AI training

[147] Thomas, J., McCosker et al, ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University, Swinburne University of Technology, and Telstra, *Measuring Australia's Digital Divide: Australian Digital Inclusion Index: 2023* (Report, 2023). Digital inclusion requires a level of access, ability, and capacity to pay for digital technologies and online connection to effectively use digital services. Those who cannot meet this threshold are considered to be digitally excluded. A person who is digitally excluded typically experiences compounding barriers to economic, civic and social participation. The report defines a person as excluded or highly excluded where they receive an Index score of below 61 on a scale of 0 to 100 (where 0 is least; 100 is most included.)

[148] 'Mapping the Digital Gap', *ADM+S* (Web Page).

datasets,[149] an inability to participate in development and governance of AI systems, and limited access to AI's benefits that AI systems can provide.

More broadly, the risks and benefits of AI are unlikely to be evenly distributed or safely managed without due care for, *and the active involvement of*, diverse groups of Australians.[150]

The need to address different actors in AI supply chains

A key challenge in the effective implementation of a risk assessment system is the complexity of AI supply chains. Effective regulation of AI and ADM systems must attend to issues arising throughout the supply chains that sit behind AI deployments by any single particular company.[151] This includes: providers in the technology stacks and cloud services associated with complex applications; the scraping and aggregation of training data by large tech companies; the various processes (surveillance, aggregation, labelling etc) and industry actors involved in the creation and circulation of AI training datasets (a multi-billion dollar market[152]); and labour issues associated with data labelling and annotation.[153] Responsibility cannot only sit with the final actor in the chain, especially if they are not the one best placed to address any risks.

Internationally, policymakers have started paying attention to these questions: for example, the EU AI Act[154] as revised over the course of its development includes differentiated obligations for developers and deployers of AI technology, and also requires that building models not be otherwise in violation of EU law, including copyright and privacy.

A risk-based approach can provide a useful, *ex ante* method for reducing harms, as part of a system to promote safe and responsible AI

As noted above in the discussion of Consultation Question 2, the capacity of AI to cause harm on a scale, and at a speed not previously possible is a strong argument in favour of **intervention *ex ante:*** requiring mitigation of risks at the design and development stage of AI.

As we also pointed out there, when we adopt a risk-based approach, we are requiring firms to identify, and mitigate, certain risks of harm. But for this to be effective, **there must be a 'what then'?** For a risk-based approach to lead to genuine improvement in the technologies applied, there must be a likelihood of consequences if organisations fail to take mitigating action and harm results. That means not just enforcing the obligation to undertake risk management, but laws prescribing the act that creates the risk of harm, and a credible threat that that law can be enforced. Australia will need to address current gaps in both our *rules* (ie laws/legal frameworks) and enforcement capacities, as discussed above.[155] Proposals like the EU AI Act are part of a much broader set of laws, and law reforms, that already provide people, and advocacy and other collective organisations to seek remedies for AI-imposed harms.

---

[149] Although the risk of extraction and exploitation of culturally-sensitive First Nations knowledge is also a problem: see Keoni Mahelona et al, 'OpenAI's Whisper is another case study in Colonisation' (Blog, 24 January 2023).
[150] On gender and racial diversity in the AI industry and its links with algorithmic discrimination' see, eg, AI Now Institute, Sarah Myers West, *Discriminating Systems: Gender, Race, and Power in AI* (Report, 1 April 2019).
[151] See, eg, Jennifer Cobbe, Michael Veale, and Jatinder Singh, 'Understanding accountability in algorithmic supply chains' (ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), Chicago, IL, USA, June 12–15, 2023).
[152] According to market analysts: see, eg, Grand View Research, *AI Training Dataset Market Size, Share and Trends Analysis Report, By Type (Text, Image/Video, Audio), By Vertical (IT, Automotive, Government, Healthcare, BFSI), By Regions, And Segment Forecasts, 2023 – 2030* (Report, 2023) on the AI training dataset market.
[153] Josh Dzieza, 'Inside the AI Factory', *The Verge* (Web Page, 20 June 2023); Perrigo, Billy, 'The $2 Per Hour Workers Who Made ChatGPT Safer' (18 January 2023) *Time*.
[154] Proposed EU AI Act (n 70).
[155] See discussion above, *Consultation Question 2: current regulatory settings and gaps.*

In addition, even within the EU AI Act itself, risk management is only one of the requirements imposed on high risk systems: other requirements relate to data quality, transparency, documentation, human oversight, robustness, accuracy and cybersecurity. Only some of these are reflected in the Discussion Paper elements. Australia should consider whether it needs to establish baseline standards (beyond an obligation of risk management) which AI systems should meet - whether generally, or perhaps initially for systems that will be used in the public sector? If so, are the European standards the right starting point? Product safety standards, and 'fitness for purpose' product and service standards are notoriously difficult to apply to software; there is no established history of safety testing for most software/data products. Some of these standards have some overlap with privacy law (where information is personal information), domain-specific rules (eg record-keeping required in highly regulated industries such as banking), and Australia's voluntary AI Ethics Principles (which make reference to robustness and security, for example).

Assuming Australia chooses to set some baseline requirements for AI systems, a question arises: is Australia prepared, like the EU, to rely on technical standard setting and conformity/assurance systems for testing compliance of AI systems with requirements across *all* standards? Or are there some issues (such as impacts on human rights) not well suited to conformity/assurance systems as currently configured? We have argued in this submission that Australia lacks the developed conformity and assurance framework on which the EU approach depends - and (as noted in the separate submission by ADM+S researchers Fraser et al) it is not even clear whether Europe's more developed approach is readily adaptable to AI.

A deliberative weighing or balancing test is a possible alternative to impact or harms-based risk assessment

The question of whether risk-based approaches are appropriate for the regulation of AI is the subject of active debate. As Kaminski has argued, in a longer academic piece[156] and a shorter policy paper,[157] one issue with a 'risk-based approach' is the framing, which tends to direct, and narrow, thinking about the impacts of AI. One problem is that it tends take use cases almost as a given, while focusing on harms and their mitigation. This can tend to downplay a focus on - and scrutiny of - potential benefits and/or how the technology could be directed to maximise those benefits. The use of AI has important potential benefits, which need to be considered and pursued. Claimed benefits should of course also be subject to scrutiny to see if they are in fact achieved, and we need ways to adjudicate between varied risks and these benefits.

Especially in the public sector context, a deliberative weighing process or 'balancing test' of public interest considerations is one way to think about technology from the perspective of the public interest, such as is used in the context of Freedom of Information laws. This would require decision makers to record the weight of relevant considerations against others, for example the relative weight of a potential privacy risk against projected customer service benefits. Similar processes to capture competing values through deliberative processes are

[156] Margot Kaminski, 'Regulating the Risks of AI' (2022) 103 *Boston University Law Review* (forthcoming).
[157] Margot Kaminski, 'The Developing Law of AI Regulation: A Turn to Risk Regulation', *Lawfare*, (Web Page, 21 April 2023).

already known in the law, as in records of decisions by courts and tribunals.[158] Furthermore, the need for balancing is already mentioned in the existing NSW AI Assurance Framework.[159]

Some possible advantages of a formalised balancing test approach may include:

- Structuring/listing factors for consideration which must (or must not) be taken into account (with respect to which a process of public consultation may be informative with respect to community concerns and values);
- Requiring decision makers to articulate active priorities of a project against residual risks in a more nuanced manner than a standard impact assessment, supporting sound decision making by requiring record-making with respect to potentially subjective elements of a decision (making decisions more transparent, accountable and reviewable);
- Permitting a degree of flexibility in decision making processes, which may (i) be capable of responding to new technology (by balancing factors of public interest rather than specific modes of technological application) and (ii) provide data for future legislative update (for example, potentially revealing community priorities and values through cautious 'sandboxing' in deliberative processes and monitoring considerations most subject to administrative review); and
- Permitting cautious innovation in a way which is tethered to considerations of public interest and stated principles.

Consultation question 15: addressing some limitations of a risk-based approach

Consultation Question

15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

ADM+S Answer: The main potential benefits of a risk-based approach are (a) the ability to avoid or mitigate harms before they happen, at the design and development stage rather than waiting for *ex post* litigation; (b) promoting better (safer, more responsible) design; as well as incorporating (c) ongoing obligations on developers of systems to engage in monitoring and addressing risks.

A core question is **who decides** whether a system is low, medium, high (or very high?) risk? Risk is multi-dimensional (it varies by type of impact/harm, severity and probability, and can shift over time) meaning that fixed categories may not work well, but the party best placed to assess the risk of a system (who could be the developer, or the deployer) may have incentives to underestimate risk. There may be mechanisms to manage this, including for example by requiring publication of risk assessments for at least some systems, and/or setting 'default' categories with the ability of entities to show that their system is lower risk than the default would suggest.

[158] This is also known in the administrative law of NSW, such as the public interest balancing test in sections 12 to 15 of the *Government Information (Public Access) Act 2009* (NSW) ('GIPA Act'). The ALRC previously made a similar recommendation for a public interest balancing test in Recommendation 9-1 of its *Serious Invasions of Privacy in the Digital Era* (ALRC Report 123) (2016).

[159] NSW AI Assurance Framework (n 116) 11, 'Evaluating AI benefits and risks'. Although the deliberative process we describe requires more guidance and requirements for responsible record keeping.

What is a risk-based approach, and who decides which risk category a system falls within?

Assuming that a risk-based approach is to be adopted, it is necessary to understand what that means.

The EU Design

In the EU AI Act, for example, 'risk' plays two distinct roles:

1. **Established allocation of use cases to risk categories:** The Act embodies (legislative) judgments about which use cases are low or minimal, limited, high or unacceptable risk, with different rules applying at each level and with the most regulatory attention paid to the management of high-risk systems. Article 7 empowers the Commission to update the list of high-risk systems.[160]

2. **A firm-level obligation to apply risk management:** for developers/deployers of systems, where these firms are required to undertake risk assessment; identify certain kinds of risks, and adopt measures to mitigate them (Art 9).

One criticism of this design is that the designation of certain broad use cases (or 'use areas') as 'high-risk' does not follow risk management practice. Risk is multi-dimensional; common risk management practice would therefore dictate that both in assessing risk, and designing its mitigation, a firm should assess:

- The nature (and distribution) of any risks;
- In the case of harms, the severity of those harms (and their distribution); and
- The likelihood of those harms occurring.

Risk can also vary over time. Unless these dimensions are taken into account, a heavy regulatory burden may be applied to *all* systems within a use area, including, potentially, relatively low-risk or positively beneficial systems.[161] This could have a negative impact on innovation and take up of AI, especially by smaller and/or risk-averse entities (such as public sector entities). Not *all* AI systems that use AI in the overall processing of student exam results would necessarily be considered high risk, although ones that automatically assign a grade would be. The variability of risks within use cases has been recognised to some extent in recent amendments to the draft EU AI Act, in that additional language has been added that indicates that both baseline standards required of high-risk systems, and mitigations should be proportional to the features and use of the system.

Alternative designs

An alternative design for a risk-based system (as seems to be suggested in the Discussion Paper) allows **organisations** to self-assess the risk level of their system, taking into account the multiple dimensions noted above. The problem however is that there can be incentives to understate or miscategorise systems as being lower risk than they in fact are, especially if it means avoiding heavy regulatory burdens. The EU design avoids this by designating high-risk use cases but allowing the organisation to undertake internal risk management and then self-

[160] As necessary and upon consideration of a range of factors in consultation with the AI Office (a new EU body to support the harmonised application of the AI Act (see Art 7).
[161] This has been recognised to some extent in recent amendments to the proposed EU AI Act (n 70), in that additional language has been added that indicates that the baseline standards should be tempered in light of the particular features and use of the system.

assess whether their system meets the regulatory requirements - including by designing their own mitigations and determining the level of residual risk that is acceptable.

There is a range of possible intermediate mechanisms that could avoid the semi-rigidity of categories pre-determined by government, without leaving risk assessment entirely to developers/deployers:

- Legislation or regulation (probably regulation, to retain some flexibility as use cases arise) could identify certain use cases as 'higher risk by default' - but allow organisations to undertake (and retain for potential audit or future dispute) a risk assessment that justifies a lower rating.
- The legislation could require organisations to retain (or perhaps in the public sector, publish in a register) some description of the system with its classification (and reasons for that classification), endorsed by a designated responsible person within the organisation. This could provide some assurance that thought has been given to the question and provide a record and enable responsibility to be assessed *ex post* in the event of harms. Again, this could potentially be confined to designated classes of users, or use cases.

Relevance of earlier discussions

In discussing Consultation Question 3, we noted several emerging areas of technological development and potential risks which illustrated how contingent is much of our knowledge of how newer AI systems work. In identifying what kinds of uses of technology are low, medium, or high risk, it will be critical to bring knowledge from a range of perspectives: both technical and non-technical. The need to ensure cutting edge knowledge is made more widely available is something we noted above.

In discussing Consultation Question 14, we also noted the importance of both socio-technical approaches to risk assessment, *and* the need for diverse perspectives. When thinking about *who gets to decide* that a system is low, medium, or high (or very high?) risk, we suggest that whether though providing guidance, connecting researchers or, in larger organisations insisting on multidisciplinary and diverse teams, these issues must be part of the way we consider what precautions are needed around a proposed use of AI.

Consultation question 17: elements of a risk-based approach

Consultation Question

17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

ADM+S Answer: ADM+S makes a number of comments on elements of the risk-based approach set out in the Discussion Paper at pages 39-40:

- We suggest that three 'categories' of risk may be insufficient, and that descriptions of the different categories could give rise to some anomalous results (such as where a risk is brief *and* severe, or where it is brief and 'reversible' (for example, as a loss of social benefits is

'reversible' as payments can be restored) but has lasting impacts (say because a person has become homeless in the meantime when they could not pay their rent);
- We suggest that more guidance and deeper thinking will be required regarding the 'risks' that must be considered; some ideas are offered and we draw attention to our discussion of environmental risks (Consultation Question 2) and issues around digital exclusion in Australia (Consultation Question 14);
- In terms of the requirements listed at page 40, we draw attention to the absence of any reference to data quality considerations, suggest that further thinking is required regarding notices/transparency/explainability, and note that the appropriate role of human oversight ('human in the loop') is complex – indeed including human oversight can sometimes increase, or obscure, problems with an AI system.

What is high, medium, or low risk?

The Discussion Paper distinguishes between three categories of risk:[162]

1. Low risk: systems with minor impacts that are limited, reversible or brief;
2. Medium risk: systems with high impacts that are ongoing and difficult to reverse
3. High risk: very high impacts that are systemic, irreversible or perpetual.

One comment on those categories concerns the descriptions. It is unclear whether, for example, a severe but brief risk appropriately counts as 'low' (as appears to flow from the way the categories are defined). Or indeed whether systems with high, difficult to reverse impacts should be seen as 'medium'. The (indicative only) examples in the table communicate the message that health and safety concerns put a system into the high-risk category. Based on research by ADM+S and beyond, hard to reverse harms to well-being and opportunities, due to people not receiving services they need and are entitled to can also lead to lasting, devastating impacts that are difficult to reverse in their effects, even once the cause is removed.

We would also suggest that the 'optics' of categorising as medium risk systems that the EU would classify as high-risk is problematic. We further note that the Canadian Directive on Automated Decision-Making recognises four categories (low, moderate, high and very high) which might be a more useful model if categories of this kind are to be used.

Which risks?

We think guidance will be needed across a range of contexts as to the risks that need to be considered, not least because not all potential impacts of AI/ADM systems are obvious to all people. Not all people presented with a new AI system will be fully aware of the potential biases it may incorporate.[163]

Any guidance provided will need to be under almost constant review: understanding of the potential impacts of the use of AI is developing all the time (as indeed our discussion under Consultation Question 3 highlighted).

Figure 1 below outlines a range of risks, harms, and impacts of AI across different stages of the lifecycle of an AI system (based on, but not exclusive to, Generative AI models) - many of which ADM+S researchers have investigated. The diagram captures a key point: that there are a wide range of risks and impacts arising from the development and use of AI, not all of which will be

[162] Department of Industry, Science and Resources, *Safe and responsible AI in Australia* (Discussion Paper, June 2023) 32-33.

[163] For example, the Commonwealth Ombudsman's Best Practice Guide on Automated Decision-Making has a detailed checklist; the NSW AI Assurance Framework also sets out a series of questions that a government entity should consider.

relevant in all circumstances. Which risks, which *kinds* of risk fall within the system, and which are relevant in given contexts, requires thought, and guidance.

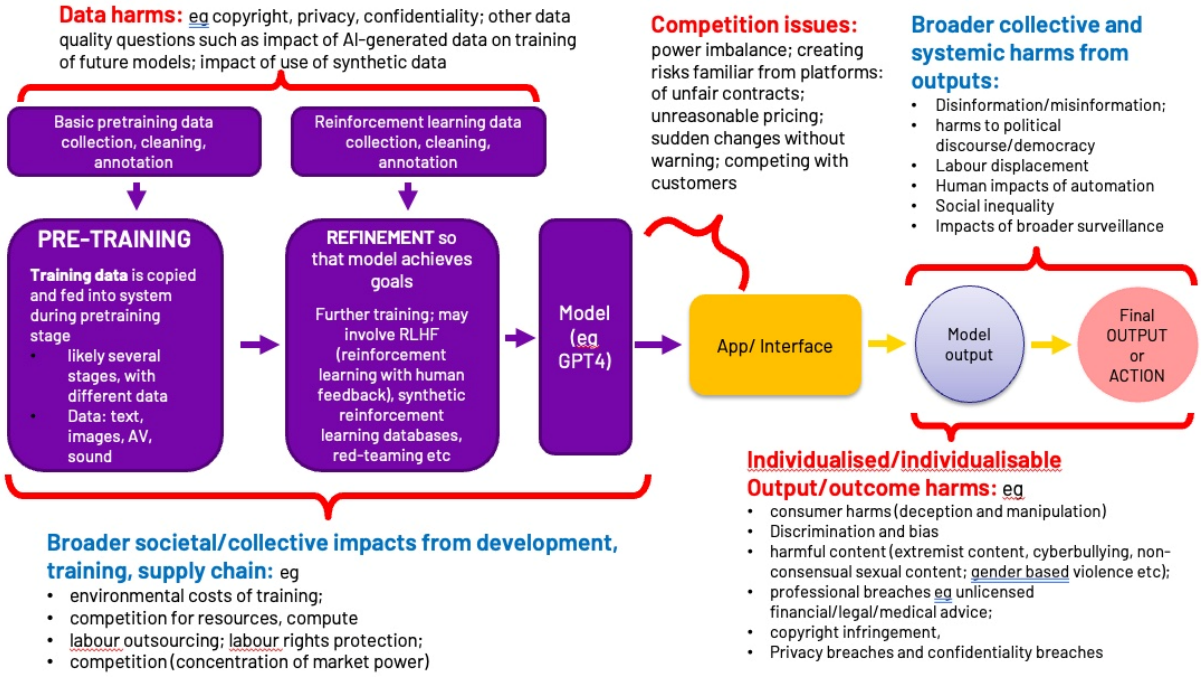

Figure 1. Diagram of AI impacts[164]

As the Discussion Paper recognises, high-level guidance (for example that talks about bias concerns, and privacy) will be of limited assistance. The OECD AI Framework for the Classification of AI Systems[165] and ADM+S' framework on ADM systems[166] emphasise that to conceptualise AI/ADM and its impact requires an understanding of the context in which AI is developed and deployed. ADM+S undertakes detailed empirical work investigating the uses and impacts of ADM in specific domains, working with our network of partners in the media, mobility, social services, disability and health sectors.

For example, at one end of the regulatory spectrum, AI is advancing in healthcare, raising complex questions of interactions with medical professional regulation and institutional risk management, creating uncertainty about how AI and data can be used responsibly, say, to assist in patient notes or recommendations for treatment pathways. This uncertainty can block adoption of even promising technology. At the other end, local governments face very new issues when they adopt systems for road maintenance, automating the identification of dumped rubbish, sign damage and other forms of routine monitoring, and are having to develop new governance practices to manage data and AI systems responsibly. Social service and community sector organisations working with sensitive client data are working on how to ensure AI tools do not breach data consent and privacy arrangements, and having to deal with how their obligations of confidentiality intersect with the policies and business models of commercial providers of technology.[167]

[164] Kimberlee Weatherall, 'Regulating (Generative) AI', (Presentation, Gradient Institute, 7 June 2023). Enlarged version in Appendix 2.
[165] OECD, 'OECD Framework for the Classification of AI systems' (2022) 323 *OECD Digital Economy Papers*, No. 323.
[166] ADM+S, Brooke Ann Coco, Paul Henman and Lyndal Sleep, *Mapping ADM in Australian social services* (Report, 15 Oct 2022).
[167] Xiaofang Yao et al, Swinburne University, *Building Data Capacity in the Not-For-Profit Sector* (Interim Report, 2021).

How best to provide the necessary assistance to a wide range of actors and a wide range of sectors will need to involve industry-specific bodies and domain regulators, but it also requires deeper engagement: a whole of society educative effort but also ongoing support. ADM+S would be happy to assist based on our experience working with a range of smaller and community actors around the adoption of these technologies.

We draw attention here too to our discussion earlier regarding the **environmental** risks associated with the use of AI (see Consultation Question 2) and **digital exclusion** (Consultation Question 14), both of which ought to be considered as raising distinct issues necessary to consider as part of any risk assessment process.

Requirements for systems at different levels of risk

The Discussion Paper suggests certain safeguards for different systems. We suggest that data quality, at least, appears to be missing; we also include comments below regarding the proposals regarding **notice/explanations,** as well as the idea of including human oversight, or a 'human in the loop', based on research conducted at ADM+S.

The need to consider data quality

Data quality is an important consideration in ensuring responsible use of AI and automation. The datasets used to train AI models have context that can determine the veracity or bias of AI models. In the EU AI Act, data quality is a baseline requirement for high-risk models, but we note that it was not included in the considerations for medium or high-risk systems in the Discussion Paper. While there are some data quality provisions in the *Privacy Act 1988* (Cth), these only apply to personal data and are in any event limited in their reach. We suggest that an assurance of the relevance and sufficient quality of data, as well as considerations around data governance should be a consideration in any AI governance system. For a further discussion of data, see above under Consultation Question 2 ('Regulatory gaps at the training stage').

Notices and Explanations

The Discussion Paper requirements for notices and explanations are, we presume, derived from an interest in implementing concepts of transparency and explainability. There are some weaknesses in these requirements:

1. The use of new terms has the potential to create confusion (given existing discourse and discussion around transparency and explainability). This does not mean the new terms are necessarily inappropriate, but it does mean that some explanation is warranted as to why new terms are required, and what they are intended to encompass (and exclude) from existing systems including the Commonwealth's AI Ethics Principles. We do note, however, that the AI Ethics Principles can be criticised for focusing on informing people that they engaging with AI (a kind of transparency) and having insufficient emphasis on the need to ensure that people can understand how the system works or how decisions were made (explanation).

2. The requirements as currently defined are much too narrow. E.g. 'notice' merely requires information be provided that AI is involved. As discussed above under Consultation Question 9, stakeholders would need more information than this (and the kind of information required may vary by context).

3. We query why low-risk AI does not need any notice at all; developing rules and policies globally generally suggest that notice is a baseline requirement.

4. There are related problems with prescriptions around explainability. For example, it is unclear what the different terms used in the Consultation Document, including 'general explainability' and 'system explainability', actually require.

We note too the need for more research on these questions. It is as yet unclear what types of explainability are appropriate for different types of decisions, or whether, with the current state of the art, explainability offers something truly meaningful to decision-makers or decision-subjects,[168] or simply whether it distracts from more meaningful interrogation as to why a decision was made.[169]

Human in the loop as a mechanism for addressing AI risk[170]

The Discussion Paper suggests that including a 'Human in the loop' is an appropriate oversight mechanism and risk mitigator for medium and high-risk systems. Human oversight requirements are common in AI and ADM regulations. A human in the loop has become a core dimension of 'human-centred' design and governance strategies addressing the ways automation exacerbates issues of transparency, accuracy, and accountability.

But frequently, the human in the loop is more a *symbol* of good governance than a provably beneficial regulatory intervention: it can make people less anxious about automation, but without actually improving outcomes.[171]

A number of studies have challenged the utility of a human in the loop for improving various types of decisions.[172] This is not to say that technical systems make better decisions without human oversight, but rather that there is real complexity in how human interaction with AI systems works. Interaction at the human-computer interface can be influenced by behavioural biases, organisational arrangements, and contextual realities, all of which will affect decision quality and performance. Research shows that as automated tasks become more complex, the role of monitoring and overseeing those processes simultaneously becomes more difficult.[173] When it comes to human oversight of automated systems, there are no clear best practices, and merely mandating a human act as the final decision-maker may cause more problems than it solves.

This highlights the need to consider, in context, whether inserting a human into the loop is going to be effective as a risk mitigation tool. Potential problems with uncritical adoption of human oversight are:

1. It can *increase* the risk of harm to people, property, and environment: eg in the case of automated manoeuvring systems for spacecraft/satellites when executing unplanned manoeuvres in the event of encountering space debris, where the lag time caused by waiting for human input to travel over vast distances would make the system less responsive to evolving circumstances. The same argument can be applied to self-driving cars. In these situations, human oversight won't be effective to avoid risks,

---

[168] Zachary C Lipton, 'The Mythos of Model Interpretability' (2017) *arXiv:1606.03490*.
[169] For opportunities and problems related to AI explainability in different contexts see, eg, Rita Matulionyte et al, 'Should AI Medical Devices be Explainable?' (2022) 30(2) *International Journal of Law and Information Technology*, 151-180; Rita Matulionyte and A. Hanif, 'A Call for More Explainable AI in Law Enforcement' (2022) IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW) 75-80.
[170] Dr Jake Goldenfein (University of Melbourne) is leading an ADM+S-wide project on the concept and implementation of the 'human in the loop', and can provide further information as required.
[171] See, eg, Yuk Hui, 'ChatGPT, or the Eschatology of Machines', (2023) 137 *E-Flux*.
[172] See, eg, Ben Green, 'The Flaws of Policies Requiring Human Oversight of Government Algorithms', (2022) 45 *Computer Law and Security Review*; Ben Wagner, 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11(1) Policy and Internet 104.
[173] Lisanne Bainbridge, 'Ironies of Automation' (1983) 19(6) *Automatica* 775-779. This emerges from research across disciplines such as industrial psychology and human factors research.

more rigorous testing of models and their behaviour under a variety of operating conditions is necessary.

2. Prescribing a human in the loop can have unintended consequences: eg the majority judgment in *Pintarich v Deputy Commissioner of Taxation* [2018] FCAFC 79, set a legal standard that requires human cognitive processes before there is a decision capable of grounding judicial review. The effect was to undermine administrative certainty and to create an accountability gap, allowing authorities to recant the (non)decision.
3. It could distract attention from faults in the system:[174] eg, if the human 'final decision' can be reviewed and contested, this might mean less attention paid to the data-driven aspect of the decision. This is especially problematic for risk profiling systems where which feed into human decisions. Focusing only on the reviewability of human decisions risks preventing us from developing an understanding of how automated elements contribute.

Even the Australian Robodebt debacle can be understood through a human in the loop lens. Critics of the system often frame the problem of Robodebt as the absence of a human in the loop, placing the blame at the door of an automated system that removed the human decision-maker. In a strange mirror of the *Pintarich* case, focusing on the absent human decision-maker allowed political actors to narrate the mistake as technical in nature. As the Royal Commission highlighted, however, the source of the issue was not technical, but the broader institutional policy of reversing the onus of debt verification in order to scale a punitive debt-recovery program.

There have been numerous efforts to identify ways to leverage the particular qualities of humans and machines to produce hybrid decision systems.[175] But these 'Men are Better at - Machines are Better at' (MABA-MABA) approaches fail to appreciate that specifying the appropriate role for a human in a decision system is as much a political question as it is a question of what people are physically and psychologically capable of.

In summary, 'human oversight' is not in itself a mechanism for ensuring safe and responsible AI. It could be, in the right circumstances, with the right understanding of the appropriate allocation of tasks and responsibilities – this is why, we would argue, regulation should be aimed at the creation of safe and responsible automated decision-making *systems*, regardless of the technology they use.[176]

For better human-computer collaboration and decision-making, we need deeper understanding of how humans and automated systems make decisions together, *and* serious engagement with the questions of responsibility that follow. We need to improve our understanding of the ways that automated systems and AI distribute cognitive tasks and decision processes, and reconfigure accountability throughout technical systems and design processes. We need better knowledge on how humans embedded in technological decision processes use and respond to different types of explanation. This requires ongoing research, which we are presently pursuing at ADM+S.

---

[174] See, eg, the consideration of the legality of SyRI in the Netherlands: <https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:RBDHA:2020:1878>.

[175] See, eg, Marion Oswald, 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power' (2018) 376(2128) *Philosophical Transactions of the Royal Society A;* Ben Wagner, 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11(1) Policy and Internet 104.

[176] To that end, an earlier draft of the EU AI Act (n 70) required that the providers of decision systems specify organisational measures guiding the function of human overseers as part of a systems validation regime. However, this has been removed in the current draft which requires only user discretion in implementing human oversight measures.

Consultation question 19: Foundation models

Consultation Question

19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

ADM+S Answer: Much of our earlier discussion is relevant to Foundation Models: see in particular our discussion under Consultation Question 2 (what is new/different about AI; what new capacities require a societal-level discussion); and Question 4 (non-regulatory actions required, in particular the need to connect government with cutting-edge research and ensure new research is incorporated into efforts to guide and educate to developers and deployers as well as the broader public).

There are concerns about the applicability of risk-based approaches in relation to foundation models, which the EU is presently grappling with. The submission discusses these developments, and ADM+S can offer further expertise as required and as regulatory positions consolidate internationally.

Finally, we note that foundation models raise genuine questions around the consolidation of power over the generation and transfer of knowledge. Steps may need to be taken to ensure research and pedagogical access for Australian researchers; it would be detrimental, for example, to the country's research efforts if researchers from certain countries where models are developed gained preferential access for the purposes of research.

General comments

Much of the discussion throughout this submission is applicable to foundation models/LLMs/MFMs. We note in particular:

- Our discussion under Consultation Question 2, which talks about what is new or different about recent developments in AI, drawing attention to the broader impact of systems being developed which 'feel more human' and so raise questions for multiple legal frameworks;
- Our discussion under Consultation Question 2, regarding new capacities requiring a public discussion;
- Our discussion under Consultation Question 2 regarding regulatory gaps at the training stage;
- Our discussion under Consultation Question 4 regarding non-regulatory actions, including in particular the need to connect government with cutting-edge research, which is more important in light of ongoing development of LLMs/MFMs;
- Our similar comments under Consultation Question 4 regarding coordination of AI governance across government: in a similar vein, precisely because our understanding of larger models is changing all the time, there is a need to connect cutting-edge research and government.

Nevertheless, some additional points are worth noting. First, we refer to the separate submission by the Gradient Institute which considers frontier models and which discusses the management of very large models.

Second, we note that a risk-based regulation is not simply transferable to the largest models:

1. Data and data quality is harder to police given the vast amounts of data involved;

2. Risk management approaches depend on trying to foresee risks, and then trying to mitigate or manage them. However, it seems, so far, that not all of the capacities and outcomes of the more powerful models *are* predictable, even to experts; risks may only

be foreseeable (and able to be mitigated) in a use context; and uses that are low-risk today can become high-risk tomorrow when model capabilities are incrementally expanded. The reverse is also true: AI-use cases that 'upset' today may become more widely accepted over time.

Fundamentally, in the context of these larger models, no-one in the value chain has sufficient information or capacity to respond to all the risks, giving rise to a significant challenge of co-ordination. Risk management, if it proceeds throughout the lifecycle will need a mechanism for aggregating information about downstream impacts of foundation models, and then distributing relevant info and responsibilities to the relevant people.

Developments in the European Union

The EU is currently facing the question of how to approach 'Foundation models', which were not included in the initial proposal from the Commission. Essentially, the two co-legislators (Parliament and Council, which need to agree for the proposal to become a regulation) came to the table with different views. For the Council of the EU (the body that is constituted of representatives from Member States), foundation or general-purpose models, should be excluded from the AI Act, and their regulation undertaken at a later stage in a more targeted regulatory tool, led by the European Commission. The logic was that the Act should focus on assessable risks generated by specific-purpose oriented systems. The European Parliament perspective was different: in this (EP) version, foundation models are brought within the scope of the AI Act, as part of an extension of the scope to cover general principles for all AI systems (those with specific intended purpose, but also those with general purpose or purely foundational).

The EP proposes amendments targeting two dimensions of foundation models. First, when provided as a service, such as through API access. In this case the obligation is one of cooperation with downstream providers, in order to enable them to implement appropriate risk mitigation all through the lifecycle of the system. Developers of foundation models will be exempted from this obligation if they transfer enough information to the downstream provider, so the latter is able to fully comply with the Regulation independently.

The second refers to the development of the model itself. Developers should assess and mitigate possible risks and harms through appropriate design, testing and analysis, should implement data governance measures, including assessment of biases, and should comply with technical design requirements to ensure appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity. Even models provided under free and open-source licences are not exempted from these obligations (Art 2. 5e). For generative foundation models, they should ensure content is generated by an AI system transparently. Finally, all foundation models should also be bound by environmental standards. How to implement these obligations is not clear. The EP amendments only refer to the need to develop new standards in this domain (recital 60h), but in the discussions, it was noted that this may require independent 3rd party assessments.

While we cannot comment on the final shape of EU law, we can offer expertise on the process and provisions on an ongoing basis where that is helpful. ADM+S has extensive connections with equivalent research entities in the EU.

Knowledge lockout and monopolisation concerns

A final concern is that current state-of-the-art models require significant investments to build and operate, elevating those with access to venture capital, large-scale computing and high-skilled workers into greater positions of knowledge brokerage over those without. The success

of these models rests on combining large existing knowledge repositories with a user-feedback loop where millions of user prompts are integrated into later training datasets to tune a model's responses. In this environment, first-movers have a natural advantage over later entrants due to increased model training times and data scope, able to consolidate large swathes of external and user-generated knowledge.

Such model and knowledge monopolisation makes it exceedingly difficult for later entrants to compete, affecting general model diversity and the availability of comparable alternate offerings. This has a number of effects:

- A primary effect from this is the consolidation of knowledge authority. Instead of 'Googling', we would prompt a knowledgeable model, the returned information similarly at risk of ranking, manipulation and obfuscation.
- The secondary effect from this is models becoming even more tightly guarded than today, as their unique training weights and model architectures increasingly set them apart from the competition.

This raises a question for policymakers: to what degree do we allow a small concentration of powerful commercial entities to become arbiters of truth? From a research and education perspective, how do we prevent citizens, educators, non-profits and other vested parties from getting locked out of the learning and knowledge transfer cycle? We may need to consider whether there will need to be pedagogical and research provisions in place for Australian researchers, that guarantee that these systems, as well as how they operate in a socio-technical context, can be studied. Questions of transparency and observability have been discussed in more detail under Consultation Question 9.

Appendix 1: Additional countries

Japan

Japan has not yet developed a comprehensive AI-specific legislative regime as is being attempted in other jurisdictions, and for now appears committed to a soft-law approach. Key policy documents include:

- *Social Principles of Human-Centric AI* (2019)
- *Governance Guidelines for Implementation of AI Principles* (Ver 1.1, 2022)
- *Governance Innovation: A Guide to Designing and Implementing Agile Governance* (Ver 2.0, 2021)

As a key trading partner with Australia and a world-leading exporter of robotics, Japan is a highly relevant jurisdiction for Australia to watch, particularly with respect to general approach (such as agile governance and innovation goals) and potentially transferable soft law and social support efforts (such as AI education/literacy goals).[177]

India

Developments in India are also relevant to Australia. India is Australia's sixth largest trading partner, chair of the Global Partnership on Artificial Intelligence, and President of the G20 where technological transformation is a key agenda item. Australia and India are currently negotiating a Comprehensive Economic Cooperation Agreement (CECA) aimed at further expanding bi-lateral economic cooperation in which removing barriers to trade in the AI industry is a key issue.

India is yet to adopt binding legislation on AI, but has recently announced the *Digital India Act 2023* which will include AI regulation. The proposal includes several interesting features including the creation of an adjudicatory mechanism for online civil and criminal offences; specific regulation for 'addictive tech' especially as it impacts minors;[178] and separate rules for specific classes of 'intermediaries' such as eCommerce platforms and search engines. India is concurrently developing its data protection regime through the *Digital Data Protection Bill 2022* with an interesting mechanism for consent of data subjects: 'Consent Managers'.

Further, Australia could also consider India's 'Draft Standard on Fairness Assessment and Rating of Artificial Intelligence Systems' developed by the Department of Telecommunication which is a voluntary framework seeking to standardise bias assessment amongst AI developers.

[177] See also above discussion, *Consultation Question 3: Non-Regulatory Actions to support responsible AI practices in Australia*.

[178] Chinese regulation on Generative AI is requiring providers to take effective measures to "prevent minor users from excessively relying on or indulging in generative AI systems". Interim Measures for the Management of Generative Artificial Intelligence Services, Cyberspace Administration of China, Order No 15, 10 July 2023, art 10.

Appendix 2: Diagram of AI Impacts