26 July 2023

Department of Industry, Science and Resources
By web: https://consult.industry.gov.au/supporting-responsible-ai/submission

Supporting Responsible AI: Discussion Paper

About us

The **UNSW Allens Hub for Technology, Law and Innovation** ('UNSW Allens Hub') is an independent community of scholars based at UNSW Sydney. As a partnership between Allens and UNSW Law and Justice, the Hub aims to add depth to research on the diverse interactions among technology, law, and society. The partnership enriches academic and policy debates and drives considered reform of law and practice through engagement with the legal profession, the judiciary, government, industry, civil society and the broader community. More information about the UNSW Allens Hub can be found at http://www.allenshub.unsw.edu.au/.

The **UNSW Business School Regulatory Laboratory** ('RegLab') is a community of researchers examining regulation and governance in the UNSW Business School. Reg Lab is a transdisciplinary lab examining the challenges faced by regulators and the regulated in the context of rapidly changing business models. It has a focus on the networked industries sector and data driven innovation. It is jointly funded by the UNSW Business School and external partners (primarily, Google but supplemented by research funding by the Commonwealth).

We have joined forces for the purposes of preparing this submission. We are also affiliated with the UNSW AI Institute who are preparing a separate submission. We encourage the Department to look closely at their submission as well, which focuses on some of the more technical issues.

About this Submission

We are grateful for the opportunity to make a submission on the Discussion Paper *Safe and responsible AI in Australia*. Our submission reflects our views as researchers; they are not an institutional position. This submission can be made public.

We focus on those questions within our collective expertise. Our main points relate to:

- Whether there is a need to define artificial intelligence as such, and tendency of any such definition to obsolesce (with consequences for using the concept in law or regulatory instruments).
- Proposals for law reform (including the possibility of a law reform commission review) in the areas of consumer protection, discrimination law, administrative law and privacy law.
- Support for standards development.
- The need for guidelines for government procurement of AI systems.
- The need for better co-ordination for policy development across government.

- The advantages of the Swiss over the EU approach to law reform in response to AI.
- The importance of different approaches to public and private sector uses of AI, particularly in the context of accountability and transparency, and some suggestions for the government's own commitment to responsible AI.
- The possibility of general "generic" laws being supplemented by technology-specific guidance, perhaps pointing to international standards.
- Some suggested contexts in which AI 'bans' should be considered.
- A proposed redirection of focus from seeking 'trust' to ensuring 'trustworthiness'.
- Some suggested characteristics for a risk-based approach.

We also wish to highlight the following points from the UNSW AI Institute submission:

- The importance of expertise within relevant agencies and regulators
- The need for special consideration of legal and regulatory responses where safety or important values are at stake.
- The importance of education, particularly critical thinking about AI and its capabilities and limitations.

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

The question about whether any particular definition is optimal depends on the purposes of the definition exercise. In our view, the definitions proffered are not problematic *per se* (although they introduce distinctions that differ from the approach of the International Organisation for Standardisation or ISO). Our primary point is that the definition of "artificial intelligence" does not correspond with the optimal scope of regulatory action.

First, the definition itself. The discussion paper indicates that it bases its definitions on those of the International Organisation for Standardisation (ISO). There are noticeable differences including in the crucial definition of artificial intelligence. For example, the discussion paper's definition would not include expert systems as a species of artificial intelligence. The submission from the UNSW AI Institute also highlights other issues with the proffered definition of AI.

This is not necessarily problematic because there is no single or optimal definition of artificial intelligence. It depends on purpose and context. One might, for example, have one definition for describing a university course on artificial intelligence and another when defining the scope of regulation. It is the latter that is most crucial here. In other words, defining artificial intelligence *in government* means asking whether there are particular risks and harms associated with a particular kind of system and, if so, how that kind of system ought to be described for the purposes of legal and regulatory instruments.

That kind of exercise is less useful here because the kinds of harms commonly associated with artificial intelligence are rarely limited to a particular type of technology. Risks associated with lack of accountability can be found in the use of systems relying on explicit programming; a good example is Robodebt. Conversely, problems associated with data-driven decision-making, such as discrimination and unfairness, need not involve an engineered system at all. There are plenty of examples where organisations have relied on traditional statistics or even stereotypes to justify decisions that have disparate impact on certain groups.

Our point here is not that the definition proffered is per se wrong, but rather that the "define a technology and regulate it" approach is only useful where the problem being dealt with aligns with

the specific technology. That can sometimes occur, for example with the decision of many governments to prohibit human reproductive cloning. Australia did not prohibit the existence of human clones; that would have involved the sacrifice of one of each identical twin pair. Rather, we prohibited a particular technological means of producing human clones because of ethical concerns around those techniques. That set of prohibited techniques thus needed to be defined. Similarly, a definition of artificial intelligence is only useful in a legal or regulatory context where it aligns with the problem that is being addressed.

While the report identifies various challenges associated with “artificial intelligence”, none of these align in scope with the definition given for artificial intelligence. Fake photographs are an old problem; so-called deep fakes are simply an example of creative technologies running ahead of detection tools. Misinformation and disinformation can be authored by humans and propagated through networks and, while artificial intelligence can accelerate generation and target propagation, it is not a necessary ingredient. Encouraging people to self-harm can be done at scale using explicit programming, say outputting "go kill yourself" whenever particular words are used in the input. Inaccuracies can be propagated at scale with explicit programming as demonstrated by Robodebt. Bias is as evident in some statistical techniques as in some machine learning techniques. A badly programmed expert system can yield false answers just like Chat GPT. Those procuring any complex system need a degree of transparency as to how it operates; this problem is not unique to artificial intelligence and, even where information about a system could be communicated, most organisations choose to rely on trade secrets or commercial-in-confidence arrangements. Finally, when people want to know if a decision affecting them is made using artificial intelligence technology, they probably mean something broader than the definition in the discussion paper, including decisions based on explicit programming.

If regulators address these identified risks and harms only in contexts where “AI systems” are used, the resulting framework will not only be fragile in the event of ongoing technological change, it will fail to deal with problems we are already facing today.

The discussion paper is correct that AI (using the paper's definition and most others) rightly generates calls for regulatory action. But that action need not be technology-specific. The question to be answered is not “how do we regulate AI” but rather “how do we ensure that our laws operate appropriately and effectively to achieve policy objectives, including in contexts involving AI”. There is no need to define “artificial intelligence” in order to address the issues associated with a diverse range of technological practices. Instead, most problems identified are better addressed through a program to reform and update privacy and discrimination legislation, consumer law, administrative law, and so forth, so that they operate to achieve their goals when applied to current practices associated with the broad frame of artificial intelligence. Some specific suggestions in this regard are set out in response to Question 2 below.

Further analysis of the challenges of technology-specific approaches to regulation can be found in:

Bennett Moses LK, 2017, 'Regulating in the Face of Socio-Technical Change', in Brownsword R; Yeung K; Scotford E (ed.), Oxford Handbook of the Law and Regulation of Technology, Oxford University Press, Oxford

Bennett Moses LK, 2013, 'How to Think about Law, Regulation and Technology: Problems with "Technology" as a Regulatory Target', Law, Innovation and Technology, 5, pp. 1 - 20, http://dx.doi.org/10.5235/17579961.5.1.1

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?

There are a number of potential risks from artificial intelligence, or particular kinds of artificial intelligence, that are not well captured by existing law. This section includes some examples, but a full audit of Australian law and its application in this area would be required to ensure a more comprehensive list.

A. Consumer law

A.1 Consumer protection in financial services

UNSW Allens Hub Senior Research Fellow, Dr Kayleen Manwaring, has recently been involved in a significant research project investigating potential harms to consumers arising from the growing use of AI-related applications in financial services (particularly insurance) and how Australia's current laws apply to these harms. The research project found that these harms range across a number of subject areas, such as discrimination, privacy breaches, digital consumer manipulation and financial exclusion.

Additionally, this research project has identified significant gaps in relevant regulatory regimes relating to these harms, such as the *Insurance Contracts Act 1984* (Cth), the General Insurance Code of Practice (2020), *Corporations Act 2001* (Cth), *Australian Securities and Investment Commission Act* 2001 (Cth) the *Privacy Act 1988* (Cth), and state and Commonwealth anti-discrimination legislation. The research project has also suggested some principles for reform.

The most relevant outputs of the project relating to this inquiry are:

> Bednarz Z; Manwaring K, 2022, 'Hidden depths: the effects of extrinsic data collection on consumer insurance contracts', 45 (July) *Computer Law and Security Review: the International Journal of Technology Law and Practice* 105667 http://dx.doi.org/10.1016/j.clsr.2022.105667
>
> Bednarz Z; Manwaring K, 2021, 'Keeping the (good) faith: implications of emerging technologies for consumer insurance contracts', 43(4) *The Sydney Law Review*, 455, http://www5.austlii.edu.au/au/journals/SydLawRw/2021/20.html
>
> Bednarz Z; Manwaring K, 2021, 'Insurance, Artificial Intelligence and Big Data: can provisions of Chapter 7 Corporations Act help address regulatory challenges brought about by new technologies?', 36 *Australian Journal of Corporate Law* 216

A.2 Digital consumer manipulation

UNSW Allens Hub members have completed significant research on exploitative and manipulative conduct by digital platforms and others providing digital services.[1] There is growing concern by

[1] Katharine Kemp, 'Concealed data practices and competition law: why privacy matters' (2020) 16(2-3) European Competition Journal 628; Kayleen Manwaring, 'Will emerging information technologies outpace consumer protection law? The case of digital consumer manipulation' (2018) 26(2) Competition and Consumer Law Journal 141.

scholars,[2] practitioners,[3] think tanks[4] and industry commentators[5] that the increase in electronic marketing and transactions, and the vast amount of data exposed to public scrutiny by ecommerce, social media, Internet-connected devices and environments and other online activities, may grant marketers a significantly increased capacity to predict consumer behaviour, and use data and behavioural research to exploit the biases, emotions and vulnerabilities of consumers.[6]

The ability of commercial entities to manipulate or exploit consumers is greatly enhanced by the use of AI-related technologies, such as machine learning. AI technologies such as machine learning are at the forefront of the significant amount of data analysis and inferencing required to predict the behaviour of consumers in any number of situations, and to be able to target them in real-time in specific ways, in particular emotional states, in such locations and at the times when manipulation is most likely to be successful.

These practices have been called 'digital consumer manipulation', that is:

> *the use of personalised consumer data collected, processed and/or disseminated by digital technologies, combined with insights from behavioural research, to exploit consumers' cognitive biases, emotions and/or individual vulnerabilities for commercial benefit*[7]

The commercial benefit firms may gain from such techniques include inducing disadvantageous purchases of products or services, extracting more personal information from consumers than is needed for the transaction, and engaging in unjustifiable price discrimination.

One example can be seen in a financial services context. Digital consumer manipulation in this industry often takes the form of 'margin optimisation', a 'process where firms adapt the margins they aim to earn on individual consumers'.[8] Even with most commercial entities' practice of

---

[2] Ryan Calo, 'Digital Market Manipulation' (2014) 82 George Washington Law Review 995; Eliza Mik, 'The Erosion of Autonomy in Online Consumer Transactions' (2016) 8 Law, Innovation and Technology 1; Natali Helberger, 'Profiling and Targeting Consumers in the Internet of Things: A New Challenge for Consumer Law' in Reiner Schulze and Dirk Staudenmayer (eds), *Digital Revolution: Challenges for Contract Law in Practice* (Hart Publishing 2016); Nancy S Kim, 'Two Alternate Visions of Contract Law in 2025' (2014) 52 Duquesne Law Review 303; Anthony Nadler and Lee McGuigan, 'An Impulse to Exploit: The Behavioral Turn in Data-Driven Marketing' (2018) 35 Critical Studies in Media Communication 151; Damian Clifford, 'Citizen-Consumers in a Personalised Galaxy: Emotion Influenced Decision-Making – A True Path to the Dark Side?' (CiTiP Working Paper 31/2017, KU Leuven Centre for IT & IP Law, submitted 15 September 2017), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3037425> accessed 30 April 2018.

[3] James Halliday and Rebekah Lam, 'Internet of Things: Just Hype or the Next Big Thing? Part II' (2016) 34 Communications Law Bulletin 4.7.

[4] Wolfie Christl, Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions (A Report by Cracked Labs, Vienna, June 2017); Nguyen and Solomon, Consumer data and the digital economy: emerging issues in data collection, use and sharing (n 646) 23–24.

[5] For example, Yael Grauer, 'Dark Patterns Are Designed to Trick You (And They're All Over the Web)' (arsTECHNICA, 28 July 2016) <http://arstechnica.com/security/2016/07/dark-patterns-are-designed-to-trick-you-and-theyre-all-over-the-web/> accessed 1 May 2018.

[6] Calo, 'Digital Market Manipulation' (n 2) 995ff; Kim, 'Two Alternate Visions of Contract Law in 2025' (n 2) 312; Helberger, 'Profiling and Targeting Consumers in the Internet of Things: A New Challenge for Consumer Law' (n 2) 140–61; Mik, 'The Erosion of Autonomy in Online Consumer Transactions' (n 2) 1ff; Halliday and Lam, 'Internet of Things: Just Hype or the Next Big Thing? Part II' (n 3) 7.

[7] Kayleen Manwaring, 'Surfing the third wave of computing: Consumer Contracting with eObjects in Australia' (PhD Thesis, University of New South Wales, 2019) 202.

[8] Financial Conduct Authority UK, 'General Insurance Pricing Practices: Interim Report' (Market Study MS18/1.2, October 2019) 21.

concealing their data-driven business practices where they can, some external evidence exists that EU, UK and US insurance firms, when setting prices, look at a consumer's willingness to pay based on their personal characteristics gained from the insights that external data provides.[9] Machine learning models and algorithms can be used to create inferences of price sensitivity and propensity for switching, based for example on the analysis of consumers' behaviour on a website or app controlled by the financial firm, the time an individual spends reading terms and conditions, or websites visited before applying to the financial services provider.

Another common example of digital consumer manipulation is so-called 'dark patterns', the design of user interfaces (such as ecommerce websites) to take advantage of certain behavioural biases. Behavioural biases are well-known psychological biases that can be exploited to make it difficult for consumers to select their actual preferences, or to manipulate consumers into taking certain actions that benefit the interface owner rather than the consumer. They are commonly used to manipulate consumers into paying for goods and services they do not need or want, or disclosing personal information that is unnecessary for the transaction and is used by the receiver for their own commercial purposes, or on-sold to third parties. Willis, in her seminal 2020 paper 'Deception by Design' details how machine learning and 'creative artificial intelligence' are used to optimise the effectiveness of the design and execution of dark patterns. Consumer experimentation can be executed much more quickly and at far greater scale with the use of AI, and website design can be both created and *personalised* by AI applications for micro-segments of consumers in response to the learnings from behavioural experimentation.[10]

Amazon has recently been targeted by the US Fair Trade Commission (FTC) for its use of these 'dark patterns'.[11] The FTC argues that these digital consumer manipulation techniques constitute unfair or misleading conduct in breach of section 5 of the Federal Trade Commission Act. The ACCC in its Digital Platform Services Inquiry and consumer advocates have recently identified them as serious issues for Australian consumers. Some of these 'dark patterns', while harmful to consumers, are not currently captured by the Australian Consumer Law (ACL). Patterns that are not misleading or deceptive (in breach of ss 18 and 29 of the ACL) can be unfairly manipulative in other ways, not currently prohibited under the ACL in the absence of an unfair trading practices prohibition.

Commentators have also raised the possibility of dark patterns fuelled by other features of AI. For example, AI applications or features designed to persuade consumers to:

1. believe that a particular sound, text, picture, video, or any sort of media is real/authentic when it was AI-generated (*false appearance*)
2. believe that a human is interacting with them when it's an AI-based system (*impersonation*).[12]

[9] Ibid; European Insurance and Occupational Pensions Authority (EIOPA), 'Big Data Analytics in Motor and Health Insurance: A Thematic Review' (Report, 2019) 12, 39.
[10] Lauren E Willis, 'Deception by Design', 34(1) Harvard Journal of Law and Technology 115.
[11] Federal Trade Commission, 'FTC Takes Action Against Amazon for Enrolling Consumers in Amazon Prime Without Consent and Sabotaging Their Attempts to Cancel' (Media Release, 21 June 2023), https://www.ftc.gov/news-events/news/press-releases/2023/06/ftc-takes-action-against-amazon-enrolling-consumers-amazon-prime-without-consent-sabotaging-their.
[12] Luiza Jarovsky, 'Dark patterns in AI: Privacy implications' (blog, 22 March 2023) https://www.theprivacywhisperer.com/p/dark-patterns-in-ai-privacy-implications

Harms from digital consumer manipulation that have attracted condemnation include its potential to:

- impair consumer choice and autonomy;[13]
- create or exacerbate information asymmetry;[14]
- unfairly disadvantage consumers;[15]
- violate privacy;[16]
- compromise the dignity of consumers;[17] and
- hinder or distort competition.[18]

As the use of machine learning techniques in data analytics increases, and transparency decreases, the likelihood of disadvantages for consumers and other data subjects is likely to increase. The new activities now made possible by hyper-personalised profiling, algorithmic microtargeting of marketing campaigns, and the growth of new data collectors and marketing media via connected devices and environments may lead to an opaqueness unprecedented in the consumer space: in other words, a mass inability to know our own minds**.**

B. Discrimination law

Data-driven influencing, whether using machine learning or statistical techniques, is based on the idea that if we can understand empirical connections between variables, we can predict other variables. When these variables involve human behaviour and result in decisions that affect those humans, fairness and anti-discrimination principles are critical. Currently, discrimination law protects against discrimination on the basis of protected attributes in a range of contexts but does not protect against many examples of "algorithmic bias" because the laws were written at a time when the primary concern was human animus and cognitive limits rather than bad models.

Thus, in the context of machine learning, discrimination law does not operate as effectively as it might. Organisations may well seek to avoid direct discrimination by removing variables without eliminating disparate impact. Organisations will often also be able to avoid accusations of indirect discrimination by relying on the reasonableness test – that is, there is no discrimination if the condition requirement or practice is reasonable in the circumstances. If an AI system is generally useful, relying on it for some organisational function may well be 'reasonable' even if there are biases.

[13] Mik, 'The Erosion of Autonomy in Online Consumer Transactions' (n 2); Calo, 'Digital Market Manipulation' (n 2); Helberger, 'Profiling and Targeting Consumers in the Internet of Things: A New Challenge for Consumer Law' (n 2); Marijn Sax, Natali Helberger and Nadine Bol, 'Health as a Means Towards Profitable Ends: mHealth Apps, User Autonomy, and Unfair Commercial Practices' (2018) 41 Journal of Consumer Policy 103.
[14] Donald Bergh and others, 'Information Asymmetry in Management Research: Past Accomplishments and Future Opportunities' (2019) 45 Journal of Management 122, 123.
[15] See also Nir Eyal and Ryan Hoover, Hooked: How to Build Habit-Forming Products (Portfolio/Penguin 2014).
[16] Tal Z Zarsky, 'Privacy and Manipulation in the Digital Age' (2019) 20 Theoretical Inquiries in Law 157, 175; Cass Sunstein, 'Fifty Shades of Manipulation' (2016) 1 Journal of Marketing Behaviour 213, 239.
[17] Ibid; Helberger, 'Profiling and Targeting Consumers in the Internet of Things: A New Challenge for Consumer Law' (n 2).
[18] Maurice E Stucke and Ariel Ezrachi, 'How Digital Assistants Can Harm Our Economy, Privacy, and Democracy' (2017) 32 Berkeley Technology Law Journal 1239, 1256–70; Calo, 'Digital Market Manipulation' (n 2) 1026.

It would be desirable to reform discrimination legislation so that the kinds of testing that are required when a decision affecting a human is made in part or entirely on the basis of data driven inference are laid out more clearly. Similarly, testing may also be required in the context of generative AI and search, to ensure that people do not, for example, only "see" white males in professional roles. Legal changes could reduce the incentive to avoid direct discrimination by deleting variables, which restricts the ability to test for disparate impact. If done well, such requirements would not only apply to artificial intelligence (as defined in the discussion paper or otherwise) but to any data driven process including reliance on statistical analysis.

C. Administrative Law

Some time ago, there was a proposal for the Australian Law Reform Commission to look at reform of administrative law in light of automated decision-making and artificial intelligence. The reasons for doing this have only expanded since. Currently, the legislative provisions in Australia that authorise the use of computers in administrative decision-making are extremely simple and broad in nature. They often authorise the use of computers to make decisions on behalf of the ultimate decision maker and deem the decision to be one that decision maker has made. In some cases, there are explicit provisions around certifications as to whether the system is "functioning correctly". However, such requirements are poorly worded for the case of an AI system that is likely to optimise against a particular rate of accuracy rather than function correctly in every case. The provisions need to be more nuanced to recognise the distinction between a program that does not meet specifications and a program that makes mistakes. Requirements around transparency and accountability for systems used in government decision-making are also critical alongside contestability. Procurement rules, for example, should prohibit government departments from agreeing to terms that require confidentiality as to crucial elements of system operation when systems are used in decisions affecting humans. System requirements should include the ability to generate explanations that mirror the requirements already existing in administrative law for decisions to be accompanied by adequate reasons.[19] In other words, legislation should not simply deem decisions to have been made but also specify requirements around issues like transparency, explainability, and sufficient evaluation and testing. As per the earlier point around definitions, these requirements need not be limited to AI systems but should extend to any situation where a system's output is deemed to be the decision of an authorised decision-maker.

One example of this is the *Therapeutic Goods Act 1989* (Cth), where section 7C(1) provides that, 'The Secretary may arrange for the use, under the Secretary's control, of computer programs for any purposes for which the Secretary may make decisions under this Act or the regulations' and section 7C(2) provides that, 'A decision made by the operation of a computer program under such an arrangement is taken to be a decision made by the Secretary.'

Set out below is a list of Commonwealth legislation which includes the phrase, "may arrange for use of computer programs to make decisions":

A New Tax System (Family Assistance) (Administration) Act 1999 (Cth)

Agricultural and Veterinary Chemicals Code Act 1994 (Cth)

Air Navigation (Aircraft Noise) Regulations 2018 (Cth)

Australian Education Act 2013 (Cth)

[19] See also Lyria Bennett Moses and Edward Santow, 'Accountability in the Age of Artificial Intelligence: A Right to Reasons' (2020) 94 ALJ 829.

Business Names Registration Act 2011 (Cth)

Customs Act 1901 (Cth)

Migration Act 1958 (Cth)

Military Rehabilitation and Compensation Act 2004 (Cth)

My Health Records Act 2012 (Cth)

National Consumer Credit Protection Act 2009 (Cth)

Paid Parental Leave Act 2010 (Cth)

Road Vehicle Standards Act 2018 (Cth)

Safety, Rehabilitation and Compensation (Defence-related Claims) Act 1988 (Cth)

Social Security (Administration) Act 1999 (Cth)

Superannuation (Government Co-contribution for Low Income Earners) Act 2003 (Cth)

Therapeutic Goods Act 1989 (Cth)

Veterans' Entitlements Act 1986 (Cth)

Student Assistance Act 1973 (Cth)

Much of the specific legislation included in this list is particularly problematic in the context of the Royal Commission into the Robodebt Scheme.

D. Privacy / data protection law

In some jurisdictions, privacy and data protection regulation is already having a significant impact on the extent to which organisations may use personal information in AI-related activities, including 'the secondary use, disclosure and retention of existing personal data for the purpose of constructing training sets (including disputes over anonymisation), and the use of personal data collected by AI systems for new training sets (including consent issues)'.[20]

In Australia, existing state privacy or data protection laws regulate some aspects of the handling of personal information inherent in the design and operation of many AI systems. However, it is widely accepted that these laws are outdated and inadequate to protect Australians' privacy and guard against the serious harms caused by privacy infringements. Throughout the Privacy Act Review conducted by the Attorney General's Department from 2020 to 2022, numerous submissions emphasised the need for urgent reform of these laws. The privacy risks introduced by certain AI systems and the widespread adoption of AI applications increase the urgency of proposed reforms. For example, the *Privacy Act 1988* (Cth) should be amended to:

20 Graham Greenleaf, 'The "Brussels Effect" of the EU's "AI Act" on Data Privacy Outside Europe' (2021) 171 *Privacy Laws & Business International Report* 3-7.

- Clarify and expand the definition of "personal information",[21] bringing it into line with international best practice in data protection regulation. This would, for example, aid in ensuring that organisations design systems to appropriately guard against new types of privacy attacks on ML models, such as model inversion attacks and membership inference attacks;
- Introduce a "fair and reasonable" test for dealing with personal information (in addition to consent requirements) to ensure that uses of personal information in AI systems, inter alia, are in keeping with the individuals' reasonable expectations and do not unduly harm the individuals concerned. This recognises the obstacles to genuine consent posed by organisations' control of choice architecture, and consumers' severely limited ability to understand data practices and their consequences, including in the context of personal information used in AI-related activities;
- Bring "small businesses" within the scope of the legislation, acknowledging that the size of the business does not reduce the privacy harms it may create;
- Update the definition of "consent" to mean "voluntary, informed, current, specific, and an unambiguous indication through clear action". Consent to use of personal information for additional purposes, including AI-related activities, should not be found to exist where the organisation makes supply of a product or service conditional upon the individual providing consent for such extra purposes;
- Provide individuals with a direct right of action, such that they can bring proceedings to protect their personal information without depending on the Office of the Australian Information Commissioner to make a determination in respect of the complaint, often after extended delays; and
- Require organisations – including private sector organisations – to undertake a Privacy Impact Assessment before undertaking any activity with high privacy risks. This is likely to include, for example, dealing with personal information in systems which may determine access to employment or education, or enjoyment of essential public or private services or benefits, or which potentially cause significant detriment to physical or emotional wellbeing.

While the Attorney General's Department Privacy Act Review Report proposed an obligation on organisations to include notices in respect of automated decision in their privacy policies, in our view, such proposals for mere "notice" or "transparency" are inadequate to address the issues associated with automated decision making. Our reasoning on these issues is discussed further in section 2(C) above and section 9 below.

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.

Government should support participation by Standards Australia in international standards development. In particular, financial support would help ensure that not for profit sectors (including

21 See further Katharine Kemp, 'Ending the Fictions in Modern Data Practices: Submission in Response to the Privacy Act Review Report' (Submission, 31 March 2023) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4521070 2-3, on the appropriate definition of 'personal information'.

consumer groups and privacy advocates) have the ability to participate meaningfully in national and international meetings.

Government should also develop procurement requirements that ensure core administrative values (fairness, accountability, etc) are factored into decisions as to which system to procure and the terms under which that occurs, including in respect to the ability to disclose important information about how systems operate. AI procurement guidelines are likely to have a significantly greater impact than AI ethical principles which lack meaningful consequences for non-compliance.

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.

Significant work has been done on this topic by ANU's Tech Policy Design Centre.[22] We agree with the importance of greater co-ordination across departments and units. This would enable, for example, a clear plan for staging law reform rather than, as occurs now, many consultations cutting across each other at the same time. The scope for such a co-ordination function would need to be flexible - whether "digital", "AI" or some other phrase is the most appropriate will likely change over time as technology evolves.

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?

Australia could perhaps look to some of the thinking in Switzerland, which encourages a distinct approach to that operating in the EU. In particular, the position paper argues:[23]

> The challenges posed by algorithmic systems are manifold and often have a new dimension or quality, but they are not unique to such systems. Therefore, these challenges should not be covered by a general "AI law" or an "algorithm law". Instead, a combination of general and sector-specific standards is appropriate. The focus here is on the selective adaptation of existing laws.

This is similar to our proposed approach, as explained in response to Question 1.

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

In many circumstances, different approaches will apply to the public and private sector, just as they do in other domains. For example, public sector decision-makers have to give reasons for (most) decisions, whereas, in the private sector, 'reasons for decision' (say, on pricing) are rarely required. Private firms are accountable to the market in a very different sense to the way in which government is accountable to citizens.

Context is relevant more generally in determining the best approach. For example, if a private organisation set up an 'AI dating site' with minimal transparency ("meet your mystery dream match"), that product will succeed or fail in the market but need not involve heavy handed government regulation around mandated explanations as to the 'reasons' for particular matches. On

[22] Johanna Weaver et al, *Tending the Tech Ecosystem* (May 2022), https://techpolicydesign.au/wp-content/uploads/2022/11/Web_TPDC_Publication_NO.1_2022-3.pdf.
[23] A Swiss Position Paper can be found at https://algorithmwatch.ch/en/position-paper-legal-framework-for-ai/.

the contrary, legislation authorising Ministers or other decision-makers to rely on systems to make decisions on their behalf should require such systems to have a degree of transparency. This is in line with broader policies around the benefits of 'sunlight' in government.

One way in which the Commonwealth could demonstrate "best practice" in the use of AI technologies would be to adopt a position of a "model user". The Commonwealth already acts as a "model litigant". The source of the obligation to act as a model litigant for Commonwealth entities is in section 55ZF of the *Judiciary Act 1903* (Cth). This section gives the Commonwealth Attorney-General the power to issue legal services directions that apply to any work that is performed on behalf of the Commonwealth. The legal directions themselves are found in Appendix B of the Legal Services Directions 2017. However, a model user approach could be made under the *Public Governance, Performance and Accountability Act 2013* (Cth) with amendment.

To ensure that the model user obligation is met, complaints should be able to be made to the Commonwealth Ombudsman after Recommendations 21.1 to 21.5 of the Royal Commission into the Robodebt Scheme have been implemented.

Similar to the dispute resolution approach in the public sector use of AI technologies, there needs to be an ombuds in the private sector with expertise to manage concerns around artificial intelligence. This may require additional resourcing.

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

There are a variety of things the Government can do in line with this objective:

- Education and training alongside clear expectations
- Developing internal policies, along the lines of the NSW AI assurance framework
- Ensure that legislation that authorises reliance on systems for decision-making include provisions setting requirements for such systems (in line with what are currently unenforceable ethical principles)
- Implement recommendations flowing from the Robodebt Royal Commission

8. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

As per our response to Question 1, we believe that problem-specific solutions are preferable to technology-specific ones. There are some circumstances where the two align - in other words, the problem relates directly to the use of particular technology. Examples include the use of biometrics for mass identification and the use of automated weapons that make 'kill' decisions.

In other cases, problem-specific, principles-based legislation can be supplemented by subordinate legislation or guidance that explains how general principles apply to particular technological contexts. Where appropriate international standards are available, guidelines can point to standards compliance with which would constitute compliance with particular legal requirements.

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.

Transparency is a concept that many people are agitating for, but the crucial questions are *what* is rendered transparent, *to whom*, *how* and in which contexts. A driver of an automated vehicle does not need a continuous output from an automated vehicle explaining the logic behind a particular automated decision to steer slightly left to stay in a lane. Rather, they want to know that the car has been evaluated (overall) as safe. On the other hand, the public should be able to find out the logic behind government systems that make decisions affecting them, the nature and quality of training data used, the testing and evaluation of systems that has been conducted (and the results of such), the assumptions on which a system relies, and so forth. Mandating uniform transparency requirements across sectors and contexts would not be helpful in almost all cases. An exception is the proposal (across sectors and contexts) to prohibit misleading uses of AI and automated systems. People should have a right to know when they are interacting with a machine rather than a human (unless they voluntarily relinquish that right for a specific activity, for example in the context of AI research). Similarly, there should be transparency about the involvement of AI in content-generation, so that (for example), an AI-generated image is labelled as such rather than represented as a human artwork.

One way in which governments can provide signalling as to best practice, would be to include the model cards[24] (where applicable) used by the Commonwealth in its use of generative AI. Similarly, a requirement in public sector procurement that model cards are a mandatory part of supply of generative AI products and services would assist with transparency.

A model card is a human-readable document that provides critical information about a machine learning model. It is used to help people understand how the model works, its limitations, and its potential biases.

Model cards usually include the following minimum information:

(a) **Model name and version**: This information helps to identify the model and to track its development over time;
(b) **Model type**: This information describes the type of machine learning model, such as a neural network, large language model, decision tree, or support vector machine;
(c) **Model inputs and outputs**: This information describes the types of data that the model can take as input and the types of data that it produces as output;
(d) **Model training data**: This information describes the data that was used to train the model. This information can be used to assess the model's performance on different types of data;
(e) **Model evaluation metrics**: This information describes how the model was evaluated. This information can be used to assess the model's performance on different tasks; and
(f) **Known limitations and biases**: This information describes any known limitations or biases in the model. This information can be used to help users interpret the model's results and to make informed decisions about its use.

24 Margaret Mitchell et al, 'Model Cards for Model Reporting' (2019) Proceedings on the Conference for Fairness, Accountability, and Transparency https://dl.acm.org/doi/10.1145/3287560.3287596.

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?

There are a variety of contexts in which high-risk AI should be prohibited (or subject to strong restrictions). Ultimately, where use of AI as the primary decision-maker, in a process that would otherwise require rigorous and nuanced human input, may result in significant harm or a burden on human rights, AI should be banned. Examples of these contexts may include where lethal force is used in police or military operations, and in formal dispute resolution which should continue to rely on human judges and juries.

11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

Trust should not be sought as an end in and of itself. It is crucial that the public remain appropriately sceptical about computer systems with which they interact so that they take appropriate measures to protect their privacy and challenge illegal decisions. We want the public to be aware not only of the benefits of AI, but also of its limitations. What the government should focus on is what is commonly referred to as *trustworthiness* - making the systems better so that the public can have confidence in their deployment. The "model user" approach set out above would also assist trustworthiness.

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

In respect of LLMs, it will be critical to apply a risk-based approach to all elements of the use of the LLM. In terms of the model, this means both disclosing the model card and analysing the risks that flow from the bias set out in that card. The effect of this risk analysis may be that a specific model should not be used. That is, part of the risk-based approach is to determine the choice of model. This approach can then be applied to any changes in the model. For example, if a model is to be fine-tuned for a specific purpose, then a further risk analysis needs to be completed on the data being used for the fine-tuning. Essentially, this risk-based approach is to analyse the tuning data in the absence of a model card. The weights and biases used in the fine-tuning will also need to be analysed in the context of the use of the fine-tuned model. One way of dealing with this is to require that any fine-tuned model produced by the public sector has an accurate model card and that each of the fine-tuning dataset, weights, and biases are published with that model card.

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to: a. public or private organisations or both? b. developers or deployers or both?

Recent experience in the provision of self-regulatory codes in the mis/dis-information area suggests that industry developed codes will be problematic. Specifically, the problems are likely to be driven by a combination of lengthy delay in code development and information asymmetries. The often-employed approach of self-regulatory codes which become mandatory if they are breached is that there needs to be a regulator in place to monitor such a breach. On the other hand, a co-designed set of codes which are mandatory would work to lessen information asymmetries during the regulatory co-design process. Ensuring a variety of stakeholders are part of the co-design will also mean that the codes will apply to all of the groups identified in the question. It is also important that the potential regulators are involved. International standards may also be appropriate for local

adoption, particularly when involvement of a diverse range of organisations and interests are represented.

The authors of this submission are able to assist in the regulatory co-design process as facilitators; one is also an active participant in standards development for AI.

Yours sincerely,

Lyria Bennett Moses, Katharine Kemp, Annabelle Lee, Kayleen Manwaring, Rob Nicholls