Safe and responsible AI in Australia - Discussion Paper Response

School of Computing and Information Systems and Centre for Artificial Intelligence and Digital Ethics, The University of Melbourne

26 July 2023

Professor Jeannie Marie Paterson, Dr Shaanan Cohney, and Professor Liz Sonenberg, with input from Dr Fahimeh Abedi, Dr Marc Cheong, Carmelina Contarino, Joe Brailsford, Professor Tom Drummond, Liam Harding, Professor Eduard Hovy, Dr Christine de Kock, Brian Martin, Aidan McLoughney, Associate Professor Olga Ohrimenko, Dr Sarita Rosenstock, Professor Ben Rubinstein, and Michael Wildenauer.

Overview

The University of Melbourne is home to world-class expertise in AI, ethics, law, and policy, represented in this submission by contributors from the School of Computing and Information Systems as well as the Centre for Artificial Intelligence and Digital Ethics (CAIDE).

The University of Melbourne welcomes the Government's interest in regulating for safe and responsible AI.

AI offers considerable opportunities across Australian society. But it is important this development is accompanied by technical, ethical, and legal best practice to ensure that AI applications are effective, fair, and safe. Overstated concerns about the risks of AI obscure its real and present risks; while undue emphasis on the importance of innovation overlooks the harms that AI enabled technologies may perpetuate.

We consider that there is no one approach that will ensure safe and responsible AI. Rather, what is required is a mix, or 'network',[1] of regulatory interventions that together promote these desired objectives. To this end, we consider that promoting safe and responsible AI requires good regulatory design, and the practical supports required to make regulation effective.

Good regulatory design means targeted responses to identified risks, backed up by 'safety-net' or 'principles-based' protections for circumstances where those more specific interventions prove inapplicable or ineffective. Overly complex or poorly designed regulation is unlikely to achieve its key goals of promoting effective, fair and safe AI, and will merely increase the costs of compliance for industry, with little real gain for individuals and society.

Practical support for the regulatory framework means properly funding regulators to ensure the measures decided upon are effectively implemented and enforced. It also requires training and education for government, regulators and society generally to 'demystify' AI. The aim should be to enable informed, inclusive, and participatory responses to the challenges and opportunities this technology presents.

Questions

Definitions

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

We do not agree entirely with the definitions in the discussion paper.

[1] On networked regulation see Christine Parker et al, 'Can the Hidden Hand of the Market Be an Effective and Legitimate Regulator? The Case of Animal Welfare Under a Labeling for Consumer Choice Policy Approach', (2017) 11 Regulation & Governance 368.

- The definition for AI reads:

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

We recommend that the underlined four words be dropped.

The underlined words 'without explicit programming' are inaccurate; there are numerous AI systems - for example robots, self-driving cars, expert system medical diagnosis engines, and others _ that have been built with explicit programming. The cited definition of AI, Def 3.1.4 of ISO 22989:2022, does not include the underlined words - it simply reads "engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives".

The ISO also does not use predictive. We don't think this term is appropriate, as there are many outputs of AI systems that are not "predictive". Systems that provide recommendations, classify images, or generate novel works all fall outside the narrow purview of predictive AI.

Generative AI and classifiers are both caught by the ISO definition which also covers automated systems which are used to guide influence or replace human decision making.

- The definition of LLM reads:

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

This is not the current meaning of LLM. An LLM is one component of a generative AI system, namely the (passive) language store. The second component is an active generative cycle that extracts fragments from the LLM and composes them into the text that is produced for the user.

- The definition of machine learning reads:

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

Machine learning does not refer to the output patterns as in the current definition. Machine learning refers to training a computer model to improve on its given task by optimising its parameters in an automatic way. Machine learning is a subfield of AI (see above our response to the first definition).

- The definition of a multimodal foundation model reads:

  'Is a type of generative AI that can process and output multiple data types (e.g. text, images, audio)'.

A MfM is one component of a generative Al system; the other being the active generation cycle that can process and output multiple data types (e.g. text, images, audio).

We also recommend adding a definition of Generative Models, of which LLMs are just one. However, there are other types of content and data that AI/ML can generate (e.g., images by diffusion models or speech, video, and audio data).

Potential Gaps in Approaches

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

We do not think all potential risks from AI are covered by Australia's existing regulatory approach.

i. First, we refer to our submissions to the Review of the Privacy Act 1988,[2] and reaffirm our support for stronger data protection laws in Australia.
ii. Secondly, we are concerned about the growing use of facial and biometric identification particularly in public spaces and by private entities.[3] We support stronger measures to curb this trend.
iii. Thirdly, as noted in the introduction above, we consider that regulatory intervention should be carefully designed to respond to the risks raised by AI, and to be compatible and coherent with Australia's existing regulatory regime.[4]
iv. Fourthly, we think there is greater scope for ex-ante protections, which should also facilitate establishing wrong doing for harms.[5]

These issues are discussed in more detail below.

2.1 The importance of privacy reform

Some risks from AI relate to privacy of individuals represented in training sets, or from the application of AI to imputing personal attributes. Such risks of intentional or unintentional privacy loss may be amplified by AI, placing an increased urgency on reforms to the Privacy Act 1988.

2.2 Biometric identification, regulation and bans

These issues are discussed below in response to question 2.4.

2.3 Regulatory design for effective responses to AI risks

Merely enacting more law or regulation will not necessarily ensure that the risks of AI are adequately addressed, or that beneficial innovation is supported. Overly complex legislation becomes costly to enforce, and for compliance, as well as potentially encouraging regulatory arbitrage.[6] Thus, in responding to the risks of AI, care must be taken in identifying the risks warranting a response, designing clear, accessible and targeted interventions having regard to existing law, and remembering the need for principles based or safety net laws that are capable of adapting to new manifestations of emerging technologies.

2.4 Strengthening the (ex-ante) regulatory response to the risks of AI harms

In principle, we think issues of liability and compensation for most potential risks from AI are likely to be

[2] See Paterson, Jeannie, Cohney, Shaanan, Kulik, Lars, and Harding, Liam 2022. "Response to the Review of the Privacy Act.". Jeannie Paterson, Shaanan Cohney, Lars Kulik and Liam Harding, Submission to Privacy Act Review – Discussion Paper (24 January 2022).

[3] Jarni Blakkarly, 'Kmart, Bunnings and The Good Guys using facial recognition technology in stores' Choice (Web article, 12 July 2022) https://www.choice.com.au/consumers-and-data/data-collection-and-use/how-your-data-is-used/articles/kmart-bunnings-and-the-good-guys-using-facial-recognition-technology-in-store.

[4] On these design issues, see also, Jeannie Marie Paterson and Elise Bant. 'Should Australia Introduce a Prohibition on Unfair Trading? Responding to Exploitative Business Systems in Person and Online' (2020) 44 Journal of Consumer Policy 1.

[5]. Jeannie Paterson, 'Misleading AI: Regulatory Strategies for Algorithmic Transparency in Technologies Augmenting Consumer Decision-Making' (2023) 34 Loyola Consumer Law Review 558.

[6] Cf Australian Law Reform Commission, Review of the Legislative Framework for Corporations and Financial Services Regulation <https://www.alrc.gov.au/inquiry/review-of-the-legislative-framework-for-corporations-and-financial-services-regulation/>.

covered by Australia's existing regulatory/legal regimes.[7] However, the aim of effective regulation should be to reduce the likelihood harm occurring. Accordingly, we think that consideration should be given to mandating a greater role for technical standards, risk assessment frameworks and strong mechanisms for accountability and governance to reduce the likelihood of harm from AI to individuals, society, and the planet.[8] Moreover, there is currently considerable uncertainty about how existing legal and regulatory regimes apply to AI, and, potentially, practical hurdles in establishing wrongdoing arising from the opacity of AI systems. Our preferred kinds of ex ante intervention may further reduce these concerns by aiding in the process of establishing wrongdoing and enforcing prohibition on the offending conduct.[9]

AI risks

Many AI risks are now well documented. It is nevertheless vital to consider these to ensure effective regulation for reducing risk and promoting beneficent outcomes. The risks of AI include:

- Technical risks:
    - Inaccuracy: produced by drift, lack of robustness, inaccuracy, bias;
    - Information leakage: Some AI and ML models make decisions based on the past data used to develop them thereby potentially disclosing information about this data to third parties using these models;
    - Adversarial manipulations: Input in AI and ML models comes from human prompts, other systems, or surroundings (e.g., videos, images, sounds). These inputs can be intercepted and manipulated, causing AI and ML models to behave in an adversarial manner (e.g., avoiding recognition of a STOP sign or generating a text with negative emotions). These risks have to be adequately addressed or at least highlighted;
- Human rights risks: including a lack of equity and access, bias and discrimination, the erosion of privacy, the undermining of rule of law values;
- Societal risks: misinformation and deepfakes disrupting democratic processes, civil society, and markets;
- Existential risks: arising from concerns about what it means to be human and how do we understand human machine interactions.

Other human systems and behaviours also raise many of these risks of harm. Thus, it might be asked why the use of AI should be subject to special attention. We suggest that the character and magnitude of the risks from AI can be difficult to predict. This is so because AI systems may:

- develop and come onto the market abruptly;
- be built and deployed by a range of actors and states;
- be put to unexpected and novel uses; and

[7] But in the context of consumer protection law see the case for a prohibition on unfair commercial practices: Jeannie Paterson, Elise Bant, Nicholas Felstead and Eugene Twomey, 'Beyond the unwritten law: The limits of statutory unconscionable conduct' (2023) 17 Journal of Equity 1.

[8] Jeannie Marie Paterson, Shanton Change, Marc Cheong, Chris Culnane, Suellette Dreyfus and Dana McKay, 'The Hidden Harms of Targeted Advertising by Algorithm and Interventions from the Consumer Protection Toolkit' 9 International Journal on Consumer Law and Practice 1; Jeannie Marie Paterson, 'Making robo-advisers careful? Duties of care in providing automated financial advice to consumers' 18 Law and Financial Markets Review.

[9] See Jeannie Paterson, 'Misleading AI: Regulatory Strategies for Algorithmic Transparency in Technologies Augmenting Consumer Decision-Making' (2023) 34 Loyola Consumer Law Review 558.

- be opaque, meaning harms such as discrimination may take time to identify and may be embedded in proxies or correlations rather than direct reliance on protected attributes.[10]

General law and statute

The use and outputs of AI systems will be subject to the law applying in the context in which the AI system is being used. For example, an AI service provided without reasonable care will be subject to the tort of negligence. AI bias may be contrary to anti-discrimination legislation or unconscionable under consumer law. Inaccurate or unlawful decisions produced using AI may breach administrative or corporations law, depending on the context.

However, and as already, noted it is desirable to undertake measures to reduce the likelihood of harms occurring or, in other words, embed accountability. It may also be desirable to introduce regulatory requirements that will assist regulators or individuals and businesses subject to harm by AI to understand where and how AI has been used, and to establish the cause of the harm, such as through requirements of transparency, explanations and audits.

These kinds of demands are key requirements of ethical or responsible AI.

Principles of AI ethics

AI principles are described as soft law because they are not formal law but nonetheless influence those subject to the law and the interpretation of the law that applies. Principles of AI ethics are the starting point for understanding the responsibilities of those developing, deploying, and using AI tools.

Key governance requirements in many formulations of AI principles are transparency, explainability contestability and accountability.

Transparency

Transparency is a requirement to provide information about where and how AI is being used to inform decision making, and the weightings or process influencing a decision or recommendation.[11]

In some higher-risk instances it may also be necessary to provide access for public auditing of model weights, training data, model outputs, and model code.

Explanations

Statements that provide clarity around how decisions or recommendations are reached.[12]

Contestability

Clear and easily navigated processes for contesting the outputs of AI that affect individual rights or entitlements.[13]

---

[10] McLoughney, Aidan James, Paterson, Jeannie Marie, Cheong, Marc, and Wirth, Anthony 2023. "'Emerging proxies' in information-rich machine learning: a threat to fairness?." 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS) doi:10.1109/ethics57328.2023.10155045. Aidan James McLoughney, Jeannie Marie Paterson, Marc Cheong and Anthony Wirth, '"Emerging proxies" in information-rich machine learning: a threat to fairness?' (Conference Paper, IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS), 18 May 2023),

[11] See Jeannie Paterson, 'Misleading AI: Regulatory Strategies for Algorithmic Transparency in Technologies Augmenting Consumer Decision-Making' (2023) 34 Loyola Consumer Law Review 558..

[12] On state of the art studies into explanations see Tim Miller, 'Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI' (Conference Paper, FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency, June 2023); Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller and Liz Sonenberg, 'The Effects of Explanations on Automation Bias' (2023) 322(9) Artificial Intelligence.

[13] Henrietta Lyons, Senuri Wijenayake, Tim Miller and Eduardo Velloso, 'What's the appeal? Perceptions of Review Processes for Algorithmic Decisions' (Conference Paper, 2022 CHI Conference on Human Factors in Computing Systems, April 2022); Henrietta

Accountability

Accountability requires systems and processes for overseeing and reducing the risks of any AI tool, including for example:

- Testing, reviewing, cleaning the data;
- Auditing outcomes to look for patterns of disadvantage against protected attributes and groups;
- Inclusive design; and
- Ongoing governance.

Notably, accountability means much more than having a 'human in the loop', which may do little to address concerns about AI due to the impact of factors such as opacity and human automation bias.[14]

Many of these principles are now laid out in technical standards, such as those developed by IEEE and Standards Australia.

We discuss these requirements further below in response to question 14.

Making standards mandatory

We note the principles of AI ethics are not solutions in and of themselves, but consider they should be part of the regulatory mix for ensuring responsible AI. However, to the extent these ex-ante principles are significant in reducing the risks of harm of AI products, we suggest that thought should be given ways of making firms and other entities in the AI lifecycle engage with them fully and impactfully.

We envisage this might be done by requiring risk assessments for AI outputs and use cases, a strategy for responding to that risk through governance and accountability frameworks consistent with best practice principles, and, possibly, providing that compliance with technical standards frameworks is evidence of best practice (although not conclusive).

Importantly, we consider that any risk-based approach should avoid prescriptive rules (see below at 14). Prescriptive rules for how firms respond to the demands of responsible AI are likely to prove unfit for purpose. They are likely to lag behind technological developments and prove a poor fit for the contribution of the different entities that might be part of the AI lifecycle.

Thus, and subject to our comments on bans below under question 10, we prefer a principles-based approach to regulation. We note that a principles-based approach is not the same as self-regulation (which we do not support) or even a soft law approach (since although it may make use of soft law standards it contains legally binding principles).

Under a principles based approach to regulation, legally binding principles would express expected standards of conduct and performance (e.g. risk assessments, governance/accountability, standards of safety). The detailed guidance as to what that conduct might require would be found in low level or soft guidance able easily to be changed in line with the market or developments in technology.

Principles based regulation allows greater responsiveness to future change. Any AI regulatory regime must be designed to be agile. The future direction of technological innovation is uncertain–whether the bulk of AI use will occur on device, or in the cloud and further, whether it will be large entities providing most models or networks of individuals. Like the early days of the internet, the direction of development impacts

Lyons, Eduardo Velloso and Tim Miller, 'Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions' (2021) 5(CSCW1) Proceedings of the ACM on Human-Computer Interaction 1.

[14] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller and Liz Sonenberg, 'The Effects of Explanations on Automation Bias' (2023) 322(9) Artificial Intelligence.

the effectiveness of various forms of regulation and a nimble approach is therefore required.

The relevance of the AI life cycle and the AI supply chain

Standards for ethical or responsible AI should be scaled to apply across the AI lifecycle.

The AI life cycle covers design, development, deployment, and operation of AI. Typically, this is an iterative process and may involve elements from different kinds of suppliers in an extended supply chain. Different suppliers may be involved in each of the following AI supply chain processes:

- Design, implementation, and dispersal of model structure;
- Collation of training and testing data sets;
- Actual training on data sets;
- Provision of pre-trained weights;
- "Fine Tuning" where weights are tweaked to fit a new context;
- Deployment of a model for inference;
- Provision of the final service to end-users.

There are at least three considerations for good regulatory design that arise from an understanding of the AI supply chain:

1. First, the liability of the suppliers in the supply chain, behind the point of supply, is often highly uncertain. For example, where an AI causes harm, the final provider of that system may be liable under general law or statute. But the responsibility of suppliers, such as the person who provided the data set, behind that provider may be unclear. The parties in the supply chain can of course deal with these risks through contract. However, uncertainty and opacity may lead to inefficient and ineffective allocations of risk as between suppliers, while also failing to adequately reduce the risks to the individuals who are subject to the outputs of the system.
2. Secondly, features of the AI lifecycle mean that any mandatory standards should be scaled to the operations and application of the AI tool. This is because the nature and scope of the risks attaching to any one part of the lifecycle will vary. Parties will need a clear account of the nature and scope of risks specific to their stage of the supply chain and the character of their contribution to an AI product in order to know what to look for. This specificity and precision in AI governance comes from adopting a risk-based approach, which we also support.
3. Thirdly, while large models provided by well-resourced actors have been the focus of much discussion, we warn that powerful models are available in open-source form. These models are publicly developed and contributed to by decentralised sets of individuals whose identities may be difficult to discern. Code and model weights are posted on public repositories under pseudonyms for other users to download and run on their own hardware. Recent results show that large language models can be fully hosted on an iPhone and have power comparable (those lesser than) those running on large server farms. This poses a challenge to regulating the technology, rather than specific offerings of the technology to consumers.

Gatekeepers

To some extent the incentive for firms to comply with ex ante risk assessment obligations comes from general law and legislation. Failure adequately to assess and respond to the risks of AI may lead to liability under these regimes, e.g. as misleading conduct, negligence, or a failure to engage in conduct that is efficient, honest, and fair under financial services regulation. We further suggest reporting obligations might attach to corporate gatekeepers in the AI lifecycle, namely firms that are supplying AI products or services (as opposed to elements of those products) to other businesses or consumers.

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.

We consider that there are at least two non-regulatory initiatives that the Australian Government might implement to support responsible AI practices in Australia: funding regulatory oversight and best practices standards for its own use of AI.

Funding in house AI expertise within regulators

The Australian Competition and Consumer Commission (ACCC) is an effective and active regulator. We draw attention to the significant success already shown by the ACCC[15] in enforcing the Australian Consumer Law in applying to technology driven services.[16]

However, effective regulation increasingly requires technical expertise. Regulators need strong powers to gather information in investigating and enforcing regulatory compliance. Additionally, these powers need to be accompanied by applied research and data-analysis capabilities–capabilities that regulators worldwide are still in the process of developing. We recommend that relevant regulators should be funded to develop and maintain this necessary technical expertise to enforce both general law and AI specific obligations. Such in house expertise should be complemented by a Standing Expert Advisory Group, as we in response to question 4 below.

One small-scale model is the Office of Technology Research and Investigation (OTECH) at the U.S. Federal Trade Commission (FTC), created to "level[] the playing field and empower[] the FTC to better tackle abuses from technology companies." The success of OTECH, despite constrained resources, has motivated legislation to fund fully staffed Bureau of Technology and the hiring of a Chief Technologist.[17] Australia might learn from this experience, by providing secure funding to develop AI expertise to key regulators.

Government modelling of best practice ethical AI governance and risk assessment

The government has a valuable opportunity to send a message to organisations that use AI through its own approach to responsible AI practices. Appropriate policy settings for central departments, portfolio agencies and SOEs in this area signal a seriousness of intent, and work to socialise and normalise such approaches in the wider corporate and non-governmental spheres, as well influence policy settings for other levels of government in Australia.

Suggested settings include visible compliance with responsibility-promoting policies in the way that public sector organisations treat AI, but also mandating practices at appropriate levels for external organisations, matched to the risk profile of the particular technology solution in question, during the tendering process and when contracting with suppliers.

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**

We support the government establishing an advisory body to oversee its own use of AI and ADM.[18] We

[15] ACCC, 'Trivago Mislead Consumers About Hotel Room Rates' (Media Release) https://www.accc.gov.au/media-release/trivago-misled-consumers-about-hotel-room-rates; ACCC, 'Trivago to Pay 447 Million in Penalties for Misleading Consumers Over Hotel Room Rates' (Media release) https://www.accc.gov.au/media-release/trivago-to-pay-447-million-in-penalties-for-misleading-consumers-over-hotel-room-rates.

[16] See eg Liam Harding, Jeannie Marie Paterson, Elise Bant, 'ACCC vs Big Tech: Round 10 and counting' Pursuit (24 March 2022).

[17] FTC https://www.ftc.gov/technologists-; : https://www.wyden.senate.gov/imo/media/doc/Wyden%20Privacy%20Bill%20Discussion%20Draft%20Nov%201.pdf (see discussion to draft bill);Federal Trade Commission, 'Office of Technology Hiring' (Webpage) https://www.ftc.gov/technologists; See also the discussion to a draft bill to fund a fully staffed Bureau of Technology, and hiring of a Chief Technologist: https://www.wyden.senate.gov/imo/media/doc/Wyden%20Privacy%20Bill%20Discussion%20Draft%20Nov%201.pdf.

[18] Recommendation 17.2 of the Report of The Royal Commission into the Robodebt Scheme recommends the 'Establishment of a body to monitor and audit automated decision-making': Royal Commission into the Robodebt Scheme (Report, 7 July 2023) vol 1, xvi.

note that many of the existing legal regimes that will apply to regulate the use of AI by private sector bodies in making decisions about and providing services to individuals do not apply to government (e.g. Corporations Law; Australian Consumer Law). An AI Advisory Group would additionally provide expert insight, advice and recommendations to government, parliament, and regulators.

There is currently no representative organisation that coordinates AI expertise across the country. The National AI Centre at CSIRO might be a convenor but currently has uneven representation (being focused on NSW). We suggest that this AI Advisory Group should have membership from a diverse range of stakeholders – industry, tech, policy, researchers, and people with lived experience of the outputs of AI and ADM. Members should be drawn from diverse backgrounds and be representative of the whole of Australia.

Responses Suitable for Australia

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

We suggest that Australia should take guidance from initiatives in other countries, including near neighbours such as Singapore and trading partners such as the UK, US, and Canada. These initiatives are all helpfully detailed in the consultation document.

We strongly recommend that Australia's actions in this field are designed with regard to best international practice, and to complement that practice. While regulation for responsible AI is a driving purpose of reform, there is no benefit in reform that increases compliance costs of innovation without proportionate improvement in outcomes. Australian firms dealing overseas and international firms operating in Australia may need to comply with multiple regulatory regimes. Ideally Australia's requirements complement and are compatible with those in key overseas markets (without lowering national expectations or standards).

Target Areas

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

Private business and government respond to different pressures and so different types of incentives and disincentives to identify and respond to potential risks of AI may be appropriate. We also consider that government should hold itself to the highest standards of ethical AI practice. As discussed above in response to question 4, governments are in a good position to set exemplary standards, trial and demonstrate best practices.

More generally, however, we consider that AI should be regulated by reference to its outputs and the character of the service provided e.g. utilities, health services, financial management, and education. Private bodies provide essential and necessary services that have profound effects on people's lives. The use of AI in these contexts should be done to high standards of governance, care, and regard to human rights regardless of the identity of the provider. Any design or deployment of AI, but particularly where it touches on fundamental rights, should proceed with regard to user experience, and human-centred design principles.

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

We think there is an imperative for government to develop its own procurement practices that are compliant with a rigorous risk assessment process and consistent with principles of AI ethics.

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 noted above at 2.3, we think effective regulation of AI is best understood as multi-faceted, or as a regulatory network, spanning across the AI lifecycle and supply chain. This means that an effective regime may include soft law guidelines, AI standards and generally applicable or sector specific law. AI specific regulation might best be used to embed ex-ante risk assessment processes and proportionate responses to this risk assessment in terms of transparency, explanations, and accountability mechanisms. AI specific law may also be required for uses of technology judged high risk and warranting additional regulation. Principles based law, found in regimes such as the Australian Consumer Law, are appropriate as a safety net to respond to those harms that arise despite a risk assessment process, and ex-ante interventions. General law and legislation also provide an incentive to take seriously the need to include safety, privacy, accountability etc 'by design'.[19] Sector specific law, such as financial services have a role in governing AI, like other interventions, according to the norms and standards imposed on all participants in that market.

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?*

We think transparency, along with explainability, is important but not the entire solution. See also our discussion above at 2.4.

Transparency

Transparency can be important in promoting safe and responsible AI. But care needs to be taken not to overload the concept of transparency in a way that undermines its utility.

The concept of transparency in AI regulation should be used precisely and not muddle the varying uses for transparency. Transparency may be used to provide visibility for:

- governance (auditing/compliance),
- internal operations (ensure companies are encouraged and able to notice the right sorts of things),
- regulatory oversight (allowing regulators to verify claims made by firms about their AI tools and processes, as well as enforce relevant law);
- advocates and lawyers in pursuing compensation and redress for AI harms; and
- journalists and public interest organisations in ensuing those who deploy AI face meaningful public accountability;
- consumers (to respect autonomy and empower decision-making or choice).

The demands of 'transparency in AI will differ between these various uses.

Importantly, transparency should not merely mean providing users/subjects of AI with "more info" (especially with huge datasets and complex code). Nor should transparency be used as a reason by firms for themselves failing to take adequate steps for responsible AI. Merely telling individuals about the AI or how it 'works' should not make users/subjects solely responsible for protecting themselves against possible risks arising from that use of AI.

A further generic solution is to require AI system builders to provide baseline information about their use of AI. Typically, this means providing the details of:

- the source data used to train machine learning algorithms: who collected the data, what measures of bias were checked and performance against these metrics, how representative training data is of the actual source materials how current is the training data, and other similar necessities;

[19] Jeannie Marie Paterson, 'Making Robo-advisers careful? Duties of care in providing automated financial advice to consumers' (2023) 18 Law and Financial Markets Review.

- the training regime used: which data preparation steps were performed, which machine learning algorithm(s) were applied, what parameters were used to guide them, what the parameter values were, what the learning outputs were;
- the achieved performance results of training: what the metrics of evaluation were, what the systems' output scores were, what gold standard material was used for comparison, who prepared the gold standard material, and how was it prepared
- the use of the trained system in the current application: how the system was adapted to the application, the parameters of run-time application, what was the impact of adaptation on performance, how well the system performs in the current application, what the gold standard is for measurement of current performance, how this gold standard material was prepared or obtained;
- any further observations and/or measurements on the bias, gaps, and other performance statistics of the system applied to the current application;

Explanations

The need for AI systems (particularly LLM-based generative AI systems) to explain their reasoning has been recognised for well over a decade.[20] But there are still no general and standard practices or methodologies that specify what an adequate explanation is exactly and how to produce one, although the issue is the topic of considerable research.[21] Existing explanation strategies (including hotspot analysis, meta-networks that interpret others, explanatory diagnosis using test cases, etc.) exist and should be encouraged where appropriate. Like transparency, explanations are not a panacea for responsible AI, and have limitations such as in identifying unlawful or unfair bias.[22]

Transparency and Privacy

The requirements of data protection and privacy can conflict with the above suggestions for transparency and explanations, since these features will often require access to personal data, and indeed may risk disclosing such information.

Machine learning models are trained on data. This data, depending on the setting, comes either from scraping the Internet, company's users (e.g., emails of customers used to train auto-complete features during email composition) or other proprietary sources. Privacy research has demonstrated the reverse-engineering of this training data from access to machine learning model APIs (e.g., by giving prompts to a language model or image generation service). To this end, one has to carefully consider and understand where training data came from and if it has privacy and/or intellectual property rights attached to it, since these rights can be blatantly violated if used in ML model.

The California Consumer Privacy Act of 2018[23] encodes a right to erasure similar to the GDPR, requiring that organisations remove data without unreasonable delay when requested. Recently the U.S. Federal Trade Commission found that this extends to derived products of data, including AI models trained on data.[24] The area of "machine unlearning" has emerged as a technical mitigation to assist organisations

[20] Wikipedia 'Explainable Artificial Intelligence' (Webpage) <https://en.wikipedia.org/wiki/Explainable_artificial_intelligence>.

[21] See eg Tim Miller, 'Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI' (Conference Paper, FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency, June 2023); Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller and Liz Sonenberg, 'The Effects of Explanations on Automation Bias' (2023) 322(9) Artificial Intelligence.

[22] Jeannie Paterson, 'Misleading AI: Regulatory Strategies for Algorithmic Transparency in Technologies Augmenting Consumer Decision-Making' (2023) 34 Loyola Consumer Law Review 558.

[23] California Consumer Privacy Act 2018 1.81.5 Cal Civ Code. 5

[24] 'California Company Settles FTC Allegations It Deceived Consumers about use of Facial Recognition in Photo Storage App' FTC (Media Release) < https://www.ftc.gov/news-events/news/press-releases/2021/01/california-company-settles-ftc-allegations-it-

with efficient removal of data from downstream models. [25] We note however that machine unlearning is still in its infancy and may raise privacy concerns of its own while remaining a valuable area for future exploration.

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

As discussed above at 2.4, we think that the use of mandatory standards should be proportionate to risk with greater expectations placed on gatekeeper firms that supply AI products or utilise AI decision-making that affects individuals or small businesses.

Model cards for AI provided as a service to consumers and (small) business

Additionally, transparency requirements may be useful in the context of individuals and businesses buying AI services (as opposed to being subject to AI outputs, such as through ADM). In engaging with AI products, at this point in time, consumers and (smaller) businesses both lack adequate information to determine if a given use of AI is appropriate in a given context.

One possibility is the use of standard disclosures or 'model cards' specifying various element of the AI being offered. Indeed, we suggest that Governments might promote the use of model cards by making this a condition of procurement.

Model cards are short displays of information provided by developers on release of a model. Model cards may provide information as to the provenance of training data, known failure modes, and basic use information. However, this practice is currently entirely voluntary and lacks consistency with respect to the depth, quantity, and quality of information provided.

Separate model cards should be provided for B2B and B2C offerings respectively. Consumers are not well placed or incentivized to evaluate the many tools they will likely encounter; however clear guidance may improve consumer ability to adopt AI in contexts where it makes the most sense. Model cards that describe intended uses, failure modes, and data provenance should be provided alongside consumer offerings of models that meet certain thresholds.[26]

10. Do you have suggestions for Bans?

We consider that bans should be targeted at uses of AI, not specific technologies. We think there are some uses of AI that should be banned. However, we consider bans should be imposed on the basis of clear compelling criteria and attenuated to context. This approach protects innovation and makes the imposition of a ban more compelling and likely to withstand the test of time. This approach is also consistent with the principles of good regulatory design discussed above. Thought might also be given to sunset provisions on bans (i.e. temporal limit).

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

Subject to the above qualifications we consider bans should be considered for uses of AI with high risk to human rights. In particular, we single out most forms of biometric surveillance in public spaces (eg FRT, iris or gait recognition, emotion detection) and further consider these technologies should only ever be used (if at all) in other contexts a high level of regulatory oversight and control.

We do not consider that bans should be imposed on particular algorithms, software or even hardware. We

deceived-consumers-about-use-facial-recognition-photo>.

[25] Chuan Guo et al, 'Certified Data Removal From Machine Learning Models' (2020) 119 Proceedings of the 37th International Conference on Machine Learning 3832.

[26] We defer discussion as to which models should be regulated to other parts of this submission.

think the potential harms from AI are real. But bans on elements as opposed to uses is futile. For example, bans on high performance compute (i.e. GPUs/TPUs) has sometimes been discussed. Such bans are easily evaded by using different hardware combinations or remote computing to reach the equivalent performance.

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

Relevant considerations for bans might include:

- applications where input into AI/ML models can be easily manipulated (see Adversarial Manipulations above 2.4)
- human rights impacts;
- public policy considerations;
- circumstances where can't rely on consent to justify use or consent is not a valid or proportionate way of justifying use given the impact on fundamental rights;
- concentrations of private power or resources.
-

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

We consider that regulatory regimes requiring good AI governance, including through risks assessments and other accountability measures will increase the trustworthiness of AI.

We also note that there is considerable uncertainty around AI within business and society generally. Therefore, education and training, as we noted in the overview of this submission, to demystify AI, are crucial.

Some degree of clarification about existing law, and even in some contexts the use of safe harbours and sandboxes may prove beneficial for business seeking to innovate with AI, provided these are matched with strong baseline standards for safety and ethical/responsible practice.

We note that industry uptake of the internet was substantially encouraged by a variety of laws, among them Title II of the U.S. Digital Millennium Copyright Act (DMCA) of 1998. This title created a safe harbour from copyright liability for online service providers from user uploaded content—so long as platforms undertook certain requirements to act responsibly. Absent this section, it is likely that substantial portions of the internet's innovations never would have happened. AI finds itself in a similar position.

Implications and Infrastructure

12. How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

We consider there is no competitive advantage in promoting harmful technologies. Australia's technology sector is capable and indeed is developing cutting edge technologies that have beneficial impacts, such as for example in the med tech and ag tech fields.

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

We do not have any comments on this question.

Risk-based approaches

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

We support a risk-based approach to address potential AI risks, however, as discussed above in response to question 2, we think careful consideration needs to be given to an adaptive and flexible approach that is compatible with Australia's existing regulatory regime. The EU AI Act uses a risk-based model. A key critique of the EU AI act is that it is overly prescriptive without being backed by strong oversight and enforcement mechanisms.

Risk based governance for AI should require firms to make a judgement about the risk threat arising from the outputs of their products and implement appropriate governance strategies in response, including as to transparency and accountability. Models for these approaches are already in existence e.g. NIST[27] or in Singapore.[28] The key is to identify mechanisms for requiring these assessments to take place (as a legal obligation) and to ensure they are robust and effective (reporting / monitoring).

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

Risk based approaches have the attraction of being able to be integrated into existing risk management and auditing regimes within corporations, both internal and statutory. Such approaches are not perfect, but would at least it force a practice of identification, quantification, mitigation, and someone having to sign-off on what residual risks remain.

However, we argue that a risk-based approach that relies on predetermined categories of risk, as opposed to a responsive model scaled to use and context as we have advocated in response to question 14, has a number of inherent caveats which require consideration.

Firstly, a key concern with AI surrounds the unpredictability of outcomes. An approach which relies solely on our ability to anticipate risks is unlikely to account for all possible eventualities.

Secondly, determining the threat level will evolve over time, as the technology develops. Maintaining a categorisation that accounts for changing technologies would be challenging, if not infeasible.

We recommend a more holistic approach that takes account of the context of the use and the profile of the entity utilising the AI. Assessment should be dynamic and ongoing.

16. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

All organisations using AI should be using a risk approach. However, we consider that approach should be scaled to the risk presented (rather than the size of the organisation). See also above comments on model cards in response to question 9(b).

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

We support the elements presented in attachment C, but would go further in what may be required from organisations.

[27] 'AI Risk Management Framework' NIST (Webpage) < https://www.nist.gov/itl/ai-risk-management-framework>.

[28] 'Companion to the Model AI Governance Framework' World Economic Forum, prepared in collaboration with Info-communications Media Development Authority of Singapore (Report, January 2020).

- We suggest a more nuanced understanding of the different roles that may be played by concepts of transparency and explanations, see comments in response to question 9.
- We agree with the importance of documentation around data, (e.g. traceability, quality, security etc).
- We support having human oversight of AI systems, particularly those impacting on human rights and interests. However, we consider the language of 'human in the loop' can be misinterpreted and undermine genuinely robust and embedded governance. 'Human in the loop may tend to suggest *a 'tech' person* responsible for the AI system.[29] This can have the effect of devolving responsibility for AI systems from management/directors to lower level individuals or departments in an AI deploying organisations. Human in the loop moreover risks having the supposed exercise of human discretion undermined by the effects of 'automation' bias.[30] We prefer firms adopt structured governance and accountability systems and processes, that ensures responsibility for risks is appropriately devolved across the firm including relevant to risk its leadership/directors.
- While work around transparency/explanations are important, we also emphasise the importance of a focus on monitoring/auditing outputs. This will often for example be a more effective way of testing for bias, drift etc.
- We note the advantages of making risk assessments and other interventions consistent with international initiatives as far as possible and prudent. (For example we point to the NIST initiatives in the US).
-

18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

This can be done by allowing flexibility and a risk response that is tailored to the application and its likely impact.

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

A standard risk-based framework may be difficult to apply to LLMs or MFMs because definitionally these have relatively unrestricted scope of application, making it hard to predict or identify particular risks and impose appropriate guardrails. We consider regulation generally, and a risk-based approach specifically, should focus on the outputs or uses of products that involve these applications (general API targeted at businesses? Public-facing chat-bot/ image generation app?) and rely on existing legal frameworks surrounding business category while demanding current disciplinary best practices for responsible development, deployment etc across the AI lifecycle and supply chain.

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation?

We consider risk assessments and proportionate responses should be mandatory, with reporting requirements imposed upon key stake holders. However, we also consider that the requirements should be principles based in most cases allowing a proportionate and adaptive approach to ex ante demands for responsible and safe AI. See at 2.4.

[29] Jake Goldenfein,'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019)

[30] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller and Liz Sonenberg, 'The Effects of Explanations on Automation Bias' (2023) 322(9) Artificial Intelligence.

And should it apply to:

a. *Public or private organisations or both?*

See above response to q6.

b. *Developers or deployers or both?*

See above 2.4. We consider all initiatives for responsible AI, including risk-based approaches, should apply to across the supply chain but with more demanding obligations on key gatekeepers.