Safe and responsible AI in Australia after generative AI[1]

Peter Leonard, Data Synergies and UNSW Business School[2]

This paper is intended to assist the Department of Industry, Sciences and Resources in the Department's consideration of a bundle of positive incentives, and regulated prescriptions and requirements, to improve AI-affected decisions made by organisations operating in Australia.

This paper responds to the Department's *Discussion Paper on Safe and Responsible AI in Australia*, June 2023.[3]

The Discussion Paper, at pages 34 and 35, asked twenty questions. Section 3 of this paper sets out our responses to many of those questions.

To illustrate the reasoning underling our responses, in section 2 we discuss a case study, being use by a medical doctor or other primary health professional of health information about a patient and a generative AI application[4] such as ChatGPT, as a task-assistant in 'writing up' a clinical note or a discharge summary about a patient.

At the end of section 2 we summarise our insights from this case study as to the appropriate balance of regulation of AI, and other incentives for safe and responsible use of AI. These insights inform our responses in section 3.

We start, in section 1, with a comprehensive overview of our recommendations on how to assure safe and responsible use of AI in Australia.

---

[2] Peter Leonard is a business consultant and lawyer advising data-driven businesses and government agencies. Peter is principal of Data Synergies and a Professor of Practice at UNSW Business School, working across the Schools of Information Systems and Technology Management, and Management and Governance. Peter serves on the NSW Government's AI Review Committee, the NSW Government's Information and Privacy Advisory Committee, and a number of corporate and advisory boards. He is immediate past chair of the AI Ethics Technical Committee of the Australian Computer Society and the Privacy and Data Committee of the Law Society of New South Wales.

[3] The Privacy Act 1988 (C'th) does not require an APP entity to consider whether an act or practice of collection, use and disclosure of personal information relating to individuals is reasonably likely to cause privacy harms to those individuals, whether having regard to mitigations of these privacy harms of those harms, or otherwise. In particular, the APPs do not state how APP entities should determine the circumstances in which rights or interests of individuals in and to privacy are affected, or how to evaluate the nature or extent of harm to those rights or interests for the purpose of application of the APPs. This leads to inappropriate focus by APP entities upon notice and consent, with the attendant limitations and shortcomings of that framework, to the neglect of proper consideration of risks of data privacy harms to affected individuals, and assessment and mitigation of risks of privacy harms. As both data privacy harms and AI are caused by uses of data, there is significant overlap between commonly accepted 'privacy harms' and 'AI harms'. As to privacy harms, see Data Synergies (Peter Leonard), 'Privacy Harms: A research paper for the Office of the Australian Information Commissioner', June 2020, https://www.oaic.gov.au/__data/assets/pdf_file/0012/1371/privacy-harms-paper.pdf; Danielle Keats Citron and Daniel J Solove, 'Privacy Harms', 102 Boston University Law Review 793 (2022), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782222. As to AI harms, see the references cited in footnote 10 below.

[4] In this paper we generally follow the Discussion Paper's approach of using the terms "general purpose AI system" to refer to both underlying foundation models and applications and services based on those models, and "generative AI application" to refer to an application or service that is enabled by use of ML foundational models such as LLMs and MFMs. See also Elliot Jones, 'Explainer: What is a foundation model?', Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

1. Overview and summary

This paper advocates a focus upon responsibility and accountability of each organisation to assess decision provenance - the interaction of people, processes, data and technologies that affects the reliability, quality and safety of a decision – for each decision by or for an organisation that is materially AI-affected. Each assessment needs to be in and for a particular decision context. Much of this paper addresses how decision context should be evaluated and applied in assessment of risks of AI harms and mitigation measures. We explain how enforced self-regulation by and of organisations should play a key role in ensuring that diverse organisations consider, internalise, address and mitigate risks of AI harms.

Assuring realisation of AI-enabled productivity and other benefits to Australian society and the Australian economy, without unacceptable externalities and other unacceptable harms to humans or the environment, requires novel initiatives by Australian governments to ensure that AI-affected decisions by organisations are safe and responsible. Regulated prescriptions and requirements are part, but only part, of the story: empowerment of organisations to sensible evaluate risks of AI harms, and education and training of users of AI (and in particular consumers and other end users using generative AI), are also important. But uses of AI are already transforming many Australian workplaces: there is a need for prompt, but not precipitous, policymaking.

Step one is to assist diverse Australian organisations to find and implement practical strategies and programs to evaluate proposed uses of AI and to exercise control over whether, when and how AI is used by those organisations.

Implementation of safe and responsible AI by Australian organisations requires:

- top down, organisations to adopt new policies and approval frameworks,
- in the middle, adoption of practical tools to evaluate, risk mitigate and control uses of AI. These practical tools may be complex structured frameworks and methodologies for AI assessment, or more simple checklists and analytical tools such as decision trees.
  The tools needs to be appropriate to the audience, being responsible and accountable officers within organisations who need to understand and address the diversity of risks of harms to humans and the environment that may arise from the organisation's adoption and use of AI.
  It is important that practical tools are designed to enable organisations to continuously self-evaluate and re-evaluate their evolving uses of AI.
  Complex structured frameworks and methodologies for AI assessment may be suitable for use by larger organisations and those other organisations that can afford to outsource assessment to risk advisory and compliance practices of external consultancies. However, outsourcing will not change the ongoing, internal competencies of organisations, and each organisation's decision-making DNA, to the

extent required to address rapidly evolving capabilities of AI, and tasks for which AI will be used, and

- bottom up, education of prospective users of AI outputs within each organisation as to risks of inappropriate use and reliance upon AI, and in particular general purpose AI.

This paper suggests specific top-down, in-the-middle, and bottom-up initiatives, as detailed in our response to question 3, in section 3 of this paper.

There has been extensive discussion in Australia about:

- principles that should be applied to ensure that uses of AI are fair, equitable, accountable and transparent,
- further attributes or characteristics of 'safe and responsible' uses of AI,
- which applications and use cases for AI should be listed as so extreme risk of harms to affected humans or the environment that cannot be controlled or mitigated as to justify regulatory prohibition ('blacklisting'),
- which applications and use cases for AI should be characterised as potentially leading to risk of harms to affected human or the environment of a level that requires a structured risk of harms evaluation and consideration of mitigation measures,
- the extent to which new regulation is required to create sufficient incentives for organisations (1) to conduct an initial 'gating' evaluation of whether a structured risk of harms evaluation is appropriate, (2) to conduct that evaluation, and (3) to implement appropriate mitigation measures,
- whether there should be a new 'AI regulator', or re-allocation of regulatory functions and priorities between existing regulators in order to better address AI', or improved coordination and principles setting across regulators.

As well as canvassing proposals for new AI regulation, the Australian government should promote discussion as to how Australian federal, state and territory governments, through initiatives of government agencies and regulators, can effectively influence decisions by regulated entities, including other government agencies, businesses and not-for-profits, as to whether, when and how to safely and responsibly use AI.

Many submissions to this current consultation will address these areas and will therefore express views as to whether Australia should model regulation of AI on the more interventionist and prescriptive proposed EU AI Act[5] or the draft Canadian AI and Data Act[6], or the UK's proposed lighter touch, coordinated but decentralised approach.[7]

[5] References to the EU AI Act are to the draft compromise text of 16 May 2023, being the current draft as at 18 July 2023 and available at https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf.

[6] The draft Artificial Intelligence and Data Act (AIDA) introduced as part of Bill C-27, available at https://www.parl.ca/legisinfo/en/bill/44-1/c-27. The Bill remains subject to a consultation process outlined by the Government of Canada in a companion document: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

[7] UK Department for Science, Innovation and Technology, A pro-innovation approach to AI regulation, 29 March 2023 (Command Paper No. 815), https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. For recent analyses of the UK approach, see Matt Davies and Michael Birtwistle, 'Regulating AI in the UK', Ada Lovelace Institute, 18 July 2023, https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/; Elliot Jones, 'Keeping an eye on AI: Approaches to government monitoring of the AI landscape', Ada Lovelace Institute, 18 July 2023, https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/.

We advocate a lighter touch, coordinated but decentralised approach to AI regulation in Australia, but on the basis that:

- regulation should focus upon ensuring that organisations design and implement policies and programs for responsible uses of AI that are appropriate to the organisation and not the manner in which those policies and programs are implemented. Regulation should incorporate sanctions that create the incentive to ensure that the policies, programs and risk assessment methods are effective and reliably and verifiably implemented. See further our discussion below of the model of enforced self-regulation.
- regulation of uses of AI should be closely targeted to some uses cases of AI, through sector specific or use case specific prescriptions or prohibitions that are tailored to address the particular context of use ('data context' or 'AI context'),
- economy-wide prohibitions and prescriptions should build from existing statutes and other laws, notably by (1) building out from existing provisions of Australian Consumer Law[8] and the Privacy Act 1988, in order to more clearly address transparency as to limitations of particular AI applications and services, and unlawful discrimination prohibitions under human rights statutes, and (2) further statutory prescriptions addressing AI-assisted targeting of misinformation and disinformation and excessive surveillance and profiling,
- specific or use case specific prescriptions or prohibitions should not be a response to pressure to address perceived existential threats 'from AI', but be targeted to circumstances where other incentives to improve AI-affected decisions made by organisations leave unacceptable risks of harms to humans or the environment,
- a number of areas should be considered as priority areas for statutory reform, regardless of whether implementation of risk management assessment and management is mandated in relation to specified categories of AI uses (i.e., high impact systems). Priority areas might be blacklisting (prohibition) of certain manifestly harmful and therefore societally unacceptable uses of AI; new documentation, transparency and disclosure requirements applying to offering of AI systems and commercial uses of AI systems and amendments to Australian Consumer Law to ensure coverage of provision of AI applications and services;[9] statutory changes to address appropriate allocations of AI liability across supply and use chains[10] and

[8] A number of approaches are possible: one might be to supplement ACL prohibitions on misleading or deceptive conduct (e.g., ACL s 18) and false or misleading representations (e.g., ACL s 29) with statutory presumptions that persons offering or using AI systems in the course of trade are presumed to represent compliance with certain prescribed 'safe and responsible AI' principles and prescriptions.

[9] See our discussion in our response to question 3 of the Discussion Paper, in section 3 of this paper.

[10] The draft EU AI Liability Directive of 28 September 2022 proposes to make it easier for claims to be brought for redress of harm caused by AI systems and the use of AI. The proposal addresses the specific issues with causality and fault linked to AI systems. See https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf. For a policy analysis as to allocation of responsibility and accountability in AI supply chains, see Ian Brown, 'Allocating accountability in AI supply chains', Ada Lovelace Institute, 29 June 2023, https://www.adalovelaceinstitute.org/resource/ai-supply-chains/; Jennifer Cobbe, Michael Veale, and Jatinder Singh, 'Understanding accountability in algorithmic supply chains', June 2023, https://doi.org/10.1145/3593013.3594073

addressing evidentiary burdens of providers and commercial users of AI systems as to safety of use of those systems.[11]

We suggest that policymaking for regulation of AI follow similar principles to those proposed by the UK government, being:

- safety, security and robustness: The UK Competition and Markets Authority (CMA) in the CMA's response[12] to the UK government's 2023 White Paper stated that organisations in properly functioning markets should "face the correct incentives to determine and implement the appropriate level of security and testing to ensure that their systems function robustly", and the CMA may need to intervene when this incentive is missing, i.e., when "AI use affects a consumer who may not be in a position to assess technical functioning or security of the product",
- appropriate transparency and 'explainability',
- fairness: the CMA noted that this principle should be applied to "the context surrounding the AI system", including data collection, testing and evaluation practices, and not just any underlying algorithms or AI functionality,
- accountability and governance of organisations deploying AI,
- contestability and redress: the CMA noted that the "opacity of algorithmic systems and the lack of operational transparency" make it hard for customers to "discipline" firms, and stressed the importance for regulators to effectively monitor potential harms and to have the powers to act where necessary.

We particularly commend the UK government's emphasis upon the creation and operation of a bundle of positive and negative incentives for organisations that are providing or using AI in their operations "to determine and implement the appropriate level of security and testing to ensure that their systems function robustly".[13] Perceptions by some organisations that harmful effects of their uses of AI are externalities suffered by others, rather than the organisation, may need to be changed through government action to create new, or changed, negative incentives that cause those organisations to internalise risks of AI harms and then be stimulated to mitigate these risks.

Most of the initiatives that we advocate require reimagining how, why, and by whom, risks of AI harms are evaluated by organisations.

These initiatives require more than simple re-tooling or modifications of enterprise risk or project management frameworks and methodologies as today in common use by individuals within organisations that already have developed capabilities to use those methodologies.

[11] Adapting from recommendations of the Australian Human Rights Commission in its Human Rights and Technology Final Report, 2021. "Recommendation 11: The Australian Government should introduce legislation that provides a rebuttable presumption that, where a corporation or other legal person is responsible for making a decision, that legal person is legally liable for the decision regardless of how it is made, including where the decision is automated or is made using artificial intelligence." https://tech.humanrights.gov.au/artificial-intelligence/ai-informed-decision-making

[12] https://www.gov.uk/government/publications/response-to-governments-ai-white-paper

[13] In the context of this Discussion Paper, the UK use of the term 'robustly' equates to responsible evaluation of reliability of outputs of AI systems for safe outcomes upon humans and the environment resulting from likely reliance placed upon those outputs.

Safe and responsible uses of AI for myriad tasks across diverse organisations cannot be assured by modifying roles and responsibilities of currently designated privacy professionals, or the frameworks and tools that they use to conduct privacy impact assessments.

Safe and responsible uses of AI will not be reliably assured by organisations:

- beefing up current second or third line of defence[14] functions or capabilities,
- re-purposing privacy officers as AI officers, or
- outsourcing AI assessment to large professional services consultancies for episodic review of 'AI projects'.

Assurance of safe and responsible uses of AI needs to become part of the DNA of each organisation - public and private, business and not-for-profit, large and small - and consistently and reliably applied in the course of each organisation's business-as-usual processes.

A suitable regulatory model to effect changes in organisational DNA is **enforced self-regulation of and by organisations**.

Organisations could be required to develop and implement policies and programs to act responsibly and ensure safety in organisational uses of AI. As a minimum, organisations should be required to prepare an annual plan setting out what they propose to do about ensuring safety in organisational uses of AI, including specification of reasonable precautions that the organisation is putting in place.

This mandate could be supported by mandated transparency requirements, i.e., to publish risk of AI harms policies, and overviews of risk of AI harms programs.

There would need to be associated meaningful legal exposures for organisations, and their directors and their senior officers, in the event that the organisation:

- did not develop and oversee reliable implementation and operation of policies and programs,
- did not take reasonable precautions to mitigate reasonable foreseeable risks of AI harms, or
- did not comply with transparency requirements.

However, the legislative requirements would not be prescriptive as to the processes by which risk management is implemented and conducted within the organisation, including as to the form of and structured risks of AI harms assessment.

The enforced self-regulation approach thereby enables flexibility in how organisations address particular categories of types of AI risks. For example, risk assessment and

[14] See further Chartered Institute of Internal Auditors (IIA), 'Position paper: Risk management and internal audit', https://www.iia.org.uk/resources/risk-management/position-paper-risk-management-and-internal-audit; Deloitte Advisory, 'Modernizing the three lines of defence model: an internal audit perspective', https://www2.deloitte.com/us/en/pages/advisory/articles/modernizing-the-three-lines-of-defense-model.html

mitigations for availability of generative AI as a task assistant for myriad tasks performed by various staff members across an organisation could be quite different to risk management of single-task AI managed into an organisation following a structured project management process.

In addition to enforceable responsibilities of organisations and their directors and their senior officers (and meaningful legal exposures) under an enforced self-regulation model, we suggest a requirement that each organisation that is operationally using AI designate a senior responsible officer with responsibility to implement policies and programs to ensure safety in organisational uses of AI.

Ideally, that person will care about ensuring safe and responsible AI, and will have skills, authority, knowledge and access to practical tools to assure safe and responsible AI.

Of course, that person would not act alone: assurance of safe and responsible AI requires a multidisciplinary and cross disciplinary team approach that adapts existing ways of doing things and supplements existing risk frameworks, methodologies and tools.

Assuring safe and responsible AI also requires recognition that many organisations in Australia that will implement AI over the next three to five years will not have internal capabilities or resources during that period to either:

- translate high level 'principles' or 'guidance' as practical steps for risk assessment and management, or
- understand and use the cumbersome, detailed and prescriptive AI impact assessments recently promoted by many 'responsible AI' proponents, including some external consultants selling services of outsourced conduct of 'responsible AI' assessments.

The Australian government's policy and regulatory responses over the last decade to data security threats illustrates what Australian governments should not do.

Safe and responsible AI cannot be assured through a confusing proliferation of policy and legislative requirements, or through conflicting requirements imposed in parallel by multiple responsible regulatory agencies.

Safe and responsible AI does not require a 'super-regulator for AI', or 'super AI rules'. In this respect, assuring safe and responsible AI across Australian organisations is quite different from improving cyber-resilience of Australians.

The Australian Government's assessment of incentives and prescriptions for safe and responsible AI should focus upon the contexts in which AI is used by people in processes of decision making by organisations operating in Australia.

A test for whether and how Australian government should, by legislative mandate, regulator-driven prescriptions, or enforced self-regulation, intervene to require Australian organisations to assess AI-affected decision provenance, should be whether the intervention is:

- necessary and proportionate, and
- reasonably likely to lead to better outcomes for the Australian economy and Australian society than would have been the case had the Australian Government not intervened.

A key question for design of regulation to require organisations to address risks of AI harms is a policy evaluation of whether organisations are likely to consider that harms to others from uses of AI are:

- externalities that do not need to be addressed and mitigated by the organisation, or
- sufficiently internalised by responsibility and accountability of the organisation and its controlling minds (i.e., boards and senior executives), either:
  - through legal compulsion (need to comply with laws) and associated legal exposures, or
  - through operation of other incentives, such as damage that an organisation or those controlling minds are likely to suffer if the organisation does not properly mitigate risks of harms.

Applying an AI-affected decision context approach, an AI policy framework should start with a threshold analysis:

> *In:*
>
> *each decision context in which AI is being used in circumstances where a reasonably foreseeable outcome is that humans or the environment may suffer significant harm as a consequence of that decision,*
>
> *if:*
>
> *individuals causing an organisation to take AI-affected decisions, or the controlling minds of those organisation, were properly informed that AI is affecting the decision and as to risks of AI-affected decisions,*
>
> *are the incentives sufficient to ensure that in most cases where that decision context arises those individuals will reliably mitigate risks of harms to the point where, objectively assessed, residual risks of harms are very low?*
>
> *If the incentives are not sufficient to lead to that outcome, what should Australian governments do to cause these incentives to change?*

Incentives should be evaluated having regard to:

- whether the risk of harms is a transitional issue that is likely to fade once Australian organisations become familiar with the categories of new risks of AI harms and how to avoid and mitigate those risks, and
- the likely outcome of the combined interaction of top-down, in the middle and bottom-up initiatives undertaken to change the weight of incentives and disincentives.

At a time of unpredictable change, regulated prohibitions and prescriptions have a role to play as part of a broader, (positive and negative) incentives-based, policy response to new uses of AI. Regardless of the operation of the incentives structure in relation to actions of most organisations, some organisations will be irresponsible, or knowingly or negligently commit or sanction illegal harms. New prohibitions and prescriptions may be necessary to deter or punish irresponsible or bad actors. However, interventions need to be measured and adaptive, given rapid developments in functionality and reliability of AI and the myriad use cases now being trialled for use of generative AI applications. Where AI regulation is justified, there needs to be consideration as to striking the right balance between:

- **prohibitions** - AI must not be used for a specified AI-affected activity,
- **before the event (a priori) prescriptions** - AI may only be used for a specified AI-affected activity if a regulated entity first complies with specified preconditions, i.e., conduct of an AI impact assessment,
- **before-the-event requirements**, i.e., for transparency, for organisations to develop and implement policies and programs to act responsibly and ensure safety in organisational uses of AI, nomination of a responsible officer, and due consideration by or for the responsible senior officer of a regulated entity of relevant possible significant risk of harm factors[15] even where the requirement for consideration does not extend to a formal structured process for risk of harms evaluation such as conduct of an AI impact assessment, and
- **after the event (ex post) legal exposures** to damages and penalties, and the appropriate level of transparency to enable detect-and-respond (remedy) incentives to operate, or to enable detect-and-prosecute. Detect-and-respond (remediate), and detect-and-prosecute, are of course closely related. Both require (1) detection of an AI harm, (2) an evidence trail for root case analysis and for allocation of accountability, and (3) a relevant party/parties willing and resourced to respond, whether that party is the one causing the harm, a party suffering the harm but able to avoid further occurrence of the harm, and/or the regulator.

A sensible approach to AI regulation is to ask whether rules that restrict or prohibit particular uses of AI, or that mandate application of a particular risk assessment framework or methodology, are justified, or whether 'detect' and 'respond' incentives as adjusted for AI would then provide sufficient incentives to cause appropriate mitigation of AI risks by regulated entities.

[15] Such as those risk of harm factors listed in 'Box 1.2: Illustrative AI risks' of the UK Government's March 2023 Policy Paper, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. See also the lists and associated discussion in Part 3 (Harms, risks and perceptions of AI systems, pages 24 to 31) of the UTS Human Technology Institute, The State of AI Governance in Australia, May 2023, https://www.uts.edu.au/human-technology-institute/news/report-launch-state-ai-governance-australia. Another useful recent review of possible consumer harms from uses of generative AI is the Norwegian Consumer Council's June 2023 paper, 'Ghost in the machine – Addressing the consumer harms of generative AI', available at www.forbrukerradet.no/ai. See also AWO, Report for the Ada Lovelace Institute, 'Effective protection against AI harms', 18 July 2023, available at https://www.awo.agency/blog/awo-analysis-shows-gaps-in-effective-protection-from-ai-harms/; Ada Lovelace Institute, 'AI risk: Ensuring effective assessment and mitigation across the AI lifecycle', 18 July 2023, https://www.adalovelaceinstitute.org/report/risks-ai-systems/; Harini Suresh and John Guttag, 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle', 2021, https://doi.org/10.1145/3465416.3483305

Many Risks of AI harms can be addressed by getting the detect and respond incentives right, and therefore an upfront restriction or prohibition is not required or justified.

In any event, it is very difficult to design appropriate and futureproof upfront restrictions or prohibitions for AI applications, given the rapidly evolving and unpredictably diverse ways in which AI systems, and in particular generative AI applications, are already being used to assist humans in performance of myriad tasks.

If all of the following conditions apply:

- prompt detection of a significant AI harm to humans or the environment is likely,
- financial recompense to affected persons that have suffered that harm is likely to be appropriate to fully redress their loss or damage, and
- prospective defendant regulated entities are incentivised to properly mitigate risks of relevant harms, because:
  - potential awards of damages, or regulator-imposed penalties are appropriately substantial, and
  - commencement and successful prosecutions of actions for recovery of damages and penalties are sufficiently likely,

then before-the-event prescriptions or prohibitions should not be required or justified.

When any of the above listed conditions are not met in relation to listed, specified AI-affected activities that reasonably likely to lead to significant harms to some humans or the environment, then AI regulation, in the form of before-the-event prohibitions or prescriptions or both, may be justified. In that case, it is still appropriate to ask whether these risks of harms are:

- short-term and transitional - likely to be addressed through other government and industry initiatives to improve understanding as to responsible AI practices, or
- longer term or more entrenched and therefore requiring more intrusive regulation.

In relation to categories of AI activities where all of the above listed conditions are met, before-the-event requirements may still be necessary or desirable, to ensure that 'detect' and 'respond' incentives operate to assure safe and responsible AI.

Transparency is particularly important in ensuring prompt detection of a significant harm to human or the environment, and attribution of that harm to a particular AI activity conducted by a regulated entity.

For those activities for which an AI risk assessment is prescribed, a requirement for transparency of that fact that an AI risk assessment has occurred, although not necessarily as to the form or required content of the AI assessment, may create sufficient incentives to ensure that risks are mitigated by regulated entities.

Transparency is particularly appropriate in circumstances where one entity in an AI-affected decision chain ought reasonably to expect that an upstream entity in that chain should have reasonably anticipated that a downstream entity would rely upon the upstream entity to

assess and mitigate particular risks, or to warn relevant downstream entities as to residual risks. However, this is not the case for many tasks for which general purpose AI may be used. Our case study in section 2 illustrates this issue, and the challenge that this issue creates for allocation of responsibility and accountability in mitigating risks of AI harms.

2. Case study in practical good governance of general purpose (generative) AI: clinical notes and discharge summaries in Australian hospitals

There are many clinical evaluation and diagnostic tasks where algorithmic inferences or more advanced AI are already being used as an aid to medical doctors and other health professionals (collectively, clinicians) working in Australian hospitals.

Our case study involves a clinician using health information about a patient, and a general purpose AI application such as ChatGPT, as a task-assistant in 'writing up' a clinical note about that patient, or a discharge summary.

Almost all medical professionals, in all hospitals, complain that too much of their working day, each and every day, is spent writing up clinical notes and discharge summaries.

Many of the details in clinical notes involve exercise by a clinician of judgement in interpretation of factual inputs such as outputs from bedside monitors, observations of vitals and other quantitative readings, medications administered, food and fluids consumed, etc..

Discharge summaries bring together:

- relevant health information derived from prior clinical notes,
- other information about a patient, typically derived from a hospital's electronic medical record system,
- a clinician's assessment of the patient's prognosis and the clinician's recommendations as to future care and medical interventions.

At the outset, it should be noted that inputting of health information about an identifiable client by way of prompting a generative AI application generally will involve exposure of sensitive health information about an identifiable individual to the AI application.

If this exposure is an uncontrolled disclosure to a person or entity outside of the data controlling entity (viz., outside the health service for whom the clinician is working), and this disclosure is not with informed consent of the patient, there is likely to be a breach of health information privacy statutes that apply to that health service, and a breach of the clinician's duty of confidentiality to the patient.

Avoiding such breaches would require implementation of robust and verifiably reliable privacy and security by design controls and safeguards in relation to the data used to

'prompt' the generative AI. However, our case study scenario is realistic. Some providers of generative AI applications are now standing up 'air-guarded' local and client-controlled instances of generative AI for large clients such as operators of hospital systems. Other providers, notably including Meta, are making available open foundation models, including LLMs that may be stood up within an organisation, with tools that enable an organisation to itself train and control use of the LLM within the organisation to stand-up an organisation-specific AI application, whether general purpose or task specific. Accordingly, this fact scenario might take place in Australia today, with a health system deploying an air-guarded LLM, possibly trained on large data sets from outside the operator health system, but not contributing prompting data back to the AI provider's data corpus. If this is the relevant data context, sensitive health information about an identifiable patient will not leave the control of the operator of the health system.

Why would a clinician use a generative AI application to assist the clinician in 'writing up' a new clinical note, or a discharge summary?

In common with many tasks for which generative AI is being trialled or operationally used today, the principal benefit to the clinician would be reduction of time spent in performing the laborious transcription and summation elements of performance of the 'writing up' task.

However, there are also other benefits.

Risk of human errors through mis-transcription of observations, or arising through inconsistencies in terminology used in clinical notes, may be mitigated.

Use of the AI may assist the clinician to standardise language to SNOMED Clinical Terms[16], facilitate comprehensiveness of coverage, or improve succinctness and readability of the note or summary.

At the same time, risks of new machine generated errors are introduced. As now commonly understood, use of generative AI also introduces a new possible source of error. Current generation large language models (LLMs) are built upon likely association of words in machine-inferred like contexts, and not by context-specific rational and evaluative judgement of how words should be used in a particular data context.

The risk of harms from this use of AI in this fact scenario is clear. The AI output may have machine-introduced errors. Those errors may not be detected by the clinician. The errors may therefore continue into the clinical note or discharge summary. Another clinician reading the note or summary may rely upon the erroneous note or summary in determining whether and how to treat a patient. Accordingly, there may be unsafe reliance upon an AI output, that leads as a consequence to harm to a patient, unless the risk of machine-introduced errors has been appropriately risk managed through appropriate governance and associated assurance controls.

[16] See further https://www.snomed.org/

There is a further risk that a suitably informed clinician may be 'lulled to sleep' through reliable results derived from the clinician's earlier use of the generative AI application in relation to other patients. For generative AI applications, prior reliable AI outputs are no reliable guide to reliability of a particular future AI output in the same clinical setting but in relation to a different data context. The 'lulling to sleep' risk may increase over time, as new generations of LLMs become reliable for a statistically greater proportion of AI outputs. However, stochastic errors will remain, because of the inherent (word association) nature of LLMs.

The incentives for good governance of uses of AI in clinical environments are already substantial. Clinicians do not wish to make errors in their exercise of professional judgment. Health services controlling clinical environments already face substantial legal exposures if the health service fails to assess and mitigate reasonably anticipated risks of harms to patients that might be occasioned through risky work practices within those clinical environments. Given the operation of these incentives, it is therefore not self-evident that any new or further legislated prohibition, or regulatory prescriptions, are required to address this fact scenario.

Good governance of a clinical environment should involve a responsible senior officer:

- determining whether risks of machine-introduced errors are allowed into the systems of work (various decision chains) within that clinical environment, and
- if risks of machine-introduced errors are allowed in, that the risk management system settings (governance, safeguards and assurance controls) are appropriate to mitigate those risks and manage residual risks.

What assurance controls might be implemented by the health service? For this data context, assurance controls might be:

- the AI outputs are evaluated by a human that is applying context-specific rational and evaluative (professional) judgement,
- that judgement includes proper appreciation by that clinician of the risk of error in an AI output as a result of the AI output's word association being inappropriate or simply wrong in the individual patient's clinical context, and
- that the AI outputs are reliably reviewed and revised by each clinician that may incorporate those AI outputs into the clinician's patient's notes or discharge summary, each day and every day.

Many risk mitigation steps are possible and practicable, including further training of clinicians as to possible incidence of machine-introduced errors, and error detection. By their nature, clinical notes and discharge summaries are prepared by health professionals that are trained to follow familiar, well understood and circumscribed protocols and conventions. Those health professionals prepare those notes and summaries in the knowledge and expectation that these notes and summaries will be relied upon by other health professionals. Therefore, there is a 'human in the loop' of a known decision chain, and that human is applying certified skills and professional judgement to execute a familiar

process. The new risks in this fact scenario are that the clinician may not have a proper appreciation of the risk of error in a generative AI output, or the need for that clinician to mitigate that risk. Although the decision chain is familiar and unchanged, unfamiliar risks have been introduced at one link in that decision chain. These risks need to be mitigated. The decision of another clinician later in this decision chain is therefore potentially an AI-affected decision. For example, that other clinician may be a patient's GP who is reading a discharge summary and relying upon that summary as a factual actionable insight in making a decision as to what further treatment of that patient is appropriate.

In this fact scenario, the downstream clinician in the clinical decision-making chain cannot reasonably be expected to anticipate and assess risk of machine-introduced error upstream in that decision chain. It would also not be appropriate to deal with this risk through warning or disclosure published by the upstream clinician that using the generative AI to the world of possible downstream clinicians. Only that upstream clinician has access to the full data inputs and can therefore undertake the necessary checks to ensure that the clinical note or discharge summary is 'safe': that is, fit for the purpose of reliance by a downstream clinician upon the factual correctness of the note or summary. Clearly, that reliance ought reasonably be anticipated in relation to any discharge summary. It is therefore appropriate that the relevant risk is mitigated before finalisation of note or summary, by an assurance control (careful checking of the AI output against the relevant data inputs and other information in the patient's file) undertaken by the author of that note or summary (the clinician using the generative AI).

As already noted, good governance of clinical environments involves decisions by the controller of that environment as to what technologies are made available for use in that environment, for use by whom, and with what instructions, warnings and training as to safe and responsible use. To date, the many algorithmic clinical decision-making tools used by doctors and nurses in primary health facilities usually pass through procurement project evaluation – 'project gating' - conducted by or for the relevant health service. Many of these tools by their nature are also regulated therapeutic devices and thereby safety-certified, with associated documentation as to known limitations as to reliability and appropriate reliance.

In the current use case example, a clinician may have self-served the generative AI application. Project gating may not have occurred. The clinician, before deciding to self-serve and use generative AI as a task assistant, may or may not may have read relevant reports as to errors in AI outputs. The clinician is unlikely to have read any AI model card. In any event, the provider of the relevant generative AI application, when determining the coverage and content of the AI model card, cannot reasonably be expected to anticipate all tasks for which a general purpose generative AI application may be used by humans. Accordingly, that provider's documentation of known limitations may, or may not, be appropriately scoped and on point for this clinical task.

Having regard to the above identified risks, preliminary conclusions might be that generative AI is unsafe in a clinical setting, and the risk of significant harms to humans is such that a

prohibition, or a regulatory prescription as to rigorous pre-conditions for use, is required. The risk of error, and of consequential harm to patients, that is introduced through use of generative AI should be objectively evaluated, and the precautionary principle applied.[17] However, we should also consider:

- risks in the counter-factual of how things are done today, and
- possible mitigations, through implementation of good AI governance and assurance controls, of risks of generative AI-induced harms.

Errors occur in current, manual human processes, including through mis-transcription of observations, and inconsistencies in terminology used in clinical notes. Pre-AI, there is an expert human assessing data inputs, applying clinical judgment through interpretation and presentation of human-generated outputs, that are then presented to a known and circumscribed audience of other health professionals that rely upon that clinician's judgment. Post-AI, there is a new potential source of error, because the data inputs are conveniently and quickly transformed, albeit possibly creating new errors, and presented as attractive conclusions ready for acceptance. However, it is the same expert human assessing the AI outputs as would otherwise be assessing pre-AI data inputs and manually transforming those data inputs through clinical judgement into a summation. Post use of generational AI in this clinical setting, the pre-AI data inputs remain available to the clinician, who should then be exercising professional judgement in assessing reliability of the AI outputs as a fair summation of the pre-AI data inputs. Further, in this clinical setting, for both pre and post use of generational AI, the audience of other health professionals reading the clinical notes or discharge summary rely upon the professional judgment of the author of those notes or summary. Accordingly, the relevant 'AI-affected decision' should stop with the exercise of professional judgment of the author of those notes or summary.

Moreover, because the data inputs are conveniently and quickly transformed, the clinician using the generative AI (if appropriately trained and alert) should have substantially more time available to detect such errors, being time freed-up through reducing time on the laborious routine of preparing first drafts of the patient's notes and discharge summary. If appropriate safeguards and associated controls are implemented – including education as to possible first draft generative AI errors - reduction in risk of human error, and increased time available to interpret factual inputs, might be expected to lead to better overall clinical outcomes, as well as improving the work life of many healthcare professionals.

A cautionary note is that otherwise freed-up time for the clinician may be recaptured by the employer health service. Like many other human driven business processes, technologically enabled improvements in human productivity may or may not positively correlate with better quality assurance. A health service may allow a clinician less time to do 'paperwork', capturing a time gain but precluding proper review of AI outputs by the clinician. A health

[17] There are various and not fully consistent statements of the precautionary principle as applied in medical practice. One statement is that where a clinician does not have compelling evidence relating to the potential outcomes associated with a treatment choice, the clinician should take reasonable measures to avoid harms that are serious and plausible. The reasonableness of a response to a threat depends on the clinician's evaluation of factors including benefit versus harm, realism, proportionality, and consistency.

service may expect the clinician to use freed-up time by conducting more ward rounds, rather than properly evaluating AI outputs against data inputs. In short, workplace governance needs to enable the benefits of AI assistance to be captured without adverse outcomes for professional wellbeing of clinicians or the health of patients.

This use case provides some useful insights for governance of AI in this clinical context:

- Risks 'from AI' are highly specific to the context of use, and in particular affected by the skills, ability and willingness of the human in the loop to evaluate the safety and reliability of the AI outputs for the reliance placed upon those outputs.
- Clinicians will require new skills to properly evaluate the safety and reliability of use of generative AI as a task assistant in preparation of clinical notes and discharge summaries.
- Given the importance of human evaluation of the suitability and safety of a particular general generative AI service for use in a particular data context and for a particular task (that is, a particular AI-affected decision chain), the coverage and content of the AI model cards and other AI-related disclosures and warnings should address reasonably anticipated susceptibility to errors or other known limitations of the AI for tasks generally, in terms reasonably transparent to a non-technical specialist reader.
- However, providers of generative AI services cannot reasonably be expected to anticipate all tasks for which multitask general generative AI may be used by humans, when determining the coverage and content of the AI model cards and other AI-related disclosures and warnings.
- For any AI-affected decision chain, it is reasonable to postulate that if a relevant risk may reasonably be expected by a downstream 'human in the loop' to have been reliably mitigated at an earlier link in the decision chain, later decisions in the relevant decision chain should not be regarded as AI-affected.[18]
- Health services and other employers of clinicians already have legal obligations relevantly:
  - to ensure that staff that are allocated responsibility to perform particular tasks have appropriate skills and training to perform tasks in the manner in which they might reasonably be expected to perform those tasks,
  - to ensure that staff are instructed as to what not to do while performing those tasks, and
  - to put in place appropriate governance and associated assurance controls (i.e., risk management systems) to increase the likelihood that tasks are only performed as reasonably expected (viz., appropriately risk mitigated and with review, detect and respond settings that address residual risks).

[18] To put this proposition another way, risk mitigations that should be applied to use by clinicians of generational AI in this clinical setting should have the effect that the audience of other health professionals reading the notes or summary are indifferent as to whether the author of those notes or summary used generative AI as an aid in creation of those notes or summary. If risk mitigations are objectively evaluated as having the effect that risk of material errors in fully manual notes or summary is similar to LLM-assisted clinician notes or summary, there should not need to be AI-related disclosures and warnings for that downstream audience. The upstream assurance control for this fact scenario - careful checking of the AI output against the relevant data inputs and other information in the patient's file, undertaken by the author of that note or summary – is reasonably practicable.

Given the operation of these incentives, it is therefore not self-evident that any new or further legislated prohibition, or regulatory prescriptions, are required to address this fact scenario.

- Generative AI poses new challenges for governance because self-service availability to users (in this case, the clinician writing the note or summary) may circumvent an employer's (in this case, the health service's) project gating controls over whether AI is available for use by employees in a workplace (the clinical environment).
- For each data context, consideration needs to be given to whether the incentives are appropriate to ensure that controllers of the relevant work environment regain control and implement governance and associated assurance controls effective to ensure that employees performing workplace tasks do not use self-service AI applications in any way that are unsafe, irresponsible or illegal.

3. The Department's questions (*in italics*), and summary answers

Definitions

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

The definitions on page 5 of the Discussion Paper are a useful guide to discussions as to when risks of harms arise from use of inference engines, including algorithmically enable automated decision making, use of foundational models and generative AI applications.

In considering possible AI harms and incentives to assess and mitigate relevant risks of harms, a broad definition should be used. As Robodebt illustrates, many significant harms to humans or the environment may flow from inadequate governance of use of hardcoded advanced data analytics (algorithmic) systems to inform or otherwise affect human decision-making or produce automated outcomes. These uses therefore need to be brought within the ambit of new frameworks and methodologies for assessment of governance and associated controls for inference-assisted decisions, regardless of whether advanced data analytics (algorithmic) systems are considered as 'AI'.

ML foundational models introduce further categories of risks, as compared to hardcoded advanced data analytics (algorithmic) systems.

Generative AI applications built on foundation models introduce further categories of risks of harms to individuals or the environment.

For a particular AI-affected decision context, the level of assessed risks of harms may be such that a comprehensive, data context-specific, risk assessment should be conducted, in order to either:

- specify appropriate mitigation measures, or
- determine that for that decision context, mitigation measures cannot reliably and verifiably reduce risk of significant harms to low or remote, and accordingly that AI should not be used in that decision context.

In other data contexts, the risk of harms may be sufficiently clear from a desktop review that appropriate governance and assurance controls can be determined without a form context specific risk assessment.

It is generally not possible to determine in advance whether employing a technology, whether hardcoded advanced data analytics (algorithmic) systems, ML foundational models, or generative AI applications, moves a decision context across the gating threshold at which a desktop review is inadequate and at which a more comprehensive, data context-specific, risk assessment should be conducted.

Certain use cases may be considered sufficiently high risk of harms that they should be prohibited. In some of those uses cases, the risk factors that cause the use case to be so high risk of harms that cannot reasonably be expected to be mitigated relates to the lack of 'inside the box' explainability that is a characteristic of some, but not all, ML, or the error limitations of current generation generative AI. For these use cases it may be useful to distinguish ML foundational models, and generative AI applications, from deployment and use of hardcoded advanced data analytics (algorithmic) systems.

In this regard, we commend the UK approach. Instead of attempting to define "AI", and then imposing generic requirements to AI as defined, we suggest that the Australian government should focusing on setting economy-wide expectations for the development and use of AI and empower existing regulators to issue guidance and regulate the use of AI within their remit. The UK Government's 2023 White Paper[19] scopes proposed activities by reference to two "characteristics of AI", being adaptivity and autonomy:

> *The 'adaptivity' of AI can make it difficult to explain the intent* [object] *or logic of the system's outcomes:*
>
> *AI systems are 'trained' – once or continually – and operate by inferring patterns and connections in data which are often not easily discernible to humans.*
>
> *Through such training, AI systems often develop the ability to perform new forms of inference not directly envisioned by their human programmers.*
>
> *The 'autonomy' of AI can make it difficult to assign responsibility for outcomes:*
>
> *Some AI systems can make decisions without the express intent or ongoing control of a human.*

This broad characterisation is intended to future-proof proposed regulatory frameworks for AI against new technologies and to confer sufficient discretions upon individual regulators, who might then issue guidance to regulated organisations setting out their expectations about the use of AI within the regulator's remit.

The UK Government stated its expectation that regulators will:

- In the next 6 months (viz., by end Q3 2023) assess and apply the principles to AI use cases falling within their remit, prioritising principles according to the needs of their sector.

[19] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

- In the next 6-12 months (by end Q1 2024) issue new guidance or update existing guidance to businesses on how the principles interact with existing legislation and to illustrate what compliance should look like.
- Support businesses operating within the remits of multiple regulators by collaborating and producing clear and consistent guidance.

This approach allows flexibility, but also creates risk of lack of comprehensiveness of scope and coverage and inconsistency between regulators. AI-affected activities within many Commonwealth, State and Territory government agencies might fall outside the scope of existing regulators' remits.

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?*

Many responses to this Discussion Paper will express views as to applications and use cases for AI that might be listed as 'unacceptable risk' of harms to affected humans or the environment and therefore justify regulatory prohibition ('blacklisting').

This paper does not reject blacklisting of carefully defined uses of technologies (as distinct from AI itself), such as use of identification of individuals through use of facial recognition in public places or semi-public places beyond a necessary, proportionate and properly safeguarded and controlled use by law enforcement agencies: see further our response to question 10 of the Discussion Paper.

However, the focus of this paper is to assist policymaking as to scoping:

- before the event (a priori) prescriptions, i.e., AI may only be used for a specified AI-affected activity if a regulated entity first complies with specified preconditions such as conduct of an AI impact assessment, and
- before-the-event requirements, i.e., for transparency, nomination of a responsible officer, due consideration by or for the responsible senior officer of a regulated entity of relevant possible significant risk of harm factors even where the requirement for consideration does not extend to a formal structured process for risk of harms evaluation such as conduct of an AI impact assessment.

See further our response below to question 3.

Many AI applications and services commercially offered in Australia originate from outside Australia.

Australian regulation can reasonably address transparency and disclosures offered by offshore providers to organisations using these applications and services in Australia.

Data sheets, system cards, and model cards made available by some providers of commercially offered AI applications and services are today of different levels of

transparency and frankness in disclosure of known limitations, comprehensiveness and quality.[20] The quality of information from suppliers of AI systems varies widely. Some providers do not make relevant disclosures at all. There is not yet standardisation in form, level of disclosure or even expectations as to disclosure, by suppliers of AI systems and services.[21]

Many organisations deploying AI systems or AI services supplied by a third-party supplier into the organisations decision chain have limited information from the supplier to evaluate inherent limitations of the AI system or service. The organisation's project team is therefore constrained in understanding how to design their deployment around these inherent limitations. This comment is not intended to be critical of all suppliers of AI systems or services. Rather, the problem reflects still nascent standardisation as to supplier disclosures for AI systems and AI services. This is partly a transitional issue that is already being addressed in the competitive AI market. However, regulated requirements for transparency and disclosure may assist.

Transparency and disclosure requirements might be applied economy-wide to commercial offering of an AI product or service that is intended for use by a customer, whether a business customer or other organisation, or a consumer or other end user.

These requirements could take the form of new provisions in Australian Consumer Law and analogous requirements in sector specific laws, such as the Corporations Act, that ensure economy wide coverage of such provisions.

Disclosure might be required of reasonably suspected errors, bias or other limitations that should reasonably be anticipated to materially affect reliability of outputs of the AI product or service when used by or for a customer for such tasks or other uses as ought reasonably be anticipated by the commercial provider of an AI product or service.

Disclosures should be made promptly and with sufficient prominence.

Disclosures should be capable of being understood by a non-expert reader of these disclosures.

Disclosures might include warnings or cautions in data sheets, system cards, model cards or other sufficient prominent instructional or explanatory material or as to tasks of other uses for which the AI product or service is not intended by the commercial provider to be fit for purpose.

Disclosures might include checklists, tools or other assurance aids and recommendations as to evaluation and assessment by the customer to evaluate safety and reliability of the AI product or service for a particular tasks or other uses, or categories of tasks or other uses.

---

[20] See Meta AI, System Cards, a new resource for understanding how AI systems work, February 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work; Responsible Use Guide, Research Paper, Model Card and other materials available at the Llama 2 launch site, Llama 2; GPT-4 System Card OpenAI March23,2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf; Hugging Face, Model Cards, https://huggingface.co/blog/model-cards; Furkan Gursoy and Ioannis Kakadiaris, System Cards for AI-Based Automated Decision Systems https://arxiv.org/pdf/2203.04754

[21] See further Hugging Face, The Landscape of ML Documentation Tools, https://huggingface.co/docs/hub/model-card-landscape-analysis

However, warnings and cautions, and recommendations as to evaluation and assessment by the customer, must not be unfair to the customer, having regard to the nature of the AI product or service and the relative knowledge, skills and capabilities of the provider and the customer to anticipate, assess and mitigate likely risks of harms to humans or the environment caused by uses of AI affected decisions.

AI products and services create issues as to gaps in coverage of Australian Consumer Law, and in particular operation of the definitions of "consumer", "goods" and "services" as used in the ACL, and the consumer guarantees under the ACL.

Many of these issues of gaps in coverage, and similar issues as to appropriate allocation as between providers and customers of responsibility and accountability to anticipate, assess and mitigate risks of harms, also arise in relation to deployment of internet of things consumer ("smart') devices and IoT device enabled internet connected ('smart') services. While noting:

- that AI adds further categories of risks of harms to those arising from smart devices and services, and
- the novel issues in allocating responsibility and accountability for multitask generative AI, as discussed in our case study in section 2 above),

we commend consideration of applicability to AI systems of the recommendations and analysis of Professor David Lindsay, Genevieve Wilkinson and Evana Wright (each of UTS Sydney) as to how the Australian Consumer Law should be adapted to facilitate safe and responsible adoption in Australia of smart devices and smart services.[22]

The draft EU AI Act proposes further notice requirements that would apply both to:

- providers placing AI systems on the EU market or putting AI systems into service in the EU, and
- users of AI systems.

Providers and users of AI systems would have new transparency obligations vis-à-vis affected individuals, subject to limited exceptions. Relevant provisions of the draft EU AI Act remain under active negotiation, but their current form would require providers to ensure individuals are informed that they are interacting with an AI system. If an AI system generated 'deep fakes', the user of the AI system would be required to disclose this.[23] Users

[22] See in particular David Lindsay, Genevieve Wilkinson and Evana Wright, Regulation of Internet of Things Devices to Protect Customers, June 2022, https://opus.lib.uts.edu.au/bitstream/10453/158337/2/ACCAN%20IoT%20Project%20Final%20Report_20622_Clean_Accessible%20%281%29.pdf; David Lindsay, Genevieve Wilkinson and Evana Wright, 'Who is responsible for an internet of unsafe things under the Australian Consumer Law?', Australian Journal of Competition and Consumer Law, Volume 31 Issue 1, March 2023; Kayleen Manwaring, 'Will Emerging Technologies Outpace Consumer Protection Law? The Case of Digital Consumer Manipulation', Competition and Consumer Law Journal, 2018, Vol 26, Issue 2, pp 141-181

[23] The May 2023 draft of the AI Act uses a tiered approach to proposed review and transparency obligations, using three key terms: "general purpose AI", "foundation models" and "generative AI". General purpose AI is broader than the latter two, but not all foundation models and generative AI systems fall within the term general purpose AI. Generative AI appears to be used as a subset of foundation model, although a better description is that it is the application (i.e., ChatGPT) built on top of the foundation model (i.e., GPT-3.5). The obligations on generative AI appear intended to ensure that users always know when content they see or hear is AI-generated. The rules

of an emotion recognition system or a biometric categorisation system would be required to inform affected individuals. Providers and users of generative AI would be subject to additional transparency requirements, including disclosing that the content was generated by AI and preventing the generative AI from generating illegal content. Providers of generative AI would be required to publish a description of copyright material used for training the foundational model.

We commend consideration of further notice requirements similar to these EU proposals, while noting that:

- the categories of AI systems and relevant uses that should be subject to such requirements should be the subject of further consideration,
- notice fatigue of users is already a well-recognised problem of data privacy regulation.[24] Users should not be unreasonably burdened with further notices with any expectation that users should read and engage with such notices in order to protect themselves from harms that were reasonably foreseeable to the drafter of the notice. Notice to users should not be allowed to become a substitute to AI system providers and users exercising reasonable diligence to protect affected individuals from AI harms, by those providers and users taking risk mitigation measures that are reasonably within their capabilities.

We also note heightened risk of harms through many uncontrolled and opaque uses of AI enabled facial recognition to enable identification of individuals. We commend the proposal for a AI facial recognition model statute as made in the UTS Human Technology Institute's report of September 2022.[25]

We recommend caution before Australian follows the proposed approach in the EU AI Act of economy-wide, mandated uniform requirements in relation to the broadly described categories of applications defined as "high-risk AI systems". The draft EU AI ACT requires providers and users to prepare a wide range of new documentation and internal processes in relation to high-risk AI systems.[26] For providers, this will include preparing:

- systems for risk management, quality management, and post-market monitoring,
- processes to ensure data quality and logging,
- tools for human oversight,

for foundation models are wider in scope. The current draft requires providers of foundation models to "demonstrate through appropriate design, testing and analysis that the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development with appropriate methods such as with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development."

[24] See further Peter Leonard (Data Synergies), 'Notice, Consent and Accountability: addressing the balance between privacy self-management and organisational accountability: A research paper for the Office of the Australian Information Commissioner', June 2020, available at https://www.oaic.gov.au/__data/assets/pdf_file/0003/2010/notice-and-consent-paper-for-oaic.pdf.pdf

[25] UTS Human Technology Institute, Facial recognition technology: Towards a model law Facial recognition technology: Towards a model law, September 2022, https://www.uts.edu.au/human-technology-institute/projects/facial-recognition-technology-towards-model-law

[26] High-Risk AI within scope of the current draft includes products covered by EU health and safety laws that require third party conformity assessment (e.g., medical devices, radio equipment, cars, toys, aviation); and AI systems used for one of the following purposes: remote biometric identification; regulation of road traffic, water, gas, heating, and electricity systems; determining access to education or assessing students; recruitment, termination, and other job-related decisions; determining eligibility for benefits; evaluating creditworthiness; despatching emergency first response services; and certain specific purposes in the area of law enforcement, justice, and immigration.

- measures to ensure accuracy, robustness, and cybersecurity, and
- technical documentation and instructions for use.

Providers must also carry out a conformity (self-)assessment, affix the CE marking, and report incidents to the relevant regulator.

The draft Canadian AI and Data Act proposes measures that to be applied at each stage of the lifecycle of a high-impact AI system, applying criteria to be specified in regulations under the Act. These measures would not apply to an organisation distributing or publishing open-source software or models, as these are not considered to be a complete AI system (as compared to an open-access, fully functioning high-impact AI system). The design and development requirements, which would be detailed in regulations under the Act, would need to be met before a high-impact system is made available on the market for use. If the provider did not fulfill these obligations, it would be exposed to penalties, including civil penalties imposed by a regulator. If the provider made the system available for use knowing that it was likely to cause serious harm, they could be prosecuted for a criminal offence. If a user organisation puts the provider's system into operation for their own commercial purposes and manages the operations, the user organisation would need to comply with the requirements for managing operations (e.g., ensuring that this use is appropriate given the risks and limitations documented by provider A, monitoring the system, publishing a description of the system). If the system as operated by the user causes harm, the user would be only liable if they did not meet the obligations related to managing operations. If in operating the system the user showed reckless disregard for the safety of other persons, they could be prosecuted for a criminal offence.

Regulated activity	Examples of measures to assess and mitigate risk
System design - includes determining AI system objectives and data needs, methodologies, or models based on those objectives.	• Performing an initial assessment of potential risks associated with the use of an AI system in the context and deciding whether the use of AI is appropriate • Assessing and addressing potential biases introduced by the dataset selection • Assessing the level of interpretability needed and making design decisions accordingly
System development - includes processing datasets, training systems using the datasets, modifying parameters of the system, developing and modifying methodologies, or models used in the system, or testing the system.	• Documenting datasets and models used • Performing evaluation and validation, including retraining as needed • Building in mechanisms for human oversight and monitoring • Documenting appropriate use(s) and limitations

Making a system available for use – deployment of a fully functional system, whether by the person who developed it, through a commercial transaction, through an application programming interface (API), or by making the working system publicly available.	• Keeping documentation regarding how the requirements for design and development have been met • Providing appropriate documentation to users regarding datasets used, limitations, and appropriate uses • Performing a risk assessment regarding the way the system has been made available
Managing the operations of a system – supervision of the system while in use, including beginning or ceasing its operation, monitoring and controlling access to its output while it is in operation, altering parameters pertaining to its operation in context.	• Logging and monitoring the output of the system as appropriate in the context • Ensuring adequate monitoring and human oversight • Intervening as needed based on operational parameters

Although we support structured risk management for applications of AI that create significant risks of harms to humans or the environment from AI-affected decisions, we also express some concerns:

- We are sceptical of the value of prescribing how risk of AI harms assessments are undertaken, by whom or their form. See further our discission in section 1 as to why we instead advocate policies and program focussed enforced self-regulation model which includes prescriptions (as described on pages 6 and 7 of this paper) to ensure that reasonable precautions are implemented by each organisation.
- Creation of a bounded list of high-risk AI products or uses that are required to be subject to individual structured prescriptive requirements for documentation and processes may lead to organisations considering that other applications do not require prudent risk assessment and management. Risk of significant harms is highly contextual. A bounded list may 'lull to sleep' organisations as to significant risks in other contexts.
- As discussed elsewhere in this submission, capabilities of organisations to conduct AI risk assessment are likely to remain highly variable over the next three to five years. Risks of harms from uses within one industry vertical (i.e., retail energy, or financial services) sector may be more better addressed through providing incentives for a relevant industry association to develop an industry specific framework or industry code, or by enforced self-regulation of and by organisations, or prescriptive intervention by an industry-specific regulator, while, for example, uses of AI by small businesses may be more effectively addressed through provision of checklists and

practical assessment by bodies such as State and Territory Small Business Commissioners. Some applications, such as uses of AI enabled facial recognition to enable identification of individuals, may be better addressed by use case specific prohibitions.

Given continuing unpredictability as to many of the possible uses of AI, regulated prescriptions and requirements should be graduated, to address now reasonably foreseeable likely risks of significant harms, and then revisited over time. This periodic revisiting should apply a precautionary principle at that point of time to determine which then reasonably foreseeable risks of harms from particular categories of uses are of sufficient magnitude to then justify imposition of a prior (a priori) prohibition or other prescriptive requirements. Other then-identified risks of harms should be monitored and assessed on the next review cycle, to determine whether further a priori regulatory intervention is appropriate. An appropriate review cycle might be every three years, subject to exceptional circumstances.

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.***

Interventions by Australian government and its agencies to improve AI-affected decision provenance by organisations operating in Australia could be targeted to improve:

- **information and educational resources available to relevant humans**: improving competencies of people within organisations, and of other stakeholders (including industry associations, consumer organisations and civil society organisations), in AI-affected decision problem definition: that is, improving their competency to foresee a risk of harms to humans or environment that is reasonably attributable to a proposed use of AI,
- **processes for gating of AI and evaluation of AI**: frameworks (including standards), methodologies, tools, checklists to assist myriad organisations to better assure that they adopt a considered approach in determining whether, when and how they use AI, for which use cases,
- **governance of AI-affected decisions within organisations**: who considers what, in consultation with whom, before use of AI,
- **safeguards and associated assurance (reliability and verifiability) controls and feedback loops**, including ensuring post-implementation monitoring and reassessment based upon results and outcomes,
- assurance that there are clear, and clearly understood, **allocations of responsibility and accountability within organisations for AI-affected decisions**: that is, identification of main individual actors and sub-units who are accountable, focussed upon the outcome or result of a decision chain that includes an AI link in the decision chain, and not just the output from the AI link in the decision chain,
- context (industry sector, or use case) specific assessment of AI affected decisions, also having the most cost effective and impactful way to influence outcomes of AI-affected

decisions. For example, one particular industry sector may be more better addressed through providing incentives for a relevant industry association to develop an industry specific framework or industry code, when some applications such as uses of AI enabled facial recognition to enable identification of individuals may be better addressed by use case specific regulation.

Interventions should endeavour to strike a sensible balance between 'top-down, in the middle and bottom-up' initiatives:

- **top down**:
  - higher level guidance and guidelines as to safe and responsible deployments of AI,
  - enforced self-regulation of and by organisations (see the description of enforced self-regulation on pages 6 and 7 of this paper).
  - a requirement that each organisation (possibly excluding small businesses) designate a senior officer who is responsible for ensuring safe and responsible deployment of AI,
  - coordination between regulators to ensure commonality of approach and avoid duplication and conflict in requirements,
  - support for complementary (but not conflicting or duplicative) initiatives to assure safe and responsible deployments of AI in particular contexts: i.e., ASX guidance for listed corporations, APRA and ASIC standards and mandated requirements for regulated entities, international and Australian standards,
  - prohibitions (blacklists) in relation to use of AI in particular contexts,
  - prescriptions as to when and how a structured risk of harms assessment should be conducted by organisations where decisions by an organisation are to be assisted or influenced by uses of AI,
  - requirements to provide appropriate transparency and 'audit trail' for subsequent scrutiny by a regulator as to risk of harms assessments that have been conducted by an organisation.

- **In the middle**
  - support for further development and modification of enterprise risk and operational risk frameworks and methodologies to ensure that AI-affected decisions are fully addressed by these frameworks and methodologies. This initiative is particularly important and time critical. There is currently a gap between enterprise risk and operational risk programs, and technology project management frameworks and methodologies, in relation to evaluation of AI-affected decisions, as distinct from deployment of AI as a technology project,
  - support for development of best practice in data sheets, system cards, and model cards for AI applications and services, particularly focussed upon non-enterprise AI. The need in non-enterprise AI sector is greater than for enterprise AI applications, as enterprise AI applications are likely to be evaluated by resources and skilled project managers and demand-side competitive pressure upon

providers of enterprise AI is likely to lead to continuous improvement in transparency and disclosures,

- support for development of international and Australian standards for AI impact assessment, including sector-specific, application-specific and task-specific standards that focus upon practical steps for non-specialist personnel in evaluation and mitigation of risk of harms from AI-affected decisions,
- support for development of better understanding within organisations of the respective roles of C-suite executives, generalist managers, HR professionals, technology professionals, data science and AI/ML engineers, and lawyers, privacy professionals and prudential and regulator specialists, in assuring safe and responsible use of AI by organisations.

- **Bottom up**:
  - an information campaign and publication of explanatory materials about safe and responsible use of AI, particularly targeting small to medium businesses that are unlikely to have internal capabilities or resources for AI impact assessment and that may be considering using self-service generative AI applications for business dealings and interactions,
  - educational resources and self-assessment leading programs for designated senior officers who are to be allocated as responsible for ensuring safe and responsible deployment of AI, particularly focussed upon organisations that do not have internal project management capabilities and resources, or developed prudential/regulatory teams (being most medium to small businesses and not-for-profits)
  - support for complementary initiatives to assure safe and responsible deployments of AI in particular contexts: i.e., for upskilling/cross skilling of (1) IT professionals to develop their capabilities to project manage AI impact assessment (Australian Computer Society and like professional associations might lead), (2) human resource professionals to develop capabilities to inform and manage uses by personnel within organisations of AI applications, particularly uses of generative AI for task assistance in circumstances where the use of particular AI for a particular task is not being project managed into an organisation following AI impact assessment (Australian HR Institute (AHRI) and like professional associations might lead), (2) marketing professionals to develop capabilities to inform and manage uses by personnel within organisations of AI for consumer marketing (Association for Data-driven Marketing and Advertising (ADMA) and like professional associations might lead).

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.***

This paper commends the UK's lighter touch, coordinated but decentralised approach, as described in the UK Government's Policy paper *A pro-innovation approach to AI regulation* of 29 March 2023[27] and *AI Regulation Policy Paper* of 18 July 2022,[28] bit with the modifications that we describe in section 1 of this paper.

These Policy Papers note the risk of multiple regulators being asked to interpret and enforce a set of common principles is that regulated entities will be given inconsistent or contradictory guidance or guidance which leads to duplication of efforts. The UK Government proposes creation of central functions to support the multi-regulator, decentralised frameworks, including by:

- developing a central monitoring, evaluation and risk assessment framework,
- creating a central guidance to businesses looking to navigate the AI regulatory landscape in the United Kingdom,
- offering a multi-regulator AI sandbox, and
- supporting cross-border coordination with other countries.

While no announcement has been made, the UK Government's Office for Artificial Intelligence,[29] a unit within the UK Department for Science, Innovation and Technology, may take on some of these central functions. The UK Government currently addresses regulatory coordination through activities of the Digital Regulation Cooperation Forum (DRCF) and the Centre for Data Ethics and Innovation (CDEI).

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?

Australia should take its customary approach of charting its own course but also adapting good ideas from Australia's peer nations.

Our answers to questions 3 and 4 above explain why we commend the UK approach, but also propose a range of further initiatives to improve governance of AI.

Target areas

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

The Australian government is already considering how to responsibly address the recommendations of the Royal Commission into the Robodebt Scheme.

[27] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[28] https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

[29] https://www.gov.uk/government/organisations/office-for-artificial-intelligence

Many of those recommendations apply more generally than governance of algorithmic decision-making by government agencies and therefore should also be considered by federal, state and territory departments and public sector agencies in relation to prospective uses of AI, including LLMs and MfMs.

Australian government agencies will also need to consider and apply administrative law requirements in determining whether uses of AI are legal, as well as safe and responsible.

The above paragraphs identify restraints and considerations that apply to public sector use of AI technologies that are additional to those that should be applied by other organisations such as private sector entities. We are not aware of any reason why other assessment and management of Risks of AI harms by public sector agencies should be less than assessment and management by other Australian organisations. Given recent revelations as to extent of outsourcing and dependencies of public sector agencies upon external consultancies, we also note the public sector-specific challenge for agencies in building their internal competencies, and changing their organisational DNA, as required to responsibly address evolving capabilities and diverse uses of automated decision-making and AI within those agencies.

If any public sector agency considers that the agency should not be required to manage risk of harms in like manner to risk assessment and management by any large private sector organisation in Australia, it would be appropriate for the agency to:

- publicly state its case in relation to addressing a particular decision context, and
- be ready to demonstrate why separate treatment and lesser requirements are necessary and proportionate to address that particular decision context.

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

See our responses to question 3 and 6 above.

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.***

See our responses to question 2 above, and our proposed methodology for AI policy making as outlined in section 1 of this paper.

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.***

See our specific proposals in response to question 2 above, and our comments in response to question 6.

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?***

Many uses of AI create high risks of harms if uncontrolled and before mitigations are reliably and verifiably applied and residual risks then objectively assessed.

Blacklisting of AI applications may be seen as politically attractive, but discussion as to blacklists readily becomes politically contentious as to edge cases (*why was X included, but not Y?*), and the objectively assessed mitigation effects of risk measures and assurance controls may not be broadly understood.

In our view, blacklisting of particular categories of AI applications should only be considered in relation to uses of AI that are of such extreme risks of harms to humans or the environment as to be unacceptable to Australian society regardless of whether safeguards and assurance controls are reliably and verifiably applied.

Inclusions in the proposed blacklist for the draft EU AI Act remain controversial. The list was significantly extended by Members of the European Parliament in May 2023 MEP committee stages. It is unclear whether the extended list will be the final list.

The list as then adapted includes:

- systems that deploy subliminal or purposefully manipulative techniques or exploit people's vulnerabilities,
- systems used for social scoring (classifying people based on their social behaviour, socio-economic status, personal characteristics),
- use of real-time remote biometric identification systems in publicly accessible spaces,
- post-time remote biometric identification systems, with the only exception of law enforcement for the prosecution of serious crimes and then only after judicial authorisation,
- biometric categorisation systems using sensitive characteristics (e.g., gender, race, ethnicity, citizenship status, religion, political orientation),
- predictive policing systems based on profiling, location or past criminal behaviour,
- emotion recognition systems in law enforcement, border management, workplaces or educational institutions,

- indiscriminate scraping of biometric data from social media or CCTV footage to create facial recognition databases and by so doing violating human rights and right to privacy.

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

See our specific proposals in response to question 3 above.

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?*

No specific comments at this time.

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

No specific comments at this time.

Risk-based approaches

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

Yes. In summary:

- a risk-based approach is the best way to address risks of AI harms,
- there is no best one-size-fits-all approach to risk assessment and management,
- we advocate a policies and programs focussed enforced self-regulation model as described on pages 6 and 7 of this paper, and not a prescriptive specification of how AI risk assessments should be undertaken, or the form that they should take.

Each AI-enabled, or AI-assisted, decision requires consideration of decision provenance: the interaction of people, processes and technologies that effect, or affect, that decision.

Most organisations operating in Australia that are implementing AI within the organisation will not be developers and suppliers of AI solutions. Typically, an organisation operating in Australia will be tailoring a third-party AI application or service and using it:

- more commonly, to inform or otherwise aid people within an organisation to perform a decision -making task, or
- much less commonly, to enable a fully automated (self-actuating) outcome.

Senior executives and managers within organisations may not see a business process (decision chain) within the organisation as an 'application of AI', or even as significantly affected by AI.

Most organisations, and particularly those that are not large businesses, do not have internal competencies to reliably translate higher level 'ethical AI' principles into practical business decisions.

Most organisations are not large businesses that have experience, settled procedures, internal capabilities and resources to reliably evaluate a third-party AI application or service for fitness for purpose for reliance in a particular business context, and assess the quality of data inputs used by the AI provider to train that AI.

Frameworks, methodologies and tools for assessment of AI are the essential translational layer between:

- higher level 'ethical AI' principles, and
- practical business decisions that effect safe and responsible uses of AI.

This translational layer need not be complex or highly structured, depending upon capabilities and willingness of organisations:

- to recognise that there are relevant risks,
- to understand the nature of risk of harms, and
- to allocate responsibility to address and mitigate those risks, and thereby demonstrate accountability and trustworthiness.

One danger for organisations to navigate when implementing AI is that because AI is a novel technology, the temptation is to focus too much on the technology, and not the relevant humans using that technology in a particular context to make a particular decision.

If the context can be appropriately circumscribed and evaluated for that context, and the skills of relevant humans reliably pre-assessed, risks of uses of AI can be substantially mitigated.

Each neighbourhood florist must water cut flowers and plants and each day changes where watered pots sit within a store. Each florist store manager manages a known slip and fall hazard zone. Their management of that zone is a combination of inherent human competency, procedures (i.e., promptly mop up split water), and simple checklists. Management of that zone is usually effective to mitigate those risks, because the risks are familiar and well understood by non-specialist risk managers.

Each airline pilot manages a more complex work environment, aided by rigorous context (cockpit specific) training, structured checklists, and knowledge as to critical dependencies upon other people conducting other checks and exercising safety-related responsibilities. The airline pilot requires checklists to aid learned knowledge and skills. The airline pilot does not require a complex structured risk management framework or methodology.

Design and deployment of:

- advanced data analytics (data and algorithms) to inform or assisted decision-making by humans,
- task-specific AI/ML applications, and
- multi-purpose generative AI,

should cause organisations to consider change control, and address management of deployment and use of a new technology. Some of these organisations will have capabilities and experience to apply enterprise or technology risk frameworks, methodologies and tools. This experience may have been applied by the organisation to deployment of earlier technologies, such as cloud platform enabled integration of diverse data sets, use of social media, COVID accelerated take-up of video-conferencing, and COVID and post-COVID remote access by personnel to an organisation's information systems and trade secret and sensitive data. Typically, an organisation addressing a new technology should promptly consider:

- whether new gating criteria are required – whether the new technology will be taken up by the organisation, and if so, who should be permitted to do what, using the new technology,
- whether prohibitions (no-go zones) need to be created, either to ensure that the technology is only used by the organisation's personnel for lawful purposes, or to ensure that there are not significant harms to the organisation or to others (customers, suppliers, citizens and stakeholders),
- design and implementation of appropriate guardrails that ensure that personnel that are authorised to use the new technology to perform particular tasks operate only within those guardrails.

In other words, an organisation's governance framework needs to be fit for purpose to enable an organisation to promptly determine, in response to the new technology, how to adjust each of the three interrelated elements of people, processes and technology, in order to ensure that having regard to the interaction of the three elements, the organisation's use of the new technology is lawful and does not cause significant harms to the organisation or to others (customers, suppliers, citizens and stakeholders).

Within an organisation, introduction of a new technology will often require significant change management led by the human resources team, because the human element of the socio-technical decision chain within the organisation is so critical in ensuring safe and responsible use.

Assuring implementation of safe and responsible AI for Australian citizens requires organisations:

- to understand what is 'safe and responsible AI',
- to internalise and address risks of harms to others, to the extent that those harms are reasonably attributable to the organisation's provision, deployment or use of AI.

Organisations are as diverse as the tasks, processes and decisions for which they will be using AI. Consider the diversity of organisations now implementing AI:

- large, medium, small;
- businesses, not-for-profits and government agencies;
- data mature, technology native or other;
- subject to sector specific or application specific regulation, or qualifying for small business or other exemptions;
- subject to other jurisdictions' data and AI regulations, familiar with data regulation, or new to data regulation;
- risk averse, high risk appetite but rational, or irresponsible and fly-by-night.

Within these diverse organisations, significant change management will be required to implement safe and responsible AI. For example, many of the following will often be required:

- decision flows and other internal business processes, including ways of personnel interacting with other staff, will need to be respecified,
- ways of human staff interacting with customers, suppliers, service providers and other third parties, will need to be respecified,
- staff will need to be trained or reskilled or reallocated,
- an organisation's automated processes will need to be changed,
- adjustable features and functionality of the new technology will need to be specified and set as appropriate to mitigate risk of harms from uses by those personnel that are authorised to use that technology for approved tasks,
- new technical, operational (including policies and processes) and legal safeguards specified and implemented to give effect to guardrails and prohibitions,
- assurance controls specified and implemented to ensure that safeguards are reliably and verifiably implemented and accordingly that the governance oversight is effective as planned.

In larger organisations with more mature risk management frameworks, new technologies are typically passed through a project initiation, evaluation and project management process. Often these processes are managed by technology professionals that are skilled in applying either an enterprise-wide risk framework, or a technology focussed framework, and associated methodologies and tools. Frameworks, methodologies and tools typically have been developed for use in organisations where structured assessment and management of:

- enterprise risk,
- operational risk, and
- technology risk,

is a familiar competence and capability. This is not the case for the large majority of Australian businesses and social enterprises that are now implementing AI applications and

services. This fact creates a challenge for Australian policymakers. A majority (by number) of Australian organisations are unlikely to develop capabilities to implement complex structured frameworks, methodologies and tools for management enterprise risks, operational risks, or technology risks, within the next three to five years. Many of those organisations will implement and use AI applications and services policy within the next three to five years.

Further, for most organisations addressing risks of AI harms will be quite different from managing technology disruptions in the past. Assurance by organisations that their uses of AI are safe and responsible requires a team approach bringing together different disciplines.

Many possible AI harms arise because outputs from AI or hardcoded algorithmic data analysis are statistically based, so outputs are not universally reliable for use in the broad range of contexts in which decisions based upon those outputs may be made. Inference based outputs may be statistically reliable for strategic or product positioning decisions by an organisation, where 'statistical errors' are not sufficiently material in scale or effect to undermine reliability of the inference as an insight for a business decision. The same inferences may be wrong and unreliable when used for more granular or targeted decisions, particularly where data depth is limited or data quality not fully assured (i.e., 'at the tails of the Bell curve'). Decision context should affect assessment of whether earlier links in the decision provenance chain[30] are sufficiently robust to be reliable inputs for a particular decision. An informed understanding of the quality and reliability of each link in the chain of data inputs, people, processes, and technologies used to create AI/algorithmically enabled output, and then apply that output in a way that affects, or makes, a particular decision (an outcome), is crucial to ensuring that an AI/algorithmically assisted decision is appropriately reliable for the reliance that is placed upon it. Evaluation of the AI/algorithmic links within this chain of decision provenance is important, but only part of an evaluation of the quality and reliability of decision provenance that needs to be made by an organisation responsible for an AI/algorithmically assisted decision.

'Statistical errors' may be addressed by a 'human in the loop': organisational reliance upon appropriately skilled humans to review the outputs and detect and override insufficiently robust results. An assessment needs to be made as to the level of statistical error that can be reasonably accommodated, having regard to the harms that may flow from reliance upon erroneous outputs. Governance and assurance controls for advanced data analytics and AI therefore needs to be different from governance and risk assurance of most other technology or engineering projects. In addition, information technology risk frameworks and

[30] Decision provenance entails using provenance methods (recording information about the nature and flow of data and the contexts in which it is processed) to analyse data and analytics affected decision chains or pipelines (chains of inputs to, the nature of, and the flow-on effects from the decisions and actions taken, at design and run-time through systems). By making transparent the connections and data flows driving systems, the context in which they are operating, their effects, and the entities involved (of which there may be a number), it then become possible to allocate responsibility, and accountability, for purposes of compliance, oversight, and increasing user agency, which (depending upon incentives) may contribute to better system design, operational (run-time) management, and risk mitigation. See further Jatinder Singh, Jennifer Cobbe, and Chris Norval, Decision Provenance: Harnessing data flow for accountable systems, 2019, https://doi.org/10.1109/ACCESS.2018.2887201; also Ian Brown, Allocating accountability in AI supply chains: a UK-centred regulatory perspective, Ada Lovelace Institute, June 2023, https://www.adalovelaceinstitute.org/resource/ai-supply-chains/; Jessica Newman, A Taxonomy of Trustworthiness for Artificial Intelligence, January 2023, https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/

methodologies have been developed over three decades. AI project frameworks and methodologies are nascent, less developed and standardised, and therefore less understood than standardised information technology project frameworks. In the last two years we have also seen emergent, albeit still work-in-progress, best practice exemplars for AI project assurance frameworks, such as the NSW AI Assurance Framework[31], national and international AI assurance standards[32] and work by entities such as the Turing Institute, Ada Lovelace Foundation, World Economic Forum, Gradient Institute and CSIRO.[33] These exemplar AI/algorithmic assessment frameworks and methodologies are generally designed for deployment within a system of a project initiation and approval, where appropriately skilled and experienced individuals evaluate the suitability, safety and legality of a proposed implementation of AI. This 'gating and penning' process reduces risk of AI being inappropriately deployed. Effective 'gating and penning' requires an organisation to ensure that:

- there is a gate,
- the gate is manned by humans with appropriate skills,
- candidates for assessment are identified early and required to pass through the gate,
- a suitable assessment framework reliably and rigorously applied for each proposed use of AI that is 'within the pen',
- evaluation within the pen ensures that AI outputs are fit for purpose, having regard to the likely reliance that will be placed upon them, such that decisions enabled or affected by AI outputs (i.e., outcomes) reasonably reflect the quality and other provenance of the AI outputs,
- 'out of the pen', real world outcomes from uses of AI outputs are assessed for fit to expectations,
- proper change evaluation controls are applied before any subsequent changes in data inputs, data processes, AI/ML/algorithmic functionality or uses of outputs are made,
- adverse consequences suffered by others from uses of the AI by an organisation are not treated as externalities and ignored by the organisation, or left for some party to address.

However, emergent AI project assurance frameworks have a number of limitations:

- They are new and therefore unfamiliar.
- They are quite complex, and typically require multi-disciplinary input in order to be done well. Typically, an experienced project manager is required to manage the process. The project manager will require the skills, experience and conferred authority to obtain, manage and evaluate input from a diverse range of stakeholders,

[31] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework

[32] See further U.S. National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AIRMF1.0); Turing Institute, AI Standards Hub, https://aistandardshub.org/; Standards Australia, An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard, https://www.standards.org.au/documents/r-1515-an-artificial-intelligence-standards-roadmap-soft.

[33] E.g., Gradient Institute and CSIRO, Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources, June 2023, https://www.csiro.au/-/media/D61/NAIC/Gradient-Report/23-00122_DATA61_REPORT_NAIC-ResponsibleAITools_WEB_230620.pdf; Mina Narayanan Christian Schoeberl, A Matrix for Selecting Responsible AI Frameworks, June 2023, https://cset.georgetown.edu/wp-content/uploads/CSET-A-Matrix-for-Selecting-Responsible-AI-Frameworks.pdf.

typically including data scientists, algorithmic/AI engineers, operational process specialists, human resource personnel, prudential and regulatory risk advisors, privacy professionals and legal counsel.

- Oversight governance personnel need to be also appropriately familiar with AI risks and harms assessment. In understanding how use of AI affects decisions made by or on behalf of an organisation, each decision context affected by the use of AI needs to be considered, having regard to the decision chain: the links of people, processes (including rules and policies) and technologies that make up a chain that leads to an AI affected decision. Analysis of risk of harms of AI requites a detailed, contextual understanding of often diverse business processes within an organisation. In many organisations, many business processes are inherent (skills based) and not documented.
- When AI affected decisions are irresponsible or unsafe, harms can be caused at scale and velocity. Because AI tools are now available as self-serve applications available to all organisations, even smaller organisations can quickly cause harms at scale.

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

See section 1 of this paper, and our response to question 14.

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?*

Yes. See section 1 of this paper, and our response to question 14.

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

The elements summarised in Attachment C are relevant but incomplete, partly because the focus in Attachment C is upon assessment of the AI system itself, rather than the decision context that is being affected by the AI.

We disagree with the apparent presumption in Attachment C that all impact assessments should be published, or that "peer review" of impact assessments "by external experts" would significantly reduce risks of AI harms. We discuss the appropriate scope for transparency and disclosures, and incentives for transparency and disclosures, in section 1 and in our response to question 2.

In section 1 of this paper we further discuss how to improve incentives for organisations to adopt a risk-based approach that is appropriate to the decision

context, the capabilities of that organisation and the level of risks of AI harms if not appropriately risk managed and mitigated.

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?*

See section 1 of this paper, and our response to question 14.

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

There are at least three relevant levels at which risks from general purpose AI systems need to be assessed:

1. The underlying foundation models and the data inputs used to fuel those models. The developers of these foundational models need to have appropriate incentives to make fair disclosures as to limitations of those models, so providers of generative AI applications built upon those models may consider the reliability of the foundational model, assess the reliability and safety of the application, and in turn make appropriate disclosures and ensure that their generative AI application offering is safe and complies with law.
2. Provision of the generative AI application built upon a foundational model.
3. Use of a generative AI application for a particular user-determined task.

We have already made specific suggestions as to transparency and disclosure requirements to enable risk assessment at levels 1. and 2: see in particular our response to question 2 above.

As to 3., we note that there are a number of reasons why generative AI applications are 'fast fashion' for many prospective users:

- generative AI applications are readily available at low cost,
- they can be accessed by anyone within many organisations without first needing the organisation to buy, the IT department to deploy, or the boss or the HR team to approve. This 'ungated' use of AI within organisations is sometimes referred to as use of 'shadow AI' or 'stealth AI',
- they are easy to play with,
- their outputs are compellingly useable, even when unreliable, and
- they readily demonstrates their usefulness as an aid to performance of many tasks, even when unreliable as used as an aid for those tasks.

Many would-be users will experiment, including in their own time and without use of an organisation's IT resources. Warnings and cautions issued by organisations should be expected to be ignored by some would be users within those organisations. Bans or controls should be expected to be circumvented by many users.

Provision and use of generative AI applications therefore amplify some of the categories of AI harms as discussed elsewhere in this paper. Take-up of generative AI applications will likely expand the range of risks of harms, notwithstanding providers of generative AI applications also improving reliability of these services and disclosures as to limitations in reliability of these services. Take-up will also be fuelled by developing generations of general purpose AI that enable 'air-guarding' of prompting data from the data corpus of a provider of the AI model or functionality, thereby enabling deployments in many data contexts where under current laws regulated data sets could not be used either to train the underlying foundational model, or to prompt the generative AI.

Many Australian organisations that do not have capabilities to implement complex structured risk management frameworks or methodologies will be implementing generative AI to assist non-technical humans to perform myriad tasks. Larger organisations that have structured functional teams will usually only introduce a new technology after risk assessment and with an associated change management program led by the human resources team. That program is likely to include changes to policies and process documentation; changes in oversight and internal review processes; re-designation of roles and responsibilities of staff members; new training; new instructional materials; new warnings and 'no-go zones', and so on. Because addressing the human element in socio-technical decision chains within the organisation is critical in ensuring safe and responsible use of AI, the risks of unintended and unanticipated AI harms are much greater for the majority of organisations that do not have a project evaluation program, a change management program, or a human resources team.

This creates an important policy challenge in addressing widespread use of generative AI, given material prevalence of unanticipated errors within AI outputs that are compellingly presented as ready for use. Generative AI applications therefore create novel governance risks of unknowing amplification of disinformation and misinformation, as well as opportunities for deliberate use for disinformation and misinformation.

Consider one well-publicised recent example. Steven Schwartz, a New York attorney with over 30 years of post-admission experience, represented Roberto Mata in an action against Avianca Airlines for injuries sustained from a serving cart while on the airline in 2019. At least six of the submitted cases by Schwartz as in a brief to the court of the Southern District of New York court "appear to be bogus judicial decisions with bogus quotes and bogus internal citations," said Judge Kevin Castel in a May 2023 order.[34] In a June 222 ruling, Judge Castel ordered Schwartz and his law firm to pay $5,000 for submitting without checks a brief

[34] The cases, generated by ChatGPT, included Varghese v. China South Airlines, Martinez v. Delta Airlines, Shaboon v. EgyptAir, Petersen v. Iran Air, Miller v. United Airlines, and Estate of Durden v. KLM Royal Dutch Airlines. Neither the judge or nor the defence lawyers could find reports of these judgements: they did not exist, although generated by ChatGPT.

with fake cases and then standing by the research.[35] As one response to this case, a federal judge in Texas is now requiring lawyers in cases before him to certify that they did not use artificial intelligence to draft their filings without a human checking their accuracy.[36]

One view might be that this example illustrates a transitional problem, and not a lacuna in regulation. Regardless of any view as to the small penalty imposed, no sensible lawyer would wish to suffer the reputational damage flowing from global reports as to the lawyer's failure to understand and mitigate the misinformation risks of reliance upon an LLM. Smart people often do dumb things when they don't know that they are doing a dumb thing. Having read a media report as to Mr Schwartz's failure to be the responsible human in the loop, other lawyers are unlikely to replicate this error, regardless of whether a legislature creates a relevant prohibition or by a professional standards body makes a code about use of LLMs, or a ruling of professional misconduct.

However, the difficulty is that analogous inappropriate reliance upon erroneous outputs from generative AI may arise in many of the myriad tasks (completely unrelated to the conduct of litigation) for which generative AI is now being used by individuals without those individuals knowing to exercise:

- appropriate caution as to the possibility of such errors, and
- responsibility to take appropriate steps to mitigate such risks.

A focus upon decision provenance as to use of generative AI as a task assistant - the interaction of people, processes and technologies in the decision chain of an decision affected by outputs of generative AI – may lead to counterintuitive conclusions. Many Australians, if asked today, would express the view that possibly 'hallucinating' ChatGPT should have no role in expressing a clinical opinion to another clinician as to an individual's health status or prognosis. As the clinical use case discussed in section 2 of this paper illustrates, that view might be wrong, if (and only if) the people and process settings are appropriately assessed managed. As that example illustrates, a key consideration in that decision context is the quality of the professional judgement brought into the loop of AI-affected clinical decision-making. In a highly controlled hospital environment, where clinicians following a well-understood and structured process, generative AI might be used safely and responsibly, if (and only if) used by clinicians appropriately skilled as to possible errors in the AI outputs, and with appropriate steps by the organisation controlling the clinical environment to mitigate risk of such errors.

[35] Debra Cassens Weiss, "Lawyers who 'doubled down' and defended ChatGPT's fake cases must pay $5K, judge says", ABA Journal, 26 June 2023, https://www.abajournal.com/web/article/lawyers-who-doubled-down-and-defended-chatgpts-fake-cases-must-pay-5k-judge-says

[36] In late May 2023 Judge Brantley Starr of the U.S. District Court for the Northern District of Texas updated his judge-specific requirements to include a section titled "Mandatory Certification Regarding Generative Artificial Intelligence." Specifically, Starr orders all attorneys appearing before the court to file a certificate attesting that either: (1) no portion of any filing will be drafted by generative artificial intelligence; or (2) that any language drafted by generative artificial intelligence will be checked for accuracy by a human being.

In summary, a risk management approach, coupled with an enforced self-regulation model, should be applied to general purpose AI services, designed to address the challenges associated with:

- lack of organisational control over how it is likely to be introduced into and used in many organisations for a myriad of tasks,
- the role of individuals within those organisations in determining when and how general purpose AI services are used as a task assistant, and the best ways to ensure those individuals exercise appropriate restraint and care,
- the key role that transparency can play in building awareness of risks and capability to mitigate risks.

See further our overview in section 1 and response to question 2 above.

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?***

See our discussion in section 1 of this paper (in particular, on page 6) of the enforced self-regulation model as a way to change organisational DNA of organisations and enable flexibility as to the processes by which organisations adopt a risk-based approach for responsible AI.

A risk-based approach for responsible AI should be applied by:

- both developers and deployers of both foundational models and algorithmic decision-making systems and generative AI applications built upon those foundational models or algorithms, and
- organisations, both public or private, that are users of third party supplied foundational models, algorithmic systems, at least to determine whether risks of AI harms are sufficiently likely to have been assessed and mitigated by upstream developers and providers.

19 July 2023

Peter Leonard
Principal, Data Synergies and Professor of Practice, UNSW Business School