Bradley Holland*
Associate General Counsel, Senior Director
B.Eng (Hons), JD., LLM., M.Intl Tax., GradDip Marketing, GradDipArts (Indonesian)

The Australian Government is to be supported in its approach to strengthen the governance in the use of Artificial Intelligence (**AI**) by industry and government departments. Guidelines, regulations and governance are necessary where the use of AI may result in adverse impacts on people, particularly those who are vulnerable. It is well recognized that the application of AI-enabled systems and processes can greatly benefit society[1] and its well-being,[2] increase economic output and productivity[3] and improve access to services across a wide range of sectors from commerce, health, transportation, cybersecurity, environment management, science, engineering, technology through to the arts, border control, education and legal services.

In addition to providing responses to the consultation questions posed under Section 5 of the paper on 'Safe and responsible AI in Australia' (**Discussion Paper**), this submission has particular interest in, and will focus on, AI systems which produce responses that are: (i) text-based, including document generation and classification; and (ii) authorization-based, such as yes-no, or approve-reject type responses.

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

This submission generally agrees with the definitions as proposed in the Discussion Paper, noting the following:

A. The proposed definition of AI substantively aligns with Article 3 of the yet to be adopted AI Act of the EU.[4]

B. It may syntactically be incorrect to state that AI makes a 'decision', but more accurately makes a 'prediction' and in so doing, provides a 'response'. These predictions and responses are inputs to decision-making, and should not be deemed to be decisions themselves. Any decision-making would ultimately be the result of a process of the responsible persons or entities making use of the AI system. This is the ***accountability*** principle.[5] Associating decision-making with AI could have the effect

* Any views, thoughts, and opinions expressed by the author of this submission are solely those of the author and do not reflect the views, opinions, policies, or positions of the author's employer.

[1] Tabassi, E. (2023), 'Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI', *National Institute of Standards and Technology*, Gaithersburg, MD, https://doi.org/10.6028/NIST.AI.100-1 (Accessed July 31, 2023), 1.

[2] OECD, 'Recommendation of the Council on Artificial Intelligence', OECD/LEGAL/0449, http://legalinstruments.oecd.org (Accessed July 31, 2023), 3.

[3] Australian Government, Department of Industry, Science, Energy and Resources, 'Safe and responsible AI in Australia: discussion paper', 1 June 2023, https://consult.industry.gov.au/supporting-responsible-ai (Accessed July 31, 2023).

[4] European Commission, *Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, 2021/0106 (COD) (AI Act).

[5] Above n 3, 14.

of demonizing AI system algorithms and, in turn, displaces the decision-making responsibilities and accountabilities of those persons or entities who are deploying the AI systems. A previous definition of the European Commission attempted to avoid the use of the term decision in preference for setting a threshold or standard by which the output of AI could be regarded as valid as an input for decision-making, for example, descriptive material associated with the definition clarified the status of AI output viz-a-viz decision-making:

> "Many AI technologies require data to improve their performance. Once they perform well, they can help improve and automate decision making in the same domain."[6]

Hence, any 'decisions' that are made by the relevant organizations, government departments, entities or persons deploying AI are external to the AI system itself.

The decision-making capacity of AI needs to be made clear in the Australian Government's proposed definition, that is, whether the AI system makes a decision (which is capable of being disputed at law, for example under administrative law processes) or whether it is an input to a decision-making process of the organizations, government departments, entities or persons deploying AI.

C. The Australian Government's proposed definition of 'AI' makes no reference to any threshold or standard as to the intelligence of the system, and in that respect, the definition is broad, and could capture non-intelligent systems – the only standard is that the system be *engineered*, operate without *explicit programming*, and at *varying levels of automation*. The proposed definition of the European Commission overcame this situation by imposing a standard that an AI system "…display ***intelligent behaviour*** by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. (Emphasis added)"[7] This submission considers it necessary to clarify the type of systems to be captured under an AI definition to avoid overreach and unintended consequences. As an example:

> *There is a circumstance where an 'Uninstall Program' could be classified under the proposed definition as an AI system. The Uninstall Program generates a series of instructions to remove the program, after analysis of its environment. Under the definition, the Uninstall Program is clearly engineered, runs autonomously (generates its own code) and use algorithms to generate the uninstall code. Viruses and malicious code execute in a similar fashion. However, this is similar to how some recognised AI systems generate content.*

In this example, it is doubtful whether simple functions should be able to be classified as AI systems and therefore a standard of 'intelligence' should be required.

D. The listed set of 'predictive outputs' do not appear to include 'responses to stimuli', such as humanless systems that control vehicles, devices that monitor, analyses and apply medical therapies to reduce tremors, and other similar operator-less controlling functions. Although these may be included under 'decisions' (noting that this may be

[6] European Commission, *Communication from the Commission to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of the Regions*, Artificial Intelligence for Europe, COM(2018) 237 final, 1.
[7] Ibid.

an inappropriate term), perhaps this should be expanded to read 'decisions, or other predicted responses to stimuli', since there remains the question of whether decisions should be included in the definition at all.

E. While the definition of 'Generative AI models' includes generation of 'code in response to prompts', it is unclear where these prompts are human-generated (i.e. user input), or external system generated (e.g. monitored systems 'trigger' a response) or self-generated, where the AI-system rewrites its own code or system, usually in response to a trigger (regenerative or self-generative code) or through learned efficiency. It may be unnecessary to include 'in response to prompts' at the end of that sentence.

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?*

It appears that to be able to build trust within the Australian community to engage safely with AI, individuals must be aware when and how they are engaging with AI. This is the ***transparency and explainability*** principle.[8] To be meaningful, the amount of information to be disclosed may vary either by stage in the AI lifecycle or type of interaction with the AI system. The higher the level of transparency, the higher the level of confidence in the AI system.[9]

The Discussion Paper highlights an example of a significant risk to consumers and users of AI systems, with respect to the situation that occurred in the *Trivago* case.[10] There are several actions that can follow from the misapplication of AI responses, including negligent misstatement, malpractice, misrepresentation, misattribution, defamation etc. Some of these actions will depend on the classification of AI output at law, and what rights, obligations, and liabilities that flow from that output.

The Discussion Paper correctly points out that Australian Laws may be deficient in certain circumstances to provide an effective remedy. Given the speed at which AI is developing and the proliferation of output, there is a need for customers and end users to highlight their concerns and the impact of the responses on them as quickly as possible. Adequate access to support services will relieve anxiety and improve trust and confidence in the use of AI systems. The European Commission considered that for unfair results, "…users should be informed on how to reach a human and how to ensure that a system's decision can be checked and corrected."[11] To provide redress to affected parties, this submission suggests that businesses that benefit from AI deployments also share in the burden of dealing promptly and appropriately with AI complaints.[12] This could be dealt with through the nomination of

[8] Above n 3, 14, see point 8 of Box 3.

[9] Above n 1, 15.

[10] Above n 3, 12.

[11] Above n 6, 16.

[12] Tabassi et al considers that any measures dealing with transparency and accountability need to be tempered with the "…impact of these efforts on the implementing entity, including the necessary resources…", above n 1, 16.

an AI-Officer[13] who is charged with being the single-point-of-contact of the business to respond to customer complaints in connection with AI use or responses. If the affected party is not satisfied that the complaint has been sufficiently addressed, an escalation could be made to a government-sponsored party such as an Ombudsman. This improves the supervision of the AI ecosystem and makes visible any gaps in regulations, guidelines and policies or any deficiencies in the implementation of such regulations, guidelines or policies. Note the role of an AI-officer need not be the same as 'human in the loop' as set forth in Attachment C of the Discussion Paper, as the AI-officer does not need to be an 'approver or decider' in an organisation's decision-making process (incorporating AI). The AI-officer only needs to be the public-facing contact point.

Regulations could prescribe which statistics need to be gathered and reported to ensure that the policy and regulatory structures that the Australian Government decides to put in place from time to time are complied with. These procedures will also provide comfort and confidence to the community that AI deployments are under control and supervised.

Where the Australian Government decides to implement AI-based systems, an impact assessment should be developed for independent review on the effectiveness and impacts on that environment. Institutions that would require detailed impact analysis would be in the areas of Education, Health and Aged-Care facilities.

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?*

This submission proposes that the Australian Government consider having the Australian Law Reform Commission (**ALRC**) undertake a review of Australian and International Laws to determine the appropriateness of Australian Law to the opportunities and challenges presented by AI. In reviewing the fitness of, and recommending changes to, Australian Law to cope with the new AI landscape, the Australian Government can build trust within the Australian community in the application and usage of AI-based services. Further, the ALRC report findings also serve as educational tools to the legal and business communities in their preparation to deal with processes and procedures in AI-based environments.

At a minimum, for government deployed AI-based services, education programs should be developed to support the community generally and those most vulnerable in the advantages and disadvantages of AI use. This should include training and awareness campaigns on how to seek assistance in the management of complaints and also what community and support groups are available to help affected parties in understanding the impact of AI-based systems, such as how their rights may be affected, what information is collected and stored, and so on.

[13] Similar in concept to the requirements of Australian businesses to nominate a Public Officer, for the purposes of handling queries and correspondence from the Australian Taxation Office (ATO), and the Privacy Officer, for the purposes of handling privacy matters and concerns that may arise under the various State privacy regimes.

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 submission is of the view that Australia needs a National AI Centre of Excellence (**NAICE**) to implement a blueprint Code of Conduct (CoC) and Method of Procedures (MoP) or Recommendations in the guidance and supervision of AI-based systems, similar to the approach taken by the NSW Government under the NSW AI Assurance Framework. The NAICE, or similar body, would also be responsible for overseeing the application of the eight (8) AI Ethics Principles to the relevant AI-system, and the compliance of that system to the principles, CoC and MOPs and the reporting, complaint and escalation procedures highlighted in the answers to questions 2 and 3 of this submission.

The importance of transparency in AI systems to build community trust was highlighted under section 2 of this submission, but it is equally important that processes, progress and accountabilities of government are openly demonstrated with respect to the regulation, management and enforcement of AI safeguards. Reporting metrics are to be defined to ensure that government oversight and regulation is delivered at, timely intervals, and to expected standards.

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?*

The US proposes test, evaluation, verification and validation (**TEVV**) processes which are objective, repeatable and scalable to measure AI-based systems.[14] The TEVV approach ensures that metrics, methods, and methodologies are established, implemented and monitored and adapt over time. The relevant metrics and methodologies should "…adhere to scientific, legal and ethical norms and be carried out in an open and transparent process."[15] This submission considers that the TEVV framework could be worthy of consideration by the Australian Government to provide structure to risk governance. Some minor mapping of the US Framework[16] to the implementation of the Australian AI Ethics Principles will be needed.

The successful passing of the TEVV would 'certify' the AI system, such certification is a publicly recognisable symbol of confidence that the AI-system works in accordance with the objectives set for that particular system. This is no different and analogous to certification of electrical appliances, food health standards, emissions levels, etc.

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

[14] Above n 1, 28.

[15] Ibid.

[16] Ibid, 12, see generally the framework set out in Fig. 4.

This submission is of the view that risk and governance approaches to AI should not necessarily differ between public and private sector use of AI technologies, although governments may consider thresholds (e.g. turnover, number of employees, etc.) which need to be exceeded before regulations apply to that organisation. However, AI-based systems benefit large and small organizations and users of those systems could be impacted to the same extent regardless of the size of the implementing organization.

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

In the context of AI responses from Australian Government agencies that are text-based or authorization-based processes, the output of an AI system is a ***prediction***, which is the core function of any AI system. The ***accuracy*** of the prediction will depend on a number of factors, including the: (i) quality of the inputs to the system (e.g. end user's or customer's information); (ii) quality of the training dataset; (iii) performance of the selected model or algorithm of the AI system; and (iv) type and length of training that is applied to the selected model or algorithm. To establish trust in the Australian Government's use of an AI-based system, accuracy is a key criterion and therefore, government agencies will need to be completely transparent with respect to items contemplated in (i) to (iv). The manner in which government agencies can demonstrate accuracy is through the provision of test results to test scenarios; by having independent subject matter experts evaluate and report on the accuracy of the AI system; establishment of a certification or level of standards of operation that all AI systems need to meet which is consistent with world standards; introduction of a complaints, escalation and ombudsman process as discussed under question 3 of this submission; and provision of clarity (and enforceability) to the nature of 'decisions' made by AI systems as discussed under question 1 of this submission.

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.*

An advisory body that consists of technologists, lawyers, social and behavioural experts, risk managers, strategists and regulators/policymakers should be charged with making determinations as to what solutions should apply to risks in AI systems. The advisory body would also set rules and guidelines for any certification process of AI-systems. This can help avoid situations (e.g. Robodebt or the Climate Change debates) where management or motivated-interest groups are either incapable or unwilling to absorb and understand detail and complexity, resulting in the inability to make appropriate and informed decisions. Failure to appoint a suitable cross-discipline group, with the requisite number of experts such as lawyers or ex-judges who can formulate discrete questions, may result in an inability to construct questions-under-test to prove an accurate output. Particularly in high-risk cases, the questions that will be asked on AI-systems will be of greater importance, and hence, the propensity to have the question-answer/prediction tested in Court will be greater. The difficulty in assessing the capacity and accuracy of AI systems by any decision-making body in a legal context:

> "… may lie in formulating the question correctly. Once the question is accurately formulated, the answer is often obvious. But the quintessentially *intellectual* process involved in framing the question is much harder than it might appear to outsiders or to software programmers and coders — or to some of the managers in the upper reaches of the bureaucracy who commission automated systems. Those managers do not necessarily have hands-on experience of the complexity of a particular decision-making process or the real-life circumstances that may be encountered."[17]

AI governance and risk management demands a multi-disciplinary approach, not just the cross section of the advisory body, but perhaps also in respect of each of the members in that advisory body.

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.

The importance of transparency has been highlighted in questions posed in this submission, but the point at which transparency is most critical and valuable in mitigating potential AI risks is at the points when the AI system: (i) is implemented; (ii) updated (in terms of its algorithms and learning processes); and (iii) re-operationalized after it has been found to be inaccurate.

The three points above are points of proof of an acceptable working AI system, and the points at which an AI system is certified. There would likely be very little trust or confidence in a system that is uncertified.

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?

Candidate AI-systems that are used in diagnosis/diagnostics or which impose restrictions, limitations or obligations on persons or entities should undergo close examination and assessment when deployed in high-risk applications. In high-risk AI applications or technologies, the costs of errors can bring about disastrous results for affected persons, as highlighted by Bernard:

> "The likelihood of legal error increases exponentially when the decision-making process involves more steps, or more complicated enquiries. The potential for error

[17] McCabe, Bernard, "Automated decision-making in (good) government" (2020) 100 *AIAL Forum* 106, 118-119.

> — and the cost and inconvenience that results — has been amply demonstrated by the Robodebt initiative."[18]

Whether AI systems are to be banned in these high-risk applications depends on whether the response or result of the AI-system will be left unchecked or unvalidated. For AI-systems, its scope to make predictions is primarily restricted by its programming, and as a result, "…it may be fiendishly difficult to accurately translate the rule into programming language that is capable of predictable application in all circumstances." The probability of making an erroneous judgment or to be able to cope with novel situations can be handled satisfactorily by persons trained in the art of decision-making in the applicable field:

> "Human decision-makers, for all their faults, can reason from a rule to deal with new, unusual or nuanced circumstances."[19]

Hence, if an AI-system can make predictions for the relevant high-risk scenario, then the output response should be checked by a natural person authorised and trained in the relevant field as a first preference, before it is provided to the customer or end user. Whether the AI-system should be certified to provide these responses autonomously is something the advisory group, as contemplated under question 8 of this submission, could provide guidance on.

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

This question has been answered in detail earlier in this submission. It is clear that public trust will only be established, and people encouraged to engage in AI-based systems, if the TEVV results have integrity and are transparent.

Education and awareness campaigns can change people's perceptions, along with case studies and use cases that resulted in positive outcomes:

> "By promoting higher levels of understanding, transparency increases confidence in the AI system."[20]

In the situation where the application of AI-systems results in not meeting the standards acceptable to society, appropriate and proportionate remedial action should take place. For example, in the erroneous use of ChatGPT to generate affidavits and case filings which resulted in the creation of non-existent judicial opinions with false quotes and citations, the court ordered the perpetrator to officially inform the judges of the misattributed and fake opinions.[21] The naming and shaming of the perpetrator's activities, and well as adverse findings on character, in such a public and highly visible forum resulted in an extremely embarrassing (and possibly career ending) situation for the lawyers involved. Bringing these

[18] Ibid, 119.

[19] Ibid, 118.

[20] Above n 1, 15.

[21] *Mata v Avianca*, 22-cv-1461 (2023).

seemingly bad situations to the public's attention may in fact generate public confidence that mishandling and wrongdoing which results from poor AI-practices is being monitored and dealt with appropriately.

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?

It is the view of this submission that the banning of any AI in high-risk activities creates a competitive deficit in Australia's educational capabilities and potential negative effects on technical and economic growth. Each candidate AI system or area of activities for ban should be evaluated, perhaps by the advisory group suggested under question 8 of this submission. Again, some of the downside of deploying AI in these areas can be managed or mitigated by adopting humanistic controls as contemplated in the answer to question 10 of this submission.

However, this submission believes that where organizations use AI in hiring processes, facial-recognition, social scoring and similar high-risk activities, there should be a simple requirement for that organisation have a public disclosure statement which can be readily accessed by the public, setting out the organisation's nominated AI-officer who can be contacted by an affected person. Please refer to the discussion under question 2 of this submission regarding the role of the AI-officer.

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

Review and risk assessments will need to be made with respect to long-held views that software 'backdoors' or 'security bypass' code is prohibited coding practice. The law, codes of conduct, regulations and educational standards that relate to this prohibition (including all government procurement guidelines) will also need to be reassessed. This is based on the same reasoning that high-risk applications should have natural persons checking and validating outputs before imposing restrictions, etc. on end users or customers (see the discussion under question 10 in this submission). In the case where AI-enabled critical infrastructure operates autonomously, not to design intent, and outside of acceptable operational objectives, the system will need to be intercepted and brought back to a normal and acceptable state of operation. In the worst case, it will need to be halted and restarted (if that is possible). Suitably coded backdoors and security bypasses, properly managed of course, can be used to bring AI-systems back under control or halt the system safely. Runaway AI-systems are in no-one's best interests.

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

A risk-based approach is supported and achievable, provided there are clear guidelines and assistance to businesses and organisations to identify, review, assess and implement controls to manage these new risks.

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

As stated above, a risk-based approach is supported and achievable, provided there are clear guidelines and assistance to business and organisations to identify, review, assess and implement controls to manage these new risks.

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

Most likely yes, given the costs, effort and investment of organisations to make AI operate appropriately are sizeable. Larger organisations may be in a better position to absorb these costs and have the resources to develop and prove these systems are more likely to implement and manage a suitable risk-based approach.

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

This submission supports the Australian Government's selection of elements in Attachment C but the 'Human in the loop/oversight assessments' should include a human contact in an organisation, particularly government agencies, where affected persons can voice their concerns or lodge complaints/disputes where an adverse response has been provided by an AI-system. Please refer to the discussion under question 2 of this submission regarding the role of the AI-officer.

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?

Most organisations will have a structured review and risk assessment procedures in order to evaluate and, if necessary, amend their internal risk management plans, including business continuity and disaster recovery plans. Good corporate governance subjects the results of these assessment procedures for review by the board (possible through audit committees or its equivalent). The Australian Government should also give consideration to whether other standards-based organisations (e.g. Responsible Business Alliance) are likely to introduce guidelines and requirements with respect to AI management and usage.

Accordingly, this submission is of the view that organisations with good governance structures should be able to incorporate an AI risk-based approach into their existing assessment frameworks as standard business practice.

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

This submission has not answered this question.

> *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?*

This submission is of the view that the Australia Government should adopt a 'wait and see' approach before imposing regulatory compliance requirements on private organisations on the adoption and use of AI-systems. The implementation and operation of AI systems across the varied industry and public sectors is not homogenous, which makes policy-making, compliance implementation and monitoring burdensome and difficult. Hence, the risk-based approach to be initially adopted by industry should be ***voluntary***, however the implementation of an AI-officer as suggested under question 2 of this submission should be required as a minimum in all cases.

That said, this submission believes that the Australian Government should lead by example, and subject its departments and agencies to an AI compliance framework (which should include the elements set out in Attachment C of the Discussion Paper, and the AI-officer as suggested in the answer to question 2 of this submission), particularly where those departments and agencies impose restrictions, limitations or obligations on persons or entities.