Submission to Safe and Responsible AI in Australia discussion paper, 2023

Raymond Sheh, Adjunct Associate Research Scientist
Robotics and AI Researcher
Johns Hopkins University
https://www.linkedin.com/in/raymondsheh/

Karen Geappen
Cyber Governance, Risk and Compliance Consultant
https://www.linkedin.com/in/karen-geappen-4491094/

We are pleased that the Australian Government is working towards a consistent and managed approach to AI technology in Australia. The ability for Australia to safely and responsibly utilise AI from an individual to national level is key to protecting Australian interests, culture and way of life. We are unique in the world and therefore should not be blindly trusting or relying on other nations to define what is 'Safe and Responsible'. We hope our response to the 20 questions posed on the topic support existing views or elicit some thoughts on this small subset of AI considerations for its opportunities and risks.

Definitions

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

The use of ISO/EIC 22989:2022 is a great start, particularly in its emphasis that AI is a broad field, of which Machine Learning (ML) is only a part. The use of an internationally recognized definition also ensures broad compatibility of terms.

However, care should be taken to actually use this definition throughout the discussion paper. While the ISO definition encompasses ML as well as a wide range of other non-ML based AI techniques, a lot of the discussion paper uses AI synonymously with ML, which can be confusing and perpetuates misconceptions that the challenges and issues around ML exist in all of AI.

This issue flows over to the implied definitions of other terms. For example, Generative AI is a broad term that encompasses the use of AI to generate content of some sort, via both Neural Networks and other means. However, much of the discussion paper, and other documents, are written as if the problems that exist in Neural Network based Generative AI exist for all of

Generative AI. Such misconceptions are important to dispel because Generative AI itself is a valuable capability. Regulations that lump Generative AI technologies that do not have the problems of Neural Network based Generative AI with the same brush can only serve to reduce Australian competitiveness on the world stage.

The definition for bias given in the discussion paper is also somewhat troubling, because the paper ties it to the notion that increasing the comprehensiveness of a dataset reduces its bias. There are many applications for which there is simply not enough data to remove all bias and others where more data, in highly skewed populations, can increase bias, by drowning out poorly represented examples. Data is also historic, it illustrates where we have been but it doesn't necessarily reflect where we wish to go. This is how training data can perpetuate historic bias.

In our view, the problem with this definition of bias in part stems from the aforementioned misconception that AI is ML and ML makes decisions based on data. In fact, many forms of AI and ML (although generally not Neural Network based Deep Learning, at least without substantial modifications) can incorporate background information that is separate from the training data. For example, background information could be the assertion that there should not be a correlation between skin colour and the probability that a system detects that a person is holding a weapon. This information can be used to initialise, constrain, and otherwise modify the way in which the ML system makes use of the data to manage bias. It also allows the system to understand where we want to go, to reduce real biases in society going forward. This mechanism in the definition of bias is often forgotten and deserves to be highlighted.

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?

There are three major risks that we see in the use of AI that do not appear to be covered by the discussion paper and by existing regulations (at least not until it's too late).

The first relates to AI interfering with the ability to perform root cause analysis of failures[1]. In this context, examples of failures include an AI based medical diagnosis system that provides bad data to a doctor, or a car with an AI based pedestrian detector that fails to detect the pedestrian and causes a collision. In this context, different forms of AI provide different abilities to perform the root cause analysis often required by regulation. Some, including many ML techniques, make it difficult or impossible to tell where the failure occurred, be it in the AI system, in the

[1] Sheh, R. and Monteath, I. (2017) Introspectively Assessing Failures through Explainable Artificial Intelligence. In IROS-17 Workshop on Introspective Methods for Reliable Autonomy (IMRA).

wider system, in the training of the system, in the sensors, or so-on. In contrast, some other AI techniques, such as some that derive mathematical models or express their decision making process in the form of first-order logic, can trace their decisions back to first principles that can satisfy such requirements. Such risks are best tackled at the “by design” stage. By the time existing regulators might encounter this risk, the systems have usually already gone past the point at which these risks can be addressed.

This should be accompanied by regulations regarding legal obligations to address these failures and make them right. Of course such regulations are not limited to AI and should cover all Automated Decision Making (ADM) systems. The well publicised Centerlink Robodebt incident[2,3] is an example of a non-AI based ADM system that should similarly be covered. It demonstrates the urgent need for well defined regulations surrounding how to expeditiously address an automated system that fails to make the right decisions, especially where it has significant legal implications for a sector of the population and can result in tragic adverse outcomes, particularly if legal remedies take a long time to resolve.

The second is related and concerns the ability for AI to hide or obfuscate other problems. For example, an unethical person or organisation could blame AI for something that they were actually responsible for, to take advantage of greater leniency, or the propensity for society to either over-trust AI, or be resigned to accepting a decision “because of AI” and not investigate further. We already see this when people “blame IT” for problems that have nothing to do with IT and the blaming of ChatGPT (classed by some as Generative AI) for incorrect citation in legal cases due to over-trust[4]. Regulations and public education around AI risks should harmonise with existing approaches to minimise any difference in real or perceived leniency to mistakes and to avoid AI being used as an excuse to not offer a reason or explanation.

The third acknowledges the cross cutting nature of AI and its ability to both draw data from, and influence decisions across, a wide range of areas that are traditionally regulated individually. For example, datasets that include a person’s health, spending, and telecommunications practices would fall into a range of areas that are regulated differently. Regulators will need visibility and the ability to co-operate and harmonise to deal with such situations and avoid over-regulation, regulatory turf wars, or oversight falling into the gaps between areas.

---

[2] Information about Robodebt - https://www.servicesaustralia.gov.au/information-about-robodebt last accessed 2023-07-25

[3] Robodebt scheme - https://en.wikipedia.org/wiki/Robodebt_scheme last accessed 2023-07-25

[4] Lawyers in the United States blame ChatGPT for tricking them into citing fake court cases - ABC News - https://www.abc.net.au/news/2023-06-09/lawyers-blame-chatgpt-for-tricking-them-into-citing-fake-cases/102462028 last accessed 2023-07-25

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.

A topic that is mentioned in passing in the discussion paper and deserves greater attention is more broadly funding the development of AI expertise in Australia that is not connected with industry. We hope that social obligations form a part of a company's conduct but at the end of the day, there needs to be significant expertise within government and other organisations who can advocate on behalf of consumers and small/medium sized businesses. Initiatives that make such positions competitive with industry in attracting the best talent is vital to ensuring that any regulations regarding responsible AI practices have appropriate oversight.

One particularly important aspect in the development of expertise is the development of metrics and requirements that AI systems in different sectors should satisfy, and people who are proficient in their use and further development as technologies and applications change. These actionable definitions and quantitative measurements are vital to ensuring that everyone is actually on the same page when it comes to appropriately assessing their use of AI. They are therefore critical building blocks of effective, fair, and transparent AI regulation and development.

Funding a broad base of research into AI, particularly in techniques that may have an ethical or regulatory advantage, is also vital. For example, right now Neural Network based Deep Learning based systems and related technologies have a lot of attention and private sector funding, despite their many regulatory issues, particularly around the legal and ethical use of data, their lack of transparency or correctability, uncertainties around bias, and so-on. Funding research into alternatives that satisfy both the performance and regulatory requirements ensures that Australians have access to responsible AI systems. It also broadens the capabilities of the Australian AI industry beyond those catering to data heavy applications. Not all applications of value to Australia can furnish the "rich, large, quality data sets" required of many forms of ML, and there are many other ML and non-ML techniques within AI that are better able to deal with such situations. Such applications, which may not attract much private sector funding and yet have large societal benefits, such as managing resources and infrastructure across our wide, sparsely populated, and yet constantly varying country.

Australia's unique political situation, having cohesive government, wide-ranging government services, and a particularly strong sense of community, fairness and ethics, also provides an opportunity to cultivate an international reputation for a highly ethical AI industry on the world market. Imagine a future where Australians are known for their ethical AI, just like the Swiss are known for their precise watches or the Japanese for their quality electronics.

Educating the general public about AI in the services that they use is also vital in encouraging responsible AI practices and making them a marketing feature. Star ratings, such as exists for automotive safety, food nutritional value, and appliance energy efficiency, have been very

effective at driving both consumer awareness and improvements in the industry. An AI Ethics star rating, most likely tailored to the application (in the same way that energy star ratings for television ratings, washing machines and refrigerators differ slightly), would be a valuable non-regulatory initiative, albeit one that requires significant expertise to ensure that it appropriately captures Australian AI ethics and values. Similarly, developing and encouraging the use of a common Product Disclosure Statement, such as was introduced for financial services, can help to further educate consumers and the public. Such a document would describe the nature of the AI used in a system, opportunities for opt-out, oversight, any minimum performance guarantees, ways to correct and challenge, and so-on, in a common format and plain language. This will enable consumers to make a more nuanced choice between using different products and services, based on the risk that they are willing to take in return for the services rendered.

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.

Experienced AI practitioners with visibility across the widely varying techniques and applications of AI, including non-ML types of AI, are in scarce supply, especially for a country like Australia with a small population. The preservation of this expertise, and a coordinated approach to its sharing across different government sectors, can be crucial to ensuring the appropriate development and use of AI. It also assists in reducing unnecessary duplication and inconsistency, and increases the efficiency/applicability of AI research and initiatives that can then be leveraged to greater benefit a wider spread of Australia. An added benefit of coordinating AI governance across different sectors is fostering diversity in background related to AI where diversity is shown to produce greater and more robust innovation.

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?

The topic of datasets was mentioned in the discussion paper but one aspect that can be powerful, particularly for a small, highly distributed population such as Australia, is synthetic data. This is data, such as about a population, that is generated from the real data and has similar properties for a given application, but where no one datapoint in the synthetic data can be matched to a real datapoint. Organisations such as the US National Institute of Standards

and Technology, are performing research into the use of privacy engineering in general[5], and synthetic datasets in particular, which may be used to develop and test AI applications in situations where there is a large privacy risk. Governance measures may include requiring only synthetic datasets be used for development and testing until a given implementation has been proven to be suitably privacy preserving, only after which access to real datasets are provided.

On the privacy front, improved regulation on the protection of personal data that is, or can be, used by AI is needed. The current Commonwealth Privacy Act (1988) is old and the method of administration by the OAIC is not suitable to AI technology for the protection of an individual. European GDPR regulations, as mentioned in the discussion paper, are imperfect but they are a start. Much can be learned from the issues that have arisen in the years since adoption and how they apply to Australia, especially given the variation in privacy culture across Europe.

As was also mentioned in the discussion paper, the European proposals for improved regulation on attribution will also be highly desirable to enable end users to identify the human contribution vs the AI output.

Target areas

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

Approaches should be tailored to their impact, consistent with risk management best practice. In that sense, for applications with the same level of impact, public and private sector uses of AI should not differ.

Certainly for the same application, there should be no difference in rigour or transparency of AI in the public or private sector. As has recently been demonstrated in the Australian economic environment, the Private sector can have the same or greater influence on Australia and Australian National Security as the public sector. Instead, standards should correspond to risk, which includes a component of impact and influence. These should be rooted in the AI Ethics Principles and be quantifiable and measurable in a meaningful way.

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

The most impactful support that the Australian Government can offer to further support responsible AI practices is in growing internal talent and expertise in AI. This is a challenge as

[5] NIST Privacy Engineering - https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering last accessed 2023-07-25

this talent is also in high demand in the private sector, thus it is essential that public sector jobs for such talent are appropriately attractive. Education and communication with the public is also vital so that they can understand the conversation around AI, properly assess their risks, and inoculate themselves against marketing and misinformation surrounding its capabilities and dangers. Rolling out public education and mechanisms such as the aforementioned star ratings and Public Disclosure Statements, to ensure an informed population, and robust ways to report issues, such as an AI Ombudsperson, as will be discussed later, will also help to further support responsible AI practices within government agencies.

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.

Solutions to the risks of AI should, at an abstract level, be outcomes based. For example, risk in a critical system is reduced by the ability to perform root cause analysis in the event of failure, and verifiably roll out fixes going forward. A generic solution at an abstract level is that the AI system needs to permit root cause analysis through the system, to a given level. However, once this becomes specific to an application, it often requires solutions that are technology specific to achieve the higher level outcome. For example, performing root cause analysis, to a given standard, must be done differently in logic or mathematical model based systems as compared to systems based on Deep Learning (or not possible as the case may be and require different measures and restrictions to be put in place). So in a sense, the answer to this question is one of the level of abstraction at which the policy is applied, rather than for a given circumstance.

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?

Transparency is most vital in improving public trust and confidence when it tells them something new and gives them a practically actionable choice. Unfortunately this often goes against the interests of those deploying AI. If the consumer is happy and doesn't know that AI is involved, why tell them? In that sense, regulations around transparency must also be accompanied by regulations that provide the public with actionable agency such as the option to contest its output or decision due to its output. There must be a way to remove themselves from the AI system, or not be subject to it as part of managing their risk. If there is no option, the transparency is not actionable.

This transparency must also go hand-in-hand with common definitions for AI and transparency itself. The practice of "transparency-washing" or "explainability-washing" is already commonplace. This is where systems provide explanations for AI based decisions, with the aim

of increasing trust or acceptance of the decision, and that may be consistent with some of the decisions, but do not reflect the actual, underlying (and often black box) decision making process. Examples that have already been misleadingly sold in this manner, by major IT vendors, include systems that make use of gradient methods, such as LIME and SHAP and their derivatives. At the root of this problem are the wide and varying uses of terms like "transparency" and "explanation". Worse yet, different stakeholders have very different needs and requirements for transparency and explanation. Coming to common, technically actionable, and ethically meaningful definitions for these in different sectors will be vital before it is possible to effectively determine how this is applied[6].

One challenge surrounding transparency is in conveying to the user when AI is even being used, versus a human or non-AI based ADM system, and where there is or isn't human oversight of such a system. Imagine if the Centerlink Robodebt incident happened now, where there would be valid concern that AI was somehow also involved. In this regard, the need for transparency extends beyond AI applications, to those that could be erroneously seen to use AI.

For services that make decisions with a legal or similarly significant effect on an individual's rights, the aforementioned Product Disclosure Statement could therefore be necessary more generally, beyond AI. Such a document would outline, in a standard, plain language manner, what kind of decision making process is used (AI, non AI ADM, human, with or without oversight, and so-on), the amount of insight that the consumer can expect in the decision making process, the mechanisms by which decisions may be challenged and information corrected, options for deleting their data from the system, and any guarantees of minimum performance that may be present.

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

A major challenge with transparency requirements is in specifying, in actionable terms, what does and does not count as transparency for a given sector.

This definition will vary by application and impact but should be consistent regardless of being private or public sector. Transparency around why a particular video was recommended to a user may simply extend to noting the data that might have influenced it. In contrast, transparency around how a hiring decision was made may need to be both more specific, and provably truthful to the underlying decision process, rather than an explanation that happens to fit some of the decisions. A coroner investigating a death due to an AI failure, such as in a medical or industrial system, should require transparency that allows for true root cause analysis[7].

---

[6] Sheh, R., and Monteath, I.(2018) Defining Explainable AI for Requirements Analysis. In KI-Künstliche Intelligenz, 32:4, pp 261-266, Springer Berlin Heidelberg.

[7] Sheh, R. (2021) Explainable Artificial Intelligence Requirements for Safe, Intelligent Robots. In the 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR).

AI systems that cannot offer such capabilities should be subject to the same standards, and prospect of recall, as other life-critical industries that have components that are not understood well enough to be subject to full root cause analysis. For example, in the pharmaceutical industry, it is often the case that how a medication works, and why it has certain effects on certain people, is not fully understood. Instead, extensive studies, testing, properly informed consent, the involvement of trained practitioners, ongoing monitoring, and the prospect of recall are necessary.

These requirements need to be designed in at the start. By the time a fatal accident happens and a coroner finds that the decision making process is a black box, it is too late.

10. Do you have suggestions for:

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

Bans on specific applications and technologies should be used very sparingly, not only in AI but more generally, and only based on what is illegal regardless of AI. Beyond those, an outcomes based approach should be taken, which requires particular applications to have particular capabilities. For example, AI that is used in safety critical applications should have the capability for root cause analysis and verifiable correction if an error is found. This is no different to what happens in the automotive or aerospace industries, where if a failure happens in, say, an aircraft engine, it is necessary to find the root causes and fix them in a verifiable way. This may result in a de-facto ban of certain technologies that do not yet provide these capabilities, such as high performance Neural Network based Deep Learning, but also incentivises further development that may either add these capabilities to Neural Network based Deep Learning in the future, or improve the performance of other, compliant technologies to match those of Neural Network based Deep Learning.

Furthermore, banning technologies can stifle domestic research and innovation. If Australia is not doing our own research and development in the area, at some point this kind of technology, or at least output of if will be imported whether legally, illegally or inadvertently within another technology or device. This importation will also carry with it the risks of all the inbuilt known, unknown and unforeseen bias. Some of which will be purely on the basis that the Australian culture is unique, even compared with those most closely aligned such as the US and UK. So something that may be bias in Australia is not in those countries. This doesn’t mean that restrictions, possibly via regulation, for high-risk AI applications should not be in place, but that these should be centred again around the AI Ethics Principles and have an emphasis on being transparent and measurable in its risks. Thus, like any other dangerous goods such as certain pharmaceuticals, explosives, or nuclear technology, dangerous AI should be restricted rather than banned.

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

As mentioned above, applications that are already illegal should of course stay illegal, such as various forms of deception, regardless of the presence of AI. Progress here can be made by updating the various regulations to account for AI, and improving the capacity of regulators and enforcement agencies to understand the state-of-the-science and enforce these existing laws appropriately.

Placing additional bans on applications simply because of the use of AI is both un-necessary and impractical, in large part because it turns into a semantic argument. Instead, the aforementioned outcomes based approach should be used to drive research and development of AI technologies for a given application in a direction that is acceptable to society. There also needs to be provisions in place to ensure that Australia has the in-house skills to assess and detect and then categorise imported AI so that anything that is restricted has a greater chance of being discovered.

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

The Australian government has been very lacking in educating the public. A quick read of Australian media even goes to show there is little in the way of standardised terminology or definitions. This means that the general population will remain uneducated at best or completely confused on any further discussions related to AI risk or AI trust. This hinders the appropriate and safe use of AI and forgoes an opportunity to give industry an incentive to invest in ethical AI development.

Thus the biggest impact we can see in initiatives to increase business-to-business and public trust in AI revolve around education. In particular:

- Education at all levels around when it comes to AI, and what different forms of AI are, and are not, useful for. This includes balanced guidance on opportunities vs risks of using different forms of AI (or how to assess for one-self), or even deciding not to use AI at all[8], perhaps guided by the aforementioned star ratings or Product Disclosure Statements.
- Regulations around misleading use in advertising of the terms AI, ML, transparency, explainability, trust, and so-on. This may require updating existing regulations to more unambiguously apply to AI.
- Better and more informed enforcement, including growing in-house expertise at all levels of government. These should also be suitably disclosed so that the general public is aware that issues are being addressed.

[8] Cakes can't be unbaked: why you should think twice about AI https://anchoramconsulting.com.au/blog/f/cakes-cant-be-unbaked-why-you-should-think-twice-about-ai last accessed 2023-07-25

- Investing in home-grown AI technologies and research, including streamlining the Australian Research Council (ARC) funding process.
- Fostering competitions and other outreach activities to high schools around AI applications.

Beyond education, knowing that those who develop, sell, and deploy AI systems can be held accountable is also vital in increasing public trust, just like in other high risk sectors such as air travel or healthcare. Initiatives that could help here include:

- Requiring high risk and impact sectors to have systems that can be subject to audit and assessment.
- Having an “AI Ombudsperson” with the resources to address concerns raised and to publicise investigations and outcomes as appropriate.
- Fostering transparency, such as the star ratings or Product Disclosure Statements mentioned previously, with in-house government expertise to ensure that they are not misrepresented.
- Transparency around rulemaking and enforcement to combat misinformation, including clearly defining “misinformation” and providing means for contesting categorisation as such.
- Ensuring that people have control over their data and intellectual property, including regulation and enforcement around the mass gathering of such data to train AI systems and ensuring that those whose intellectual property is used by AI have robust means to be appropriately compensated.
- Ensuring that meaningful human review of decision making processes are not subject to de-funding, both within government and in the private sector.
- Providing a robust notifications and recalls mechanism, like for aircraft and motor vehicles, to address AI issues.

Implications and infrastructure

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

We need to be careful with banning terms rather than activities. It sounds good to ban social scoring but they are not so far removed from other scores that we are familiar with, like credit ratings, or ones that we don’t even have visibility into, such as the visibility of our posts on social media. Banning social scoring but ignoring similar use cases results in a regulatory environment that risks either encouraging loopholing, or over-regulation, depending on how broadly the regulation is interpreted.

Instead of bans, tight, technically actionable, outcomes based requirements on the capabilities of high risk activities, such as the need for real transparency and explainability, the ability for people to truly delete their data, and so-on, can be applied. This can foster Australia's tech sector as one with a reputation for producing responsible, trustworthy AI that is held to a high standard domestically and can command a premium internationally.

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

Risk-based approaches

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

Yes. AI, being such a wide ranging technology and wide applications, needs an approach that is adaptable to the technology's application which a risk-based approach caters for. The approach must be standardised against an Australianised AI Risk Management Framework that is linked back to the Australian AI Ethics Principles and the framework must be measurable. The discussion paper already mentions the US National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF)[9] is a good starting point for developing an AI Risk Management Framework suitable for Australia. Some of the concerns we raised in the development of the NIST AI RMF are still relevant[10]. Australian contributions back to the NIST Trustworthy and Responsible AI Resource Center[11], particularly in terms of playbooks for the NIST AI RMF, can also be valuable in harmonising Australia's risk management approach to AI.

It must also be logical such that the general public is able to trace the analysis if they have reasonable common background knowledge. Controls that are then applied (or not) can be logically traced back to the risks as applicable to that particular deployment.

The risk based approach must also include constant updates and reassessments, must also occur in all sectors including those that might not have inherently changed, and include those that may not be using AI at all. For example, some Australian government agencies have deployed AI that analyses the voice of callers to authenticate them. The risk of using such AI systems has changed considerably with the advent of Generative AI that can learn someone's voice and generate fake, but very believable, audio of them saying anything. Such risk

---

[9] NIST AI Risk Management Framework - https://www.nist.gov/itl/ai-risk-management-framework last accessed 2023-07-25

[10] Comments on the First Draft of the NIST AI Risk Management Framework - https://www.nist.gov/document/ai-rmf-rfi-comments-raymond-sheh-and-karen-geappen last accessed 2023-07-25

[11] NIST Trustworthy & Responsible AI Resource Center - https://airc.nist.gov/home last accessed 2023-07-25

assessment updates should occur across the board in response to developments in AI. A traditional risk-based approach will have triggers for reassessment when there is a significant change in the system, a significant change in the system's environment it operates in or a significant change in risk appetite. This must also apply to a risk-based approach to managing AI risks.

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

The main benefit is that it provides information (hopefully standardised) to the consumer on the potential for AI to cause both desirable and adverse outcomes, and what these may be. It will inform implementers, deployers and developers on what controls are applicable and pragmatic in their application of AI technology, as opposed to a compliance based approach where a control may be irrelevant or, worse, cause greater risk. Limitations is that such an approach will not account for all eventualities and all risk assessments are at some level qualitative and subjective to the assessor. There is also the need for constant reassessment to determine if controls are still applicable, proportional and pragmatic.

These limitations can be overcome for the former by having a well written Australianised AI Risk Management Framework and guidance tooling to assist consumers in the application of the framework. A possible library or catalogue of commonly used AI risk management controls (e.g. synthetic data pool) could be made available which accurately reflects the Australian operating environment along with indicators on how effective such controls are at managing various AI risks.

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?

For some high risk sectors there could be a combination of risk based and regulatory based approaches. A combination of (for example) certification against an appropriate build, transparency or ethics standard, and a risk approach for any additional controls that reduce risks to a predetermined appetite. This could, for instance, be for transport, critical infrastructure (including medical) or educational settings. The education sector is particularly relevant due to the risk of embedding bias into generative AI that can be used for external influence with an impact to National Security or Australian Cultural values.

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 are in general suitable, however without further information on measurement they can be subjective. This reinforces the need for home-grown expertise in the measurement,

requirements analysis, and capabilities assessment, of AI systems, which is still very much a topic of research, at home and abroad.

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?

Existing assessment frameworks such as the Privacy Impact Assessment, used when dealing with Personal Information under the Australian Privacy Act (1988), are relevant if considering existing frameworks to incorporate an AI risk-based approach. For such frameworks, it should be relatively easy to include impacts as a result of the AI into assessments. Similar can be said for Risk Management Frameworks where the consequence for the use of AI is included in analysis and calculations. Where there is an existing framework focusing on the outcomes, it may be more advantageous to create guidance on its application to AI technologies or services.

The guidance can be around how to assess for impacts from AI that align with impact already in the existing framework. Or the guidance can prompt ‘lines of thought’ that may be different in the AI system to non-AI systems as applicable to the framework.

By incorporating, through guidance, the AI technology considerations into existing assessment frameworks, organisations are able to compare AI and non-AI technologies and services to give a more balanced risk vs benefit vs cost analysis.

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

Risk based approaches are even more important for the developers of LLMs and MFMs. Their subsequent users may be unable to appropriately perform their own risk assessments downstream of these systems and apply any risk management controls as required. Instead they must rely on the (perhaps partial) risk assessments of those providing the LLMs or MFMs. Unfortunately, performing appropriately meaningful risk assessments of large AI systems is still a topic of novel research that requires significant funding.

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?

Both, with varying levels of rigour on the application of the tool dependant on the risk consequence of the AI in use. The higher the consequences the greater the rigour and closer to regulation of assessment and implementation of risk management controls.

Regardless of the level of rigour applied, all systems should disclose their assessed residual risk as discussed previously. Whether in a Disclosure document or a star rating. This allows a consumer to enact any additional risk management controls they may wish or to request removal from the AI system/service (as per above mention for being meaningfully actionable).

b. developers or deployers or both?

Both, as above. As with all technology, risks must be considered from the inception of the idea through to eventual deployment, sustainment and decommissioning in the technology lifecycle. With AI this is no different but in fact more critical. AI will enhance technology capability but in balance to that also enhances risks of that technology. Meaning risk considerations must be applied at all stages of its lifecycle.