Director

Technology Strategy Branch

**Department of Industry, Science and Resources**

Industry House

Canberra ACT 2601

26th July, 2023

Dear Department of Industry, Science and Resources,

Thank you for the opportunity to respond to the Department’s discussion paper on responsible AI.

I am responding as a concerned individual; I do not currently work in the field of AI, but I have been keenly following developments in the area for some years now.

I believe that AI technologies may bring unimaginable benefits to humanity over the coming years but, unchecked, it may also bring devastating impacts. I feel that it is crucial for the government to do everything possible to ensure that advances in AI proceed in a manner that is beneficial to all Australians, and to humanity.

Yours sincerely,

David Taylor

**Part 1: Initial Comments**

My primary concern with the measures outlined in the discussion paper is that these fail to address the risks associated with AI models or systems of human-level or greater intelligence.

I feel that it is likely that the capabilities of AI models / systems will continue to increase, rapidly, over the coming years. On the current trajectory, I believe that the intelligence and capability of these models will soon reach a level of general intelligence that is equivalent to that of a human. Continuing on the trajectory, shortly after this, it seems inevitable that the models / systems will achieve levels of general intelligence that exceed that of human beings.

If such a model or system’s goals and motivations are not fully aligned with the wellbeing of humans and human culture, this would represent a true existential risk to us all.

In an attempt to avoid negative outcomes arising from the scenario above, I propose the four following recommendations:

1. Regulate the release of strong artificial intelligence models, systems or applications into production use

In the same way that the TGA regulates the release of medicines and medical devices, the Australian government should establish a regulatory body to provide evaluation, assessment and monitoring for the release of strong[[1]](#footnote-1) AI products into the community.

Before an AI product is released to the community or put into production use, the regulator would need to perform a thorough assessment of the technology to ensure that the risks represented by the product are acceptable.

It should be noted that this regulation would not limit AI research activities, only the deployment of a developed product.

1. Transparency of organisations’ progress and results in AI and related research activities

Organisations should be required to report to the AI regulatory body on the results, performance and capabilities of AI systems, models or applications that they may be developing. The objective of this transparency would be to provide the government with a degree of foresight into AI advances.

Reporting to the regulatory body should include the results of AI model performance against applicable benchmarks but would not need to include details of how the models were constructed, implemented or trained. Results would not need to be verified by an independent 3rd party.

Reporting would be confidential between the organisation and the regulatory body.

1. Board level accountability of the use of AI technology within an organisation

I believe that there needs to be explicit board-level accountability for the use of AI technologies within an organisation.

The board should be ultimately accountable for the outcomes, actions, advice or decisions arising from the use of AI technologies within the organisation.

The board must be aware of the use of AI technologies within the organisation, and must ensure that these uses are fair, safe and comply with applicable regulations and legislation.

1. Self-identification of AI agents

Finally, I believe that there should be a mandatory requirement for AI technologies to identify themselves as artificial intelligence when they are interacting with humans. The objective of this measure is to somewhat ameliorate the risk of humans being manipulated by future AI technologies (i.e. by thinking that they were dealing with another human).

This requirement should apply where the AI is providing a service (e.g. customer service chatbot), where the AI agent is interacting with services provided by a human (e.g. making calls or sending emails on behalf of an individual or company), or when the AI is creating or posting content in the public domain (e.g. social media posts, authoring news articles).

The human parties in customer or service provider interactions should have the option of instead dealing with another person instead of the AI agent.

# Part 2: Answers to Discussion Paper Definitions

**Q1 Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?**

|  |
| --- |
| No response |

# Potential Gaps in Approaches

**Q2 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?**

|  |
| --- |
| I don’t believe that the existing regulatory approaches outlined in the discussion paper adequately address the risks associated with future AI systems, models or applications possessing artificial general intelligence at a level that is equivalent to or exceeds human intelligence.  My suggestions for possible regulatory actions to mitigate these risks are:   * Require governmental approval before the deployment of strong AI systems, models or applications into production. * Transparency of organisations’ progress and results in AI and related research activities.   These points are outlined in more detail in “Part 1: Initial Comments”. |

**Q3 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.**

No response

**Q4 Do you have any suggestions on the 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.**

No response

# Responses Suitable for Australia

**Q5 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?**

No response.

# Target Areas

**Q6 Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?**

Approaches to regulation of AI technologies should be consistent, regardless of whether these are used in the public or private sectors.

**Q7 How can the Australian Government further support responsible AI practices in its own agencies?**

No response

**Q8 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.**

No response

|  |
| --- |
| **Q9 Given the importance of transparency across the AI lifecycle, please share your thoughts on:**   1. **where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?** 2. **mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.** |

|  |
| --- |
| As outlined in “Part 1: Initial Comments”, and the response to question 2, I believe that there must be transparency into the state and progress of AI research within organisations.  Organisations should be required to report to a government regulatory body on the results, performance and capabilities of AI systems, models or applications that they may be developing. |

|  |
| --- |
| **Q10 Do you have suggestions for:**   1. **Whether any high-risk AI applications or technologies should be banned completely?** 2. **Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?** |

I believe that the use of AI technologies in autonomous weapons systems should be banned completely.

Beyond this, as outlined in “Part 1: Initial Comments”, I believe that the release into production of strong AI technologies should be subject to approval by a regulatory body.

**Q11 What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?**

I believe that the following initiatives would increase public trust in AI:

* Explicit board-level accountability for the actions, decisions, advice or outcomes from AI deployed within organisations.
* A requirement for AI technologies to identify themselves as AI when interacting with humans.

These points are outlined in more detail in “Part 1: Initial Comments”.

# Implications and Infrastructure

**Q12 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 response

**Q13 What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?**

No response

# Risk-based Approaches

**Q14 Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?**

I support a risk-based approach for addressing potentials AI risks.

**Q15 What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?**

I feel that the ‘possible draft risk management approach for managing AI risks’ outlined in the discussion paper falls short as a risk management system. The possible draft proposes a risk rating table, with examples and proposed mitigating controls, but the approach fails to discuss the risk assessment process required to arrive at a risk level for a given AI technology.

There could be a very high degree of subjectivity in an organisation’s determination of the risk level of their technology, with a clear benefit to arriving at a lower risk level in their self-assessment.

**Q16 Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?**

The risk-based approach should be flexible enough to be applied consistently to all organisations implementing AI technologies.

**Q17 What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?**

|  |
| --- |
| I support the elements presented in Attachment C. |

**Q18 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?**

No response

**Q19 How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?**

The risk-based approach should be flexible enough to accommodate LLMs, MFMs, and future generations of general purpose AI systems.

For general purpose AI systems, a threat modelling approach may be warranted, where the misuse or abuse cases are considered, rather than the typical usage cases anticipated by the designers.

As discussed in “Part 1: Initial Comments”, there should be a requirement to obtain regulatory approval for strong AI technologies before their release into production. This would provide the regulator with an opportunity to exercise any necessary assessment of general purpose AI systems.

**Q20 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:**

1. **public or private organisations or both?**
2. **developers or deployers or both?**

The risk-based approach for AI technology should be mandated through regulation.

It should apply to both public and private sector organisations.

It should apply at the point at which the AI technology is deployed into production. This ensures that regulation does not stifle research and innovation, but provides protect to the community.

1. I acknowledge that ‘strong’ is somewhat subjective in this context. My suggestion would be to establish a minimum capability or intelligence threshold above which this regulation would apply – perhaps a level of capability equivalent to current LLMs (e.g. GPT4, Llama 2, PaLM 2) [↑](#footnote-ref-1)