(human compatible) (submission)

> *The Australian Government seeks public feedback on potential AI risk mitigation strategies and the promotion of safe and responsible AI practices, intending to build governance mechanisms and inform appropriate regulatory and policy responses to foster public trust and maximize the benefits of AI.*

This document is a response to this request submitted by **Human Compatible**.

(synopsis)

(discussions)

The synopsis of our response:

(themes)

- Support for the definitions in the discussion paper and the overall risk-based approach to AI governance.
- Emphasis on a balanced regulatory approach that builds on existing frameworks while addressing emerging issues as they arise. Avoid overly restrictive regulation that could stifle innovation.
- Recommendations for additional non-regulatory initiatives like independent advisory councils, testing facilities, community participation programs and certification schemes.
- Support for transparency, human oversight, training, documentation and risk assessments proportionate to the level of risk.
- Caution to ensure "human in the loop" requirements in all circumstances and where efficiency, speed and scale are priorities.
- Highlighting the need for international coordination on AI governance, especially for emerging capabilities like generative models.

(observations)

- Existing laws provide a baseline level of protection but gaps exist in regulating systemic impacts and emerging capabilities.
- Non-regulatory initiatives can complement regulation in building trust and capabilities.

- Risk-based requirements should be proportional to the level of risk posed.
- Blanket human oversight requirements may not always be feasible or desirable.

(opportunities)

- Develop specialised conformity infrastructure to support standards and auditing.
- Leverage international collaboration on technical standards.
- Support SMEs to adopt AI responsibly through initiatives like the AI Adopt program.

(risks)

- Overly restrictive regulation of AI could disadvantage Australian companies.
- Failure to regulate emerging high-risk applications before widespread harm occurs.
- Lack of coordination across government creates a complex regulatory environment.

(recommendations)

- Establish an independent advisory council on AI ethics and governance.
- Develop voluntary certification programs and testing facilities with industry collaboration.
- Require impact assessments, transparency and contestability mechanisms for high-risk AI.
- Take a precautionary approach by monitoring and preparing mitigation strategies for extreme long-term scenarios.
- Support international coordination on identifying and governing emerging capabilities.

(our response)

(definitions)

> *Do you agree with the definitions in the discussion paper?*

Yes.

> *If not, what definitions do you prefer and why?*

We do not recommend any changes.

(risk abatement)

(regulatory)

> *What potential risks from AI are not covered by Australia’s existing regulatory approaches?*

Key risks not well covered in existing regulatory approaches include consideration of the systemic impacts of AI on the economy and society, mechanisms to respond to emergent capabilities and associated risks, and insuffient explainability and accountability protocols as will be needed to apply to the unique challenges of AI.

Australia will need to develop an "elastic governance architecture" to assess these macro risks effectively. This will require a new approach to auditable solutions to address the fast changing impacts of new capabilities arising from generative AI, and new AI methods as they are developed.

Regulatory steps should include mandated impact assessments analysing societal risks, expanded regulatory powers to audit AI systems against well developed standards, and bans on harmful uses of emerging capabilities, as the paper rightly touches upon.

We recommend a balanced approach, one that builds on existing frameworks, while seeking to address the gaps as they become evident.

As noted in the paper, reforms targeting issues like transparency in automated decisions are already being proposed and we support these types of initiatives.

(non-regulatory)

> *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.*

Well implemented non-regulatory initiatives are likely to have as much impact, if not more, than regulatory measures. Rapid adoption of AI in Australia will be vital to the nation's future economic prosperity and social wellbeing. However, this will not happen without a coordinated approach by Governments at all levels. The following options should be canvassed:

- Establish an independent advisory council on AI ethics and governance. This should bring together experts from industry, academia, government and the community, to provide impartial advice on AI policy issues. The field of AI is rapidly evolving and will require ongoing input to address previously unknown issues.

- Develop voluntary AI certification schemes with industry.

- Establish open and transparent AI testing facilities.

  These testing sandboxes would allow companies to trial AI safely before deployment. Facilities could be public or privately run with government support. In order to ensure market competition small companies should be well supported to access such facilities.

- Support community participation programs on AI.

  Ongoing, public and open initiatives to include anyone with an interest in AI to participate in the development of programs and regulations concerning the development and use of AI. Community participation is critical to ensure engagement and buy-in from a broad range of diverse stakeholder groups

  Ongoing, public and open initiatives to include anyone with an interest in AI to participate in the development of programs and regulations concerning the development and use of AI. This is very important because we believe that the huge impact of AI on the lives of many people in the community will demand this.

(government coordination)

> *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.*

No.

(international examples)

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

We are not aware of measures taken by other countries beyond the comprehensive list included in the discussion paper, other than anecdotally. The measures discussed seem appropriate in the Australian context as well. Australia is not a nation at the forefront of new AI model development at this time and some of the measures relating to this may not be relevant at the moment. However, it would be prudent to avoid the assumption that this will always be the case.

(public vs private)

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

In a word, yes, but the approaches for public and private sector use should not differ by much.

Where there are differences they would be because:

- The public sector has public accountability obligations in ways the private sector does not. Moreover,public sector AI use, especially in law enforcement, likely has a greater potential impact on fundamental individual rights than many private sector activities. This potential impact has been highlighted by the recent Royal Commision into failures of automated decision making deployed by Centrelink.
- The private sector is driven by competition and needs flexibility to innovate. It encompasses a diversity of contexts, requiring more tailored governance. As discussed below, our view is that a risk based approach appropriately balances legislative obligations with the need for flexibility across different types of organizations.

We do not have specific recommendations on this point and in general argue for differences beyond these two motivations to be minimal. Core principles of transparency, accountability, and contestability should apply to all uses of AI.

(in government)

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

(generic vs technology)

> *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.*

We agree that AI can be combined with other technologies in innovative products and services and that the complexity arising from this may necessitate context-specific regulatory responses.

(transparency)

> *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.*

We agree with the suggestions in the paper, especially the importance of transparency in data sourcing and testing, mandatory incident reporting, and impact assessments. We also support full transparency to end users and that model development processes align with human values.

(banning)

> *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?*

There are very few AI applications or technologies we know of that we believe should be banned completely. However, the most powerful of the new AI models should be carefully monitored and regulated as we have previously outlined and we agree with most of the banning cases discussed in the paper for various forms of AI.

For example, we support bans on biometric or other identification technologies except in well defined use-cases. In those we recommend regulations that ensure transparency with respect to the uses for which such data are collected.

We are less enthusiastic about bans on ChatGPT or similar tools in schools and universities. Such actions are a "head in the sand" response to the reality that such tools now exist and the education system needs to adapt to them, not the other way around.

We support the approach adopted in the proposed EU AI Act's ban on certain uses of AI that pose an 'unacceptable risk' like social scoring and the development of criteria to ban AI where impacts are irreversible or perpetual.

(trust)

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

This is a fundamentally important role for Government because of the disruptive impact AI will play in many people's lives. Those who avoid it will, in many cases, become economically and/or socially disadvantaged as a result. Without such initiatives, AI may engender unwarranted fear or mistrust in the public and its response to it. It is therefore incumbent on Government to encourage the safe and responsible adoption and use of AI by everyone.

We agree with the paper's advocacy of voluntary ethical principles and codes of conduct, formal public education campaigns, transparency mechanisms and conformity infrastructure and assurance processes and an emphasis on safety by design.

The key point here is to ensure Government is seen as a backstop for anyone adversely affected by AI, e.g. via loss of their job or an entire industry. AI will fundamentally change the nature of employment, and Government’s role is to ensure that all parts of the community benefit from the AI dividend, rather than that dividend accruing to a small group. We see a powerful analogy in the Government's successful response by way of JobKeeper etc during the Covid Pandemic. We believe the impact of AI will for many people be similar, even if the speed of its impact a little more gradual.

(trade)

> *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?*

The potential impacts of Australia banning certain high-risk AI applications will likely be minimal in the near term, given the current makeup of the nation's tech sector and economy. AI technology exports are not yet a major contributor to economic activity. As such, prohibiting narrowly defined applications like facial recognition and social scoring would be unlikely to significantly restrict the sector's overall performance.

However, if these technologies represent emerging opportunities, inconsistent policies compared to trading partners could incrementally disadvantage Australian tech companies in the long run. The risks would be higher if banned applications eventually make up a sizeable portion of AI-related exports but we believe the impacts will likely be modest if restrictions remain confined to demonstrably high-risk uses, rather than blanket bans.

(conformity infrastructure)

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

We is comes to conformity infrastructure for AI, Australia appears to have very litte. Developing this to include AI assurance will likely require considerable effort and investment, referencing approaches in other advanced economies like the US, UK and EU.There is and will continue to be a shortage of expertise globally in auditing and evaluating emerging AI technologies against standards of fairness, safety and accountability for quite some time. Our organisation, which has specialists in the field, is small and there are few others like us. Australia needs to fund capability building, including through tertiary and research institutions, to grow this skilled talent pool locally.

Leveraging work by international bodies to harmonise standards and conformity assessment protocols is also important to help establish unified benchmarks and mutually recognised standards. Interoperability would support trade and technological development in our view.

As noted in the Discussion Paper, work done by the International Standards Organisation and the National Institute for Standards and Technology on identifying and managing AI Risk should form the basis of domestic standards, in accordance with the Department of Industry’s Best Practice Guide to using standards and risk assessments in policy and regulation.

Deployment and use of these frameworks have two important benefits. First, they align with approaches adopted in the US and EU - large economic blocks with whom Australia has strong trade relations. Second, these standards, and the accompanying guidance on implementation provide a methodical approach to identifying, monitoring and responding to risks of the lifecycle of an AI system.

The Australian Assurance and Conformity Infrastructure should prioritise consideration of these to determine if it can be applied or adapted to an Australian setting.

We are not of the view that standards should be legislated. Rather the current risk-based approach - i.e. a requirement that AI users implement a risk management program - used in legislative regimes such as Corporations Act 2001 (Cth) 912A and Security of Critical Infrastructure Act 2022 (Cth) Part 2A is preferable. We note that this is the approach taken by both the EU's proposed AI Act and the Canadian AI and Data Act (Bill C-27).

Lastly, we note that while there are existing risk management obligations in a range of legislative regimes, these are not applicable to all organisations across the economy. Accordingly, these gaps need to be filled by specific legislation.

(risk based approach)

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

Yes, as discussed above, we support a risk based approach as advocated in the paper.

We are not aware of other approaches that have proven more effective.

However, the challenge will be in identifying and acting to mitigate future risks, before they appear. We see as unlikely, but not impossible, some of the concerns raised by others of the potential for the most powerful AI to achieve "human intelligence" without alignment to "human values". This is a risk we believe, that if manifest, would prove difficult to mitigate after the fact. Moreover, the dangers posed by these risks materialising are of a magnitude that the AI community must prioritise serious consideration of mitigation strategies.

We therefore advocate a robust risk-based framework be combined with precautionary measures against plausible long-term scenarios such as this. This needs to be undertaken with evidence-based oversight of existing AI applications to inform continued research and monitoring of AI capabilities. This will help guide appropriate safeguards against these future risks. International coordination in early detection and response should be actively pursued, especially given the rapid global pace of advancement in the most powerful AI models.

(benefits and limitations)

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

We agree with the general assessment of the risk-based approach.

The prime benefits are proportionality balancing innovation with risk abatement, flexibility being able to adapt the response to context, a focus on high risk activities as previously discussed and the development of a risk management mindset in both the public and private sectors.

The limitations are primarily focused on how risk is defined and assessed. It can be highly politically charged in some situations and risk profiles can change faster than regulatory responses can keep up, especially in the field of AI where development of new technologies is increasing at an exponential rate.

(sector based)

> *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. Absolutely, but only in those sectors that have an outsized impact on the health, financial and physical wellbeing of individuals.

We can only speculate on how the size and maturity of organizations using AI could factor into how suited they are for risk-based requirements; larger entities with more resources willbe better equipped to thoroughly assess risks, while smaller groups will find this more burdensome. And those new to AI may initially struggle with risk analysis without experience.

However, we caution that overly differential treatment could lead to imbalanced policy and there is an argument that comprehensive risk-based requirements across all sectors and organisation sizes would encourage widespread capabilities and expectations for responsible AI.

(element to include)

> *What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C? (page 40 of the discussion paper)*

In our view the components listed in the paper should all be included in a risk-based AI governance approach. We would further advocate:

- External and internal auditing requirements to provide important verification of risk controls and documentation, especially for higher risk AI systems. Mandatory independent audits improve accountability but need to be balanced against the resource cost associated that such audits would entail .

- Ongoing monitoring of AI systems by users and developers
- Requiring minimum general liability coverage for organisations deploying high or very high-risk AI seems prudent to ensure resources are available if harms materialize that require redress.
- Mandatory external certification for the highest risk applications could also improve safety and reliability assurance through independent evaluation by accredited bodies.
- Access to large, high-fidelity data sets is critical for the safe and effective deployment of machine learning systems, especially for sensitive applications like healthcare and precision medicine. Data quality frameworks would help manage risks from inaccurate or biased training data. Governments at all levels can shape industry practice in this respect through the operation of public data sharing regimes such as those instituted under the *Data Availability and Transparency Act 2022 (Cth) and the Data Sharing (Government Sector) Act 2015 (NSW).*

(incorporation)

> *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?*

AI is quite unique in many ways. Care should be taken not to dilute focus on AI-specific issues. However, developing AI-specific guidelines or criteria that can plug into prevailing risk management methodologies like ISO 31000 and CPS 230 makes sense. For example, Box 1 has some relevant proposals from the Privacy Act Review report that relate to transparency and explainability for automated decision making that impacts individuals with which we agree.

(lls and mfm)

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

These AI technologies are were the greatest risks mostly likely apply. Risk assessments will depend on the nature of the model. Supervised learning models, unsupervised learning models, reinforcement learning models, or generative models will each have different risks. For example, reinforcement learning systems can have issues related to reward hacking where the model finds unintended ways to maximize its reward function. There are many other modes of risk and they are being discovered daily. For example, https://llm-attacks.org which presents a paper that proposes an advanced method of creating adversarial attacks on large language models by generating specific suffixes that increase the likelihood of objectionable content production, highlighting the method's transferability across different models and raising important questions about the prevention of such undesirable outcomes.

(voluntary or not)

> *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?*

It should be both in all three cases. We have a vision statement that covers these questions and which we think may help inform your review.

[2023-07-30 Sun 19:59] (submission) (hmc7s3ekwko2jpav) (© 2023 human compatible) (✍️ bruce tulloch)