Good Ancestors Policy

Automated Decision-Making

Submission to

Attorney-General's Department

Automated Decision-Making Consultation

January 2025

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity. We care about today's Australians and future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world are facing. We want to help by making forward-looking policy recommendations that are rigorous, evidence-based, practical, and impactful.

Good Ancestors has been engaged in the AI policy conversation since our creation, working with experts in Australia and around the world while connecting directly with the Australian community.

Good Ancestors is proud to help coordinate Australians for AI Safety.

Our thanks go to the volunteers who provided input to this submission and who care so passionately about being good ancestors to future generations of Australians.

Technical support for oversight of ADM

Good Ancestors appreciates the opportunity to comment on Government's use of automated decision-making (ADM). While ADM is long-standing, the rise of generative AI paired with its rapidly increasing capability will drive a paradigm shift in the nature of ADM and its salience to government and society.

For an ADM framework to succeed, it needs to ensure public servants are supported by the necessary technical tools, and that the capabilities of those tools grow alongside the capability of AI used in ADM.

This submission primarily focuses on technical issues that arise in the discussion paper's questions, including as they relate to transparency, safeguards and recommendation 17.2 of the Robodebt Royal Commission.

Recommendations:

1. **Australia's ADM Framework should acknowledge that "human in the loop" will be less relevant to safeguards, oversight and accountability as AI capability increases.**
2. **Australia's ADM Framework should require the adoption of scalable oversight tools alongside AI-assisted ADM and ensure that the capability of the oversight tools keeps pace with the capability of AI-assisted ADM. Government has a legitimate role in market-shaping in this context.**
3. **Australia should create an AI Safety Institute to house technical expertise in the effective oversight of highly capable AI, including expertise on scalable oversight for ADM. This could also deliver Australia's commitments under Recommendation 17.2 of the Robodebt Royal Commission and the Seoul Declaration on AI Safety.**

An ADM framework must account for advancing AI capability

Any enduring ADM framework needs to grapple with rapid advances in AI capabilities. Government is generally under pressure to accelerate its adoption of AI. For instance, the Australian Government trial of Microsoft 365 Copilot recommended greater adoption.[1] ADM will be subject to the same pressures to use AI more and to keep up with new AI capabilities.

An ADM framework must specify, not only how oversight, accountability and transparency techniques will work, but how they will "keep up" with growing AI capability and pervasiveness.

If Government proposes an ADM framework that makes sense for today's systems but underestimates near-term trends in capability, the framework will fail under imminent, rapid change.

AI capability will disrupt historical approaches to oversight

A useful, if limited, metaphor for understanding the growing capabilities of AI models is to compare their benchmark results to typical academic performance.

Year	Model generation	Rough performance on cognitive tasks
June 2018	GPT-1	Toddler
February 2019	GPT-2	Primary school student
June 2020	GPT-3	High school student
March 2023	GPT-4	Senior school student
September 2024	GPT-o1	University student[2]
Early 2025(?)	GPT-o3	PhD candidate
2026(?)	?	Professor?
2027(?)	?	Nobel laureate?

[1] **Digital Transformation Agency**, 'Evaluation of the Whole-of-Government Trial of Microsoft 365 Copilot: Summary of Evaluation Findings' (23 October 2024). https://www.digital.gov.au/initiatives/copilot-trial/summary-evaluation-findings.

[2] OpenAI describes o1 saying, "In our tests, the next model update [o1] performs similarly to PhD students on challenging benchmark tasks in physics, chemistry, and biology. We also found that it excels in math and coding." **OpenAI**, 'Introducing OpenAI o1-preview' (12 September, 2024). https://openai.com/index/introducing-openai-o1-preview/.

While this is rough, and even narrow AI models vastly outperform all humans on many tasks, it does illustrate that any framework for tackling issues relating to openness, fairness, participation, accountability and consistency needs to have in mind AI systems that can not only ingest more data than a human and work faster than a human - but have cognitive abilities that match or exceed many, most or all humans.

This has profound implications for oversight and transparency**. A framework designed for a teacher overseeing a student is fundamentally different from one designed for a student overseeing a teacher.** If Government proposes an ADM framework that assumes humans can readily assess the outputs of an AI model, the framework will rapidly become outdated.

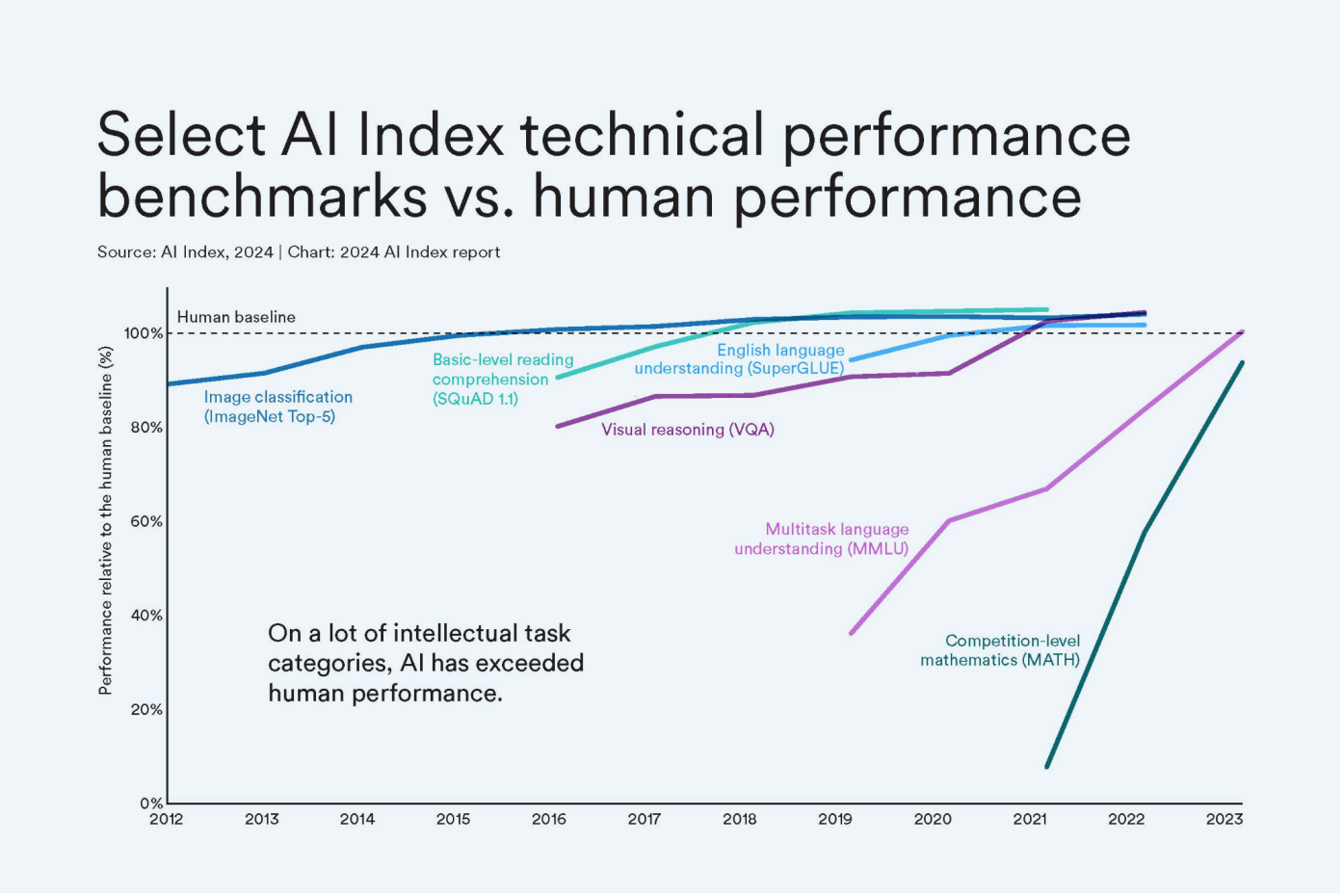

Figure 1: AI has surpassed human-level performance at a number of tasks, and the rate at which humans are being surpassed at new tasks is increasing.[3] We should expect this trend to continue.

[3] **Shana Lynch**, 'AI Index: State of AI in 13 Charts' (Stanford University, 15 April 2024) https://hai.stanford.edu/news/ai-index-state-ai-13-charts.

AI capability growth is real, rapid, and right now

AI capability growth is driven by exponential growth in investment, compute, training data and algorithmic efficiency (including recent innovations regarding reasoning models).[4]

The "**training compute**" going into AI models is increasing rapidly – a factor of 10 billion since 2010.[5] The "doubling time" of compute is accelerating. Between 1951-2010, AI compute doubled every 18 months. Between 2010-2022, it doubled every six months.

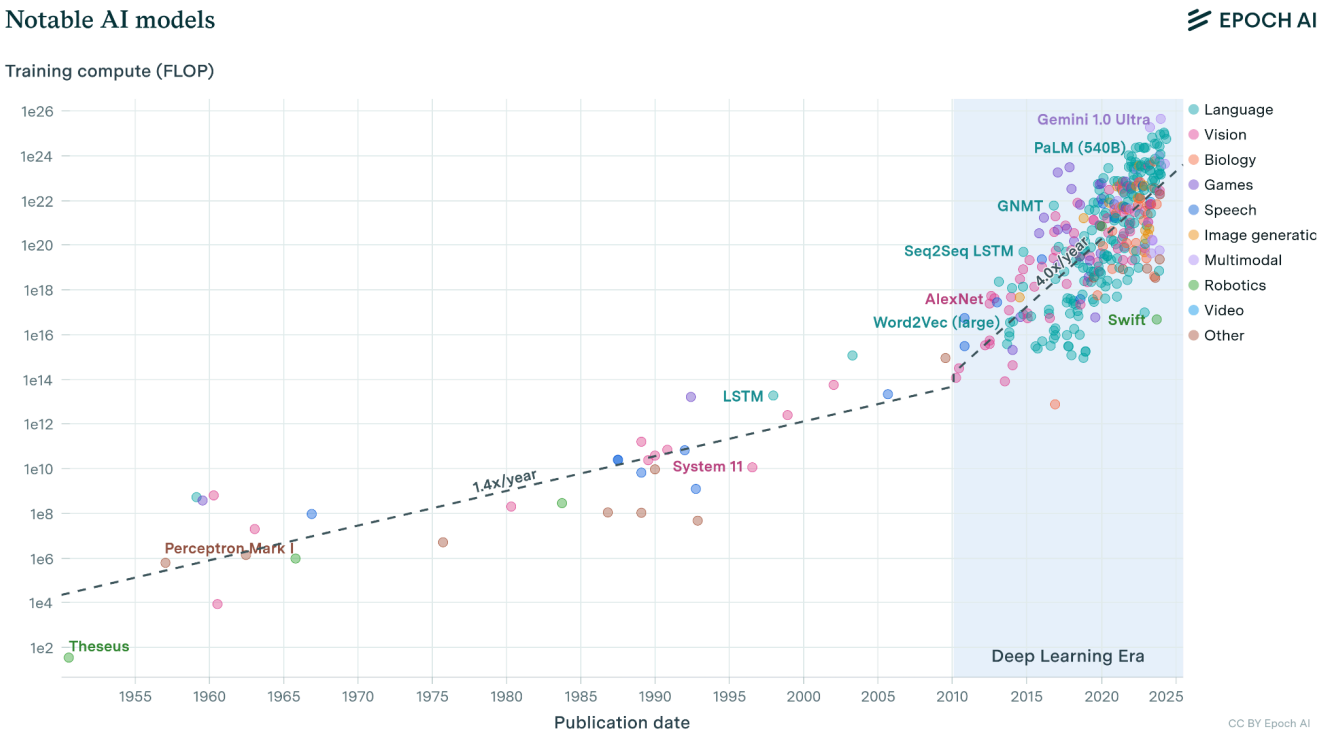

Figure 2 illustrates the recent step-change in computing power being used to train AI systems.[6]

Training datasets are growing rapidly. The best estimate is 2.2x growth per year since 2010. This trend is so rapid that it might mean that by 2040, leading labs will use the totality of public text and images to train their models.[7] Even this might not represent a limit on growth. LLMs can be trained on the outputs of other AI models,

---

[4] **Daniel Murfet and Greg Sadler,** 'Using Open-Source AI, Sophisticated Cyber Ops Will Proliferate' (17 December 2024) *The Strategist* https://www.aspistrategist.org.au/using-open-source-ai-sophisticated-cyber-ops-will-proliferate/.

[5] **J Sevilla et al,** 'Compute Trends Across Three Eras of Machine Learning' (Paper presented at the *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022) 1–8, doi: 10.1109/IJCNN55064.2022.9891914, https://epoch.ai/blog/compute-trends.

[6] Ibid

[7] **Pablo Villalobos et al,** 'Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data' (Preprint, arXiv, 4 June 2024) https://doi.org/10.48550/arXiv.2211.04325.

known as "synthetic data."[8] Synthetic data is already being used by leading labs. OpenAI's o1 model seems to use synthetic data in its 'chain of thought',[9] and Anthropic's constitutional AI has involved synthetic data for training Claude.[10]

AI algorithms are also improving. Each year, both large language models and computer vision models require only a third as much compute to achieve a given performance level.[11] The introduction of "interference time" to facilitate advanced reasoning, as demonstrated by OpenAI's o1 and o3 Models as well as Chinese competitors, has further boosted capability and unlocked the ability of models to utilise more computing power to perform better in complex tasks.[12]

These trends are **self-reinforcing**. AI is already being used to help design the chips that run AI and to help code the next generation of AI. As AI companies demonstrate that customers are willing to pay for their products, it becomes easier for them to raise funds to invest in more powerful models.

AI capability progress has several drivers – each source alone may be sufficient to drive exponential growth. Policymakers learned first-hand during the COVID-19 pandemic that exponentials can quickly turn something from a possibility to something impacting everyone we know. Governments that took expert advice earlier were better off. We should be aware of the likelihood of this happening with AI.

Finding: Given the track record of increasing AI capability and its diverse drivers, AI is very likely to continue advancing for the foreseeable future.

"Human intervention" in ADM will not be able to keep up

A key implication of this rapid growth in AI capability is that "human in the loop" – referred to as "human intervention" on page 25 of the discussion paper – cannot be an enduring solution to issues relating to transparency or the provision of effective safeguards. In the same way engineers have specific tools to assure the

---

[8] **Jiaxin Huang** et al, 'Large Language Models Can Self-Improve' (Conference Paper, ICLR 2023, 2 February 2023) https://openreview.net/forum?id=NiEtU7blzN.

[9] Trapit Bansal: "When training a model for reasoning, one thing that immediately jumps to mind is to have humans write out their thought process and train on that. When we saw that if you train the model using RL to generate and hone its own chain of thoughts it can do even better than having humans write chains of thought for it. That was the "Aha!" moment that you could really scale this." **OpenAI**, 'Building OpenAI o1 (Extended Cut)' (YouTube, 21 September 2024, 3:10). https://www.youtube.com/watch?v=tEzs3VHyBDM.

[10] **Yuntao Bai** et al, 'Constitutional AI: Harmlessness from AI Feedback' (Anthropic, 15 December 2022) https://arxiv.org/abs/2212.08073.

[11] **Anson Ho** et al**,** 'Algorithmic Progress in Language Models' (Preprint, arXiv, 9 March 2024) https://doi.org/10.48550/arXiv.2403.05812.

[12] **Zijian Yang,** 'How OpenAI's O1 Series Stands Out: Redefining AI Reasoning' (20 September 2024) *Medium* https://medium.com/@researchgraph/how-openais-o1-series-stands-out-redefining-ai-reasoning-9e499937139e.

quality of a manufactured part, decision-makers need specific tools to oversee AI. Unaided human skill is not sufficient for either task.

An example from chess engines

While AIs that play chess or other games are "narrow AI systems" and relevantly different from the increasingly "general AI" systems, they can provide valuable insight into how humans and advanced AI can interface. Specifically, they can help us think about the extent to which a human can practically oversee or assure the recommendations of an advanced AI system.

Imagine it was your job to be the "human in the loop" with a chess bot like Stockfish 17 – providing approval, oversight and accountability for its decisions. Stockfish recommends a chess move in seconds - using a neural network combined with a 32TB table of chess positions.

If Stockfish recommended a move, often you would be able to quickly confirm that the recommendation was reasonable, and you may be able to provide a general explanation of why the move is reasonable. However, sometimes, a move would seem counterintuitive and you might want to investigate further. In these cases, you could perform hours or days of analysis and perhaps consult world-leading experts to perform their own analysis. Despite that work, it would seemingly always turn out that the AI-recommended move was better than your alternative.

Faced with this task, even the most diligent "human in the loop" would rapidly lose heart, leading to "de-skilling" of the kind referred to on page 10 of the consultation paper. Importantly, "de-skilling" doesn't necessarily mean having less skill, rather it means being in a working context where deploying that skill is impractical. For instance, if Stockfish did make occasional errors, these human factors would lead the demoralised 'human in the loop' to question their own analysis and defer to the AI.

A task like making administrative decisions based on defined decision criteria and a fixed data set (like the content of an application) is the very kind of task that we should expect AI to excel at today or in the near future. A human tasked with overseeing such an AI would be in a similar position to a human tasked with overseeing Stockfish today.

The key lesson is that any ADM framework that requires an unaided human to oversee a highly capable AI will fail.

The challenges with "human in the loop" are most apparent in high-volume contexts. In these contexts, an AI can ingest and analyse data far faster than a human. A simple example would be summarising lengthy applications or attached documents: it may take today's AI moments to summarise hundreds of pages and generate conclusions and actions. Checking this could take a human hours or days, and they will probably make errors themselves. A human charged with overseeing such a system will rapidly be overwhelmed and will be forced to make practical decisions about the extent to which they "trust" the AI on foundational questions, like which data it highlighted as being decision-relevant. As AI becomes more capable and matches or exceeds the reasoning capabilities of human experts, the "human in the loop" task might be impossible.

Even if not "de-skilled", human decision-makers will fall behind

The risks of AI-ADM de-skilling human decision-makers, discussed on page 10 and in the ARC's *Automated Assistance in Administrative Decision* making, are real. However, as AI systems become more powerful, no amount of unaugmented human skill will allow humans to offer practical oversight. "More human skill", or particular approaches to deploying human skill, cannot be the solution.

Page 17 of the discussion paper, citing Burrell,[13] argues that transparency may become impractical for complex models and systems. Noting that the Burrell paper is from 2016, before the dramatic increase in generative AI performance, this problem has only become more acute. For instance, Stanford University reports that, as of 2023, AI models performed better at answering open-ended questions about images than humans.[14] Despite seeing and interpreting images being a core competency for humans, AI vision is better than human vision and we are not in a technical position to understand or communicate how or why.[15] Given that a human overseeing a basic AI vision model cannot practically provide transparency or explainability, any attempt at transparency or explainability for more complex tasks is unlikely to be possible or genuine.[16]

Given that the enduring solution to effective governance of ADM cannot be humans on their own staying ahead of AI, **technical solutions must be developed and implemented to ensure appropriate support for humans** charged with governing ADM.

---

[13] **Jenna Burrell**, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) *Big Data & Society* 1, doi:10.1177/2053951715622512.

[14] **Shana Lynch**, 'AI Benchmarks Hit Saturation' (Stanford University, 3 April 2023) https://hai.stanford.edu/news/ai-benchmarks-hit-saturation.

[15] "Sision" is an illustrative example of a human capability. We acknowledge that many individuals live with varying degrees of visual impairment or blindness. Referencing vision as a "core competency" is not intended to diminish or exclude these experiences, nor to suggest that vision is universal or essential for personhood.

[16] That is, we might be able to concoct a post-hoc rationalisation about why or how an AI made a decision, but it would not be robustly true. Government should not establish a framework that provides faux-transparency.

Finding: Unaided humans cannot provide effective oversight of increasingly capable AI models – including explainability or transparency.

Figure 3 is part of the Visual Question Answering dataset. These problems at the intersection of vision and language were once a serious challenge to leading AI systems. As of 2023, AI models perform better than humans on these tasks. AI performance exceeding human performance creates serious challenges for human oversight of AI systems and normal approaches to transparency.

ADM needs technical tools for monitoring and audit

The "AI safety"[17] community is developing tools to allow humans to provide effective oversight and control of increasingly powerful AI tools - typically called "scalable oversight". These tools allow humans to ensure increasingly powerful AI tools are aligned with their intent and assure their outputs – exactly the kinds of tools an ADM framework must incorporate. Critically, **scalable oversight development is currently not on track to keep up with AI capability development. Government intervention is essential to support scalable oversight – both by directly supporting research and by generating demand for implemented solutions.**

[17] AI safety is an interdisciplinary field focused on preventing accidents, misuse, or other harmful consequences arising from artificial intelligence (AI) systems. It encompasses machine ethics and AI alignment, which aim to ensure AI systems are moral and beneficial, as well as monitoring AI systems for risks and enhancing their reliability. The field is particularly concerned with existential risks posed by advanced AI models.

Scalable oversight explained

Scalable oversight refers to the systematic approach and methodologies required to monitor, evaluate, and enhance AI systems as they grow in complexity and capability, particularly when they exceed human performance.[18]

Most proposed mechanisms for scalable oversight, at their core, attempt to support humans overseeing AI by using AI *in* the oversight process.[19]

One avenue of scalable oversight is in decomposing or distilling the complex outputs of advanced AI. This can include:

- Task decomposition, or 'factored cognition' - having the model under evaluation (in this case, the model being used to support ADM) break down the output into smaller chunks for evaluation.[20] This may also involve training other specific models as assistants for this purpose.[21]
- Debate - having the model take on multiple roles to debate itself or using two separate models that have to test their analysis against one another.

These approaches typically still have a "human in the loop", assessing final outputs – but supporting those humans with continually improving tools would greatly extend the effectiveness and longevity of any ADM framework.

Scalable oversight would also help prepare for when "human in the loop" becomes infeasible. Researchers are exploring:

- Focusing on "objective" problems - like science (or more broadly STEM) questions, making predictions, or forecasting markets.
- Having another model take on the role of evaluator[22] - broad approaches include reinforcement learning from AI feedback and "constitutional AI" (being used today by Anthropic[23])

A theoretical criticism of this approach is that it may constitute an AI "marking its own homework". While this criticism does not present immediate practical problems for using scalable oversight techniques by Government in support of ADM, it is subject to ongoing research.

---

[18] **Deepak Babu P R**, 'Scalable Oversight in AI: Beyond Human Supervision' (Medium, 2 October 2023) https://medium.com/@prdeepak.babu/scalable-oversight-in-ai-beyond-human-supervision-d258b50dbf62.
[19] **Evan Hubinger,** 'An Overview of 11 Proposals for Building Safe Advanced AI' (4 December 2020) arXiv https://arxiv.org/abs/2012.07532.
[20] **Jeffrey Wu, Ryan Lowe, Jan Leike, Long Ouyang, Daniel Ziegler, Nisan Stiennon, and Paul Christiano**, 'Summarizing Books' (OpenAI, 2021) https://openai.com/index/summarizing-books/.
[21] **Leike, Jan**, 'Why I'm Excited About AI-Assisted Human Feedback' (29 March 2022) *Substack* https://aligned.substack.com/p/ai-assisted-human-feedback.
[22] **Collin Burns** et al, 'Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision' (OpenAI, 14 December 2023) https://arxiv.org/abs/2312.09390.
[23] **Yuntao Bai** et al (n 9).

Independent researchers and AI Lab Anthropic propose a "sandwich" method. The method intends to facilitate research today to develop scalable oversight mechanisms that will remain valid even after AI cognitive ability surpasses humans.[24] This method simulates a future where the model is more capable than its evaluators by having human *non-experts* evaluate a model. For instance, researchers would develop scalable oversight techniques that would allow a non-lawyer to effectively oversee the legal outputs of an AI model. Expert lawyers could judge the effectiveness of the oversight. If judged successful, those same techniques may remain valid even if the AI becomes more capable than the experts. This is referred to as "sandwiching" because the model is placed between the non-experts and the experts.

Scalable oversight as a field has identified a number of promising avenues for further progress. What is needed now is technical effort to find which approaches are most tractable and industry effort to turn these approaches into specific products. In other domains, experts use specialist tools for quality assurance. We don't expect engineers to assess the tolerance of a part by looking at it. Instead, they use tools like check gauges to help their assessment. An ADM framework needs to build the "check gauges" of AI-aided decision-making.

Market stewardship is core business for the Australian Government.[25] In the context of AI, many market forces are driving AI capability, but oversight and control are falling behind. Government should use its ADM framework to expressly shape the market for AI oversight. This should include an overt statement that adoption of ADM tools must be accompanied by the adoption of scalable oversight tools. This kind of conscious market shaping is essential to success.

Finding: Any ADM framework must include the implementation of scalable oversight tools alongside the use of AI. Government's ADM framework could engage in "market shaping" by signalling Government's intent to procure scalable oversight tools alongside ADM tools.

Australia needs an AI Safety Institute

The global network of AI Safety Institutes (AISIs) will be essential to developing AI safety tools and techniques, including scalable oversight. Australia contributing to

[24] **Samuel R Bowman** et al, 'Measuring Progress on Scalable Oversight for Large Language Models' (Anthropic, Surge AI, and independent researchers, 11 November 2022) https://arxiv.org/abs/2211.03540

[25] **Department of Prime Minister and Cabinet,** *Draft National Care and Support Economy Strategy 2023* (2023) 46, Objective 3.3, https://www.pmc.gov.au/resources/draft-national-strategy-care-and-support-economy.

this network of AISIs, including driving the development of scalable oversight, should be a core component of any ADM framework.

In May 2024 Australia, alongside ten other countries and the European Union, signed the Seoul Declaration.[26] The Declaration confirmed the nations' shared understanding of the opportunities and risks of AI and committed signatories to "create or expand AI safety institutes" alongside other forms of international cooperation on safe and responsible AI.

Creating an Australian AI Safety Institute would also discharge Australia's commitments under the Hiroshima Process[27], the Bletchley Declaration[28], and Australia's own AI ethical principles. Relevantly, Australia's AI ethics principle of **transparency and explainability** asks that users and third parties be able to understand their interactions with AI, which requires us to have a sufficient understanding of how increasingly advanced AI systems work.

The creation of an Australian AISI is also a natural moment to deliver Recommendation 17.2 of the Robodebt Royal Commission as it relates to the use of AI in ADM. Recommendation 17.2 highlights a technical capability gap, and calls for the creation or expansion of a body that has suitable technical expertise in the functioning of increasingly advanced AI systems.

Given the importance of scalable oversight and interpretability to ADM specifically and AI safety in general, the inclusion of this function within an AI safety institute, in support of a relevant auditor, is a natural way to proceed. Creating and empowering an Australian AISI is also essential to delivering the proposed mandatory guardrails, discussed on page 22 of the discussion paper, relating to an obligation to "Test AI models and systems to evaluate model performance and monitor the system once deployed" and "Keep and maintain records to allow third parties to assess compliance with guardrails".

---

[26] **Department of Industry, Science and Resources,** *The Seoul Declaration by Countries Attending the AI Seoul Summit, 21–22 May 2024* (24 May 2024) https://www.industry.gov.au/publications/seoul-declaration-countries-attending-ai-seoul-summit-21-22-may-2024#seoul-declaration-1.

[27] **Department of Industry, Science and Resources,** 'Australia Joins Hiroshima AI Process Friends Group' (3 May 2024) https://www.industry.gov.au/news/australia-joins-hiroshima-ai-process-friends-group.

[28] **Husic, E.** (2023, November 3). Australia signs the Bletchley Declaration at AI Safety Summit [Press release]. Minister for Industry and Science. Retrieved from https://www.minister.industry.gov.au/ministers/husic/media-releases/australia-signs-bletchley-declaration-ai-safety-summit

Conclusion and recommendations

Overall, the involvement of AI in ADM promises substantial benefits, like efficiency, speed and accuracy. The budgetary benefits of making better, faster decisions for less money are likely to be irresistible across Government and industry.

"Human in the loop" is an enticing policy solution to challenges like safeguards, oversight and accountability. "Human in the loop" seems intuitive because it's how the system works today – more senior and experienced humans provide oversight and expertise to assist more junior staff doing frontline tasks. Simply substituting AI for junior workers feels straightforward.

However, a "human in the loop" with increasingly powerful AI models faces practical problems. Frontier AI already matches or exceeds human capabilities on most benchmarks and dramatically exceeds human capability in key areas (like the speed of ingesting information). AI capability is growing rapidly. An unassisted human charged with providing oversight to AI will not only be ineffective at the task, but will rapidly "de-skill" in the face of the impossible job.

Fortunately, the field of AI safety – specifically scalable oversight – offers solutions to these problems. Urgent investment is needed to grow domestic capability and ensure the expertise is available to meet demand.

Recommendations:

1. **Australia's ADM Framework should acknowledge that "human in the loop" will be less relevant to safeguards, oversight and accountability as AI capability increases.**
2. **Australia's ADM Framework should require the adoption of scalable oversight tools alongside AI-assisted ADM and ensure that the capability of the oversight tools keeps pace with the capability of AI-assisted ADM. Government has a legitimate role in market-shaping in this context.**
3. **Australia should create an AI Safety Institute to house technical expertise in the effective oversight of highly capable AI, including expertise on scalable oversight for ADM. This could also deliver Australia's commitments under Recommendation 17.2 of the Robodebt Royal Commission and the Seoul Declaration on AI Safety.**