SUBMISSION

Submission to the Department of Industry, Science and Resources

Submission to the Supporting Safe and Responsible AI Practices Discussion Paper consultation

26 July 2023

The Australian Academy of Technological Sciences and Engineering (ATSE) is a Learned Academy of independent, non-political experts helping Australians understand and use technology to solve complex problems. Bringing together Australia's leading thinkers in applied science, technology and engineering, ATSE provides impartial, practical and evidence-based advice on how to achieve sustainable solutions and advance prosperity.

The accelerating development of Artificial Intelligence (AI) brings with it the possibility of great change in human society, impacting how we live and work, improving productivity, supporting learning and offering new entertainment possibilities. ATSE's report for the National Science and Technology Council on generative AI illuminates the risks and opportunities for Australia, as well as challenges such as balancing progress with human rights, data security and environmental impacts. The potential impact of this rapidly developing branch of technological advancement necessitates thoughtful and responsive oversight and regulation, and ATSE commends the Australian Government for commencing with the Safe and Responsible AI Practices Discussion Paper. ATSE agrees that the level of regulation applied to AI systems should be proportionate to the risk that misuse or malfunction of the technology may have to individuals, industry and society. Many companies, both in Australia and internationally, are moving ahead of the Government by developing their own ethical AI policies, principles, standards and strategies. Independent oversight and guidance are important to ensure that AI systems are used appropriately and that bad actors can be identified and prevented. The regulatory environment must remain flexible enough to adapt to rapid changes in technology, supporting the sector to grow and develop, while minimising the risk of potential negative impacts.

ATSE commends the tiered model proposed in the discussion paper as fundamentally the right approach to regulating AI in Australia. ATSE's recommendations in this submission focus on changes designed to support regulation that meets its intended aims, and identify factors to be considered if this framework is translated into law. The nature of data collected and analysed by AI models should also be more fully incorporated into the risk matrix to ensure appropriate use and storage of private or sensitive data. Regulation should address AI systems under development and those that store data offshore, and protect vulnerable people.

To this end, ATSE makes the following recommendations:

Recommendation 1: Ensure Australia's AI regulatory environment is consistent with the OECD AI principles and can adapt to future AI technology development.

Recommendation 2: Include AI systems under development within the graded regulatory requirements based on the potential impacts of the AI system.

Recommendation 3: Assess the risk categorisation of AI systems under the proposed regulatory framework against real-world outcomes of the system, as evaluated by ethical AI experts and representatives of those impacted by the AI system.

Recommendation 4: Require reporting of aggregate data for medium risk AI systems to ensure these systems do not perpetuate or exacerbate systemic disadvantage.

Recommendation 5: Amend the definition of the proposed risk categories to incorporate the risk posed by data inputs to AI systems, in addition to the current categorisation based on outputs and impact.

Recommendation 6: Require AI systems operating in Australia to meet Australian regulatory requirements for privacy and data security regardless of the location of their servers.

Regulating the future of AI

The regulation of AI must be designed to have longevity, and be as flexible and future-proof as possible. Regulations must adapt as AI technology develops, without unintentionally impeding new technologies. At the same time, regulations must provide a strong framework to ensure technology meets community expectations. The OECD has outlined that AI development should be based on five main principles – benefitting people and planet, human-centred values and fairness, transparency and explainability, accountability, and robustness, security and safety (OECD 2023). These principles align with the Department of Industry, Science and Resource's (DISR) own AI ethics principles, the existing voluntary framework for ethical AI practices (DISR n.d.). DISR's principles are based in the IEEE's ethically aligned design principles (IEEE 2017). These principles lay a strong foundation for any regulatory framework adopted by the Australian Government. As the detail of the regulatory framework is determined, these principles should form the basis of decisions made by the government and should lay the foundation for what AI developers, users and the public can expect from AI systems in Australia.

The OECD also highlights the importance of investment in AI development, developing skills and international cooperation as key to a successful transition to an AI future (OECD n.d.). ATSE supports these policy principles and encourages the Australian Government to invest in Australian-led AI and the skills necessary to make that happen. Local underinvestment in AI research will impair the nation's ability to not

Level 2, 28 National Circuit
Forrest ACT 2603
Australia

PO Box 4776
Kingston ACT 2604
Australia

+61 2 6185 3240
info@atse.org.au
atse.org.au

ABN 58 008 520 394
ACN 008 520 394

Australian Academy of Technological Sciences & Engineering

only create AI systems but will also impair Australia's ability to effectively regulate AI systems developed overseas. As ATSE's vision statement on strategic investment in AI highlights, investment in fundamental research in AI in Australia is vital to the nation's successful adoption of AI technologies. AI systems used in Australia will need to be based on Australian data and Australian models to ensure that these systems are responsible and meet Australian needs, and this will require support and investment.

The National Science and Technology Council report on generative AI identifies six major steps in AI development, only two of which occur at or after the point an AI system is released to the public (Bell et al. 2023). It is therefore necessary that ethics and regulation in AI must not only be considered at the point where products are ready for the consumer market - these principles must also be at play during the design and development of AI products. Australia's regulatory framework must therefore include a role for regulation and oversight during the development phase of AI systems. This includes the development of Australian models based on Australian data, and developers should engage with groups likely to be affected by the AI system during the design and development process to ensure the models align with our national ethical framework.

These measures must not be overly onerous, to avoid stifling research and development, but should ensure that training data, testing of AI systems, and the approach to, management and use of, and storage of data collected during those tests are managed in a manner that supports privacy, maintains human oversight, and identifies and deals with as many as possible potential impacts beyond test environments. While ongoing refinement based on real-world use strengthens AI systems, products should not be allowed to be tested where there could reasonably be anticipated likelihood of large or irreversible impacts on people or critical systems, including government services or decisions. Oversight of systems under development should follow a similar tiered approach as proposed for the public version of AI models themselves, with requirements linked to the potential impacts of the AI system under development.

Recommendation 1: Ensure Australia's AI regulatory environment is consistent with the OECD AI policy principles and IEEE ethically aligned design principles and can adapt to future AI technology development.

Recommendation 2: Include AI systems under development within the graded regulatory requirements based on the potential impacts of the AI system.

Assessing AI impact against predetermined and defined standards

The purpose for which an AI system is used, and the AI system's efficacy, should determine the regulatory framework that governs its use and development. One of the strengths of the proposed model is the delineation of three categories of regulation, scaling up the regulatory burden as the impact of the AI increases. The intended impact and the actual impact may, however, not always align and this must be recognised in the regulatory process. For example, an AI designed to deliver the most relevant advertisements online may also, intentionally or otherwise, help develop and perpetuate addiction by repeatedly showing individuals with addictive tendencies gambling advertisements. Assessments of AI impact must not be based on developer or user intentions but instead on the actual, real-world, interactions with and outcomes of these systems, with principles of social justice in mind.

AI tools present an opportunity to help break down a range of social inequalities and reduce bias, but only with the right safeguards in place (OECD AI Policy Observatory 2022). In designing safeguards, it must be recognised that the impacts of AI are not always evenly felt. A company using an AI system to assist with recruiting may select an inappropriate employee based on an erroneous AI output, causing inconvenience and small additional costs to the company. For an individual who routinely has their job applications rejected by erroneous or biased AI systems, the costs borne are much greater, losing out on vital income, opportunities and connections, which may also impact mental health and wellbeing. This kind of bias is not simply theoretical – evidence of racial bias has already been found in AI recruiting software (Zapata, 2021). Under the proposed framework, this kind of AI system would be classified as "medium risk", requiring self-assessment and explanations of decisions to be provided to individuals involved, even though an erroneous system presents a high risk to the impacted individual. Risk classification assessment must include those affected by the outputs of AI system, as well as experts in ethical AI practices (that is, those with expertise in both AI systems and ethics).

Many of the AI systems designated as "medium risk", like the recruiting software example above, have the potential to perpetuate or exacerbate pre-existing systemic disadvantage if improperly designed or trained on biased datasets (including datasets that simply reflect current systemic disadvantage). These systemic impacts may not be obvious where reporting is based on individual explanations of outcomes, particularly where these reports are generated by the AI system itself which users must trust are an accurate reflection

of a process that occurs in a 'black box' that makes it impossible to verify the rationale behind outputs. It is therefore necessary that these systems are subject to requirements to continually report outcomes across all outputs, so that biases or systemic errors in outputs can be properly identified and rectified. Producing usage-wide reports will also help individuals who wish to contest decisions to develop a stronger evidentiary basis for their claims, allowing for a great chance of successfully challenging biased AI-supported decisions. Crucially, this data should form the first step of a clear and transparent process to challenge AI-supported decisions, particularly where they affect government service delivery or are linked to human rights protections. This should be supported by public outreach to ensure people are aware of their rights and remedies.

Recommendation 3: Assess the risk categorisation of AI systems under the proposed regulatory framework against real-world outcomes of the system, as evaluated by ethical AI experts and representatives of those impacted by the AI system.

Recommendation 4: Require reporting of aggregate data for medium risk AI systems to ensure these systems do not perpetuate or exacerbate systemic disadvantage.

Protecting data privacy for users and content owners

All AI models require datasets to be trained on and data inputted into the model, either through automated inputs or manually (for example through a prompt in a generative AI model). The outputs produced by any AI model is (intentionally or otherwise) predicated on the data that it is trained on. In practice, this results in the, sometimes private, data the model is trained on entering into the public domain through the outputs of the AI system. Existing models of AI often develop their datasets by scraping information off the internet - the owners of that information may not be aware of this scraping, nor might they , nor fully appreciate its impact. This practice risks sensitive data (either personal or commercial) being caught up in web scraping and that data being exposed by the outputs of the model. Regulations must ensure that Australian data sovereignty is protected. This is particularly sensitive and important when it comes to data about Aboriginal and Torres Strait Islander peoples and communities, who are among Australia's most vulnerable, and whose data has often been wilfully or inadvertently misused without specific permission in the past. Other AI tools developed for medical applications must be trained on highly sensitive medical information to be accurate which, while typically deidentified, must be treated as highly confidential – especially as models become more complex and account for an ever-greater number of data points. It is essential that this need for privacy be balanced appropriately against the benefits these systems deliver.

Data entered into AI models can be similarly sensitive. Student assessments, medical results and private search queries are all likely inputs into AI systems designed to support professionals in their work. ATSE's submission to the National Robotics Strategy further highlighted the range of sensor and video data inputs that may be collected by AI enabled robots, highlighting the need for ethical storage and use of this data (ATSE 2023). The reuse of inputted data to continually train and improve AI models requires that this data be stored and protected against malicious use or privacy violations in a way that meets community expectations of security and privacy. This is potentially complicated by systems that transmit data and store it in central servers which may be outside Australia's jurisdiction.

The current proposed model of AI regulation focuses the defined risk categories on the outcomes of the AI outputs or decisions, without regard to the inputs on which they are based (i.e., the data being held by the AI systems to make those decisions). ATSE believes that failing to include data privacy provisions within the risk framework would serve to undermine public trust and adoption of AI systems, and a lack of regulation in this regard could lead to the next generation of major data leaks. While all data leaks are irreversible, community expectations around the use, storage and transmission of sensitive data should be reflected within Australia's AI regulation framework. Even users who have significant concerns about privacy often act in ways that undermine that privacy online (Barth and de Jong 2017). Individuals cannot, therefore, be solely responsible for their privacy when using AI systems and national leadership is needed to meet community privacy expectations. ATSE recommends that the definitions of each of the three proposed categories of Australia's regulatory framework be amended to include the sensitivity of data used by AI systems as a defining factor in determining risk classification.

Recommendation 5: Require AI systems operating in Australia to meet Australian regulatory requirements for privacy and data security regardless of the location of their servers.

Recommendation 6: Amend the definition of the proposed risk categories to incorporate the risk posed by data inputs to AI systems, in addition to the current categorisation based on outputs and impact.

ATSE thanks the Department of Industry, Science and Resources for the opportunity to respond to the discussion paper on supporting safe and responsible AI practices. For further information, please contact academypolicyteam@atse.org.au.

Level 2, 28 National Circuit
Forrest ACT 2603
Australia

PO Box 4776
Kingston ACT 2604
Australia

+61 2 6185 3240
info@atse.org.au
atse.org.au

ABN 58 008 520 394
ACN 008 520 394

Australian Academy of Technological Sciences & Engineering

References

ATSE (2023) *Submission to the National Robotics Strategy Discussion Paper Consultation*, https://www.atse.org.au/research-and-policy/publications/publication/submission-to-the-national-robotics-strategy-discussion-paper-consultation/, accessed 27 June 2023.

Barth S and de Jong MDT (2017) 'The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review', *Telematics and Informatics*, 34(7):1038–1058, doi:10.1016/j.tele.2017.04.013.

Bell G, Burgess J, Thomas J and Sadiq S (2023) *Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs)*, https://www.chiefscientist.gov.au/GenerativeAI, accessed 27 June 2023.

DISR (n.d.) *Australia's AI ethics principles*, https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles, accessed 29 June 2023.

European Parliament (2023) *EU AI act: first regulation on artificial intelligence*, https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, accessed 29 June 2023.

IEEE (2017) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems Ethically Aligned Design-Version 2*, https://standards.ieee.org/industry-connections/ec/ead-v1/, accessed 3 July 2023.

OECD (2023) *AI language models: technological, socio-economic and policy considerations*, https://www.oecd-ilibrary.org/science-and-technology/ai-language-models_13d38f92-en, accessed 27 June 2023.

OECD (n.d.) *OECD AI principles overview*, https://oecd.ai/en/ai-principles, accessed 29 June 2023.

OECD AI Policy Observatory (2022) *New AI technologies can perpetuate old biases: some examples in the United States*, https://oecd.ai/en/wonk/ai-biases-usa, accessed 27 June 2023.

Zapata, D. (2021, June 18). New study finds AI-enabled anti-Black bias in recruiting. *Reuters*. https://www.thomsonreuters.com/en-us/posts/legal/ai-enabled-anti-black-bias/