Australian National University

Submission to Senate Select Committee on Adopting Artificial Intelligence (AI)

May 2024

Responsible Innovation Lab
Australian National Centre for the Public Awareness of Science
The Australian National University

Ehsan Nabavi*, Xiao Han Drummond, Faranak Hardcastle, Chris Browne, Indigo Strudwicke, Dan Santos, Ala Barhoum

[redacted]

EXECUTIVE SUMMARY

- Responsible AI means different things to different people. The problem of responsibility has been framed in various ways, with each targeting specific types of solutions and certain leverage to effect change. The solutions range from technical fixes for addressing bias and explainability to broader social and policy efforts, such as promoting diversity in AI teams or establishing new regulatory frameworks. These efforts are often developed independently yet are all part of a larger constellation of interventions to make AI systems responsible.

- The disjointed and ad-hoc approach to responsible AI has led to the responsibility gaps in understanding and action concerning AI issues, opportunities, and solutions. These gaps highlight areas where accountability and effectiveness in addressing AI-related challenges are lacking, resulting in unorchestrated interventions to emerging issues in AI development and deployment.

- In this submission, we argue that responsible AI needs a systems scaffold. By that, we emphasise four critical aspects: considering the entire responsible AI ecosystem in policy and governance design, accounting for lifecycle factors, addressing responsibility questions across different scales, and anticipating and mitigating unintended consequences.

- The systems scaffold enables the Australian government to gain a holistic understanding of the entire AI landscape and the interventions carried out under the banner of responsible AI. This includes mapping these interventions globally, nationally, regionally, or by sector to identify AI issues and gaps. This includes the entire AI development process, from problem identification to system maintenance.

- Without a systemic view to guide AI development, implementation, and monitoring in government initiatives, Responsible AI initiatives may falter or fail to achieve their intended outcomes. This leaves the pursuit of responsible AI adoption vulnerable to unintended consequences, biases, or misuse.

1. INTRODUCTION

Responsible Innovation Lab is an interdisciplinary research team with expertise that spans the social sciences, engineering, and systems science, supporting crosscutting collaborations on emerging technologies including AI. Drawing on our collective understanding of AI, derived from deep engagement with existing research in this area, we adopt a systems approach in understanding socio-technical dimensions of AI systems.

Our submission offers a brief overview of responsible AI, discussing global and Australian interventions aimed at promoting responsible AI. It stresses the importance of adopting a systemic framework within government entities to effectively tackle the challenges of scaling AI systems and coordinating efforts. We underscore the need for systems scaffolding for responsible AI in Australia and the immediate benefits it provides. Our recommendations centre on problem-oriented approaches to adopting responsible AI.

The submission highlights three interconnected aspects:

- Acknowledging the diversity of interventions within the AI ecosystem under the banner of responsible AI.
- Emphasising the need for systems scaffolding in prioritising interventions based on their potential leverage to effect meaningful change within organisations using AI systems.
- Promoting cultural awareness of the systemic and interconnected nature of ‘AI’ and ‘responsibility’.

These elements underscore the systemic nature of both "AI systems" and "responsibility" within society, advocating for holistic approaches that facilitate the effective allocation of responsibility, whether in times of success or failure.

2. UNORCHESTRATED INTERVENTIONS FOR RESPONSIBLE AI

Global context

Advances in AI-based technologies, along with debates about biased algorithms and the ethical and regulatory challenges of autonomous systems, underscore the fact that the *management of AI is as much a social and political issue as it is an exclusively engineering challenge* (Nabavi, 2019; Coeckelbergh, 2022). This realisation has led to increased ‘**awareness of responsibility**’ in AI systems and their governance, defined broadly as including principles such as transparency, fairness, and accountability in creating AI technologies that meet legal requirements and societal expectations, norms, and values.

Over the last few years, numerous frameworks, principles, guidelines, and tools have been released by governments and international organisations across the world to address the ethical implications of AI-enabled systems (see Appendix 1). The nature, scope, and locus of influence of these initiatives vary widely: from software improvements emphasising isolated factors supporting or hindering AI adoption in government and industry, such as bias, safety, privacy, and security, to establishing new regulations and specific guardrails for AI design, development, deployment, and use.

These interventions are all happening without much coordination across the world (at national, regional, and international levels), and as such develop largely within specific sectors and are driven by varying social, political, and economic agendas. However, as our research shows (Nabavi & Browne, 2023), regardless of their scale or level, each of these interventions holds a certain ‘leverage’ to effect meaningful change. Many of these interventions have been framed as “technical

fixes" such a using a new software to improve fairness or explainability in the algorithm, and many have been reduced to procedural checklists/box-checking exercises in practice. Assuming equal effectiveness for all interventions risks diluting the concept of 'responsibility', turning it into yet another buzzword for businesses to exploit in their marketing campaigns. This could further erode the already fragile trust that the public has in responsible AI initiatives (Gillespie et al., 2023)

Australian context

The Australian government recently released its interim response to the safe and responsible AI discussion paper. This response, which also reflects an analysis of submissions to the discussion paper, provides insight into how those within the Australian AI ecosystem perceive responsible AI and, importantly, propose potential interventions. The Australian approach to intervention seems to draw inspiration from risk-based models in the EU and Canada that aim to enhance responsibility in the final product, such as labelling AI systems or watermarking AI-generated content or improve responsibility in processes by providing new training for developers and deployers and establishing clearer obligations to hold organisations accountable and liable for AI safety risks.

However, without a systems scaffold to view the big picture and orchestrate between different organisations, it might be unclear what is meant by responsible AI, how it looks like in practice, and who should be in charge of implementing it, how responsibility is distributed when things go wrong, such as in cases like Robodebt. This would also impede our ability to gain a broader, holistic view required for identifying key interventions areas within the Australian AI ecosystem that can most strategically strengthen national capabilities across, for example, education, training, and funding. Consequently, different stakeholders may advocate for specific interventions, criteria or indicators while undermining the value and efficacy of others.

Establishing a systems scaffold empowers a broad spectrum of stakeholders, both current and potential, to engage in discussions about what constitutes responsibility, what defines responsible technology, policy, and practice, what that entails in the context of Australia, and what consequences this can have globally. This would enable a more inclusive conversation that is necessary for assessing the potential leverage of any government-led intervention. Additionally, it facilitates better coordination between and among various actors and organisations, within the government and beyond, responsible for various parts of the AI lifecycle in the country.

Thus, in this submission, **we argue that responsible AI requires a systems scaffold; otherwise, it risks collapsing on itself.** By systemic scaffolding, we mean four key aspects:

1. Considering the entire AI ecosystem when designing policy and governance as it relates to addressing responsibility aspect (parts in relation to the whole).
2. Taking into account lifecycle factors, including different stages of AI design, development, and deployment
3. Addressing responsible AI at various scales, from small to large, and from local to global.
4. Anticipating and addressing unintended consequences

This systems scaffold could include practical tools, mechanisms, and frameworks that are scientifically proven as useful in helping policymakers, regulators, and practitioners:

- First, in *understanding the potential leverage of each intervention* for the responsible adoption of AI;
- Second, in *crafting a targeted intervention* tailored to the needs and objectives of the organisation adopting AI while remaining mindful of the holistic nature of the challenge at hand.

With a systemic scaffold guiding our discussions and interventions in responsible AI, we can effectively navigate the complex interplay of human, economic, environmental, infrastructural, and legislative factors in producing, deploying, and regulating AI technologies, particularly as they become integral to delivering public services. This is crucial for Australia to not only identify but also capitalise on opportunities for policy development, strengthen sustainable competitive advantages for industry, and safeguard citizen well-being.

However, without having such a structured appraoch, Responsible AI initiatives risk failing to achieve their intended goals—leaving the pursuit of responsible AI adoption vulnerable to being undermined by unintended consequences, biases, or misuse.

3. BENEFITS OF SYSTEMS SCAFFOLDING AI GOVERNANCE

Here, we outline some of the key benefits of implementing systems scaffolding in AI governance as governments begin to adopt AI in various activities and services:

A systems scaffold unifies fragmented efforts and establishes a shared vocabulary. The disconnect between various stages of responsible AI adoption–from conceptualisation to operationalisation and governance–often obscures the bigger picture, hampering problem articulation and hindering the development and adoption of effective interventions. A systems scaffold enables the cohesive integration of context-specific considerations, beyond siloed efforts, fostering a shared vocabulary across departments, disciplines, and stakeholders. This facilitates a coherent understanding and articulation of responsible AI concerns, thus, supporting the government for effective public interventions in this space.

In this work, science communication experts can play a pivotal role. Their expertise in creating a shared vocabulary and facilitating communication across diverse domains is instrumental in bridging disciplinary divides and fostering collaboration. By promoting interdisciplinary dialogue, science communication experts can help with enabling a systems approach to addressing AI-related concerns. Their involvement spans various levels, including policy formulation, identification of skills requirements, academic consultation, and more. Through their work, government entities can navigate the complexities of responsible AI adoption with clarity and efficacy, ensuring the development and implementation of ethical and sustainable AI practices.

A systems scaffold fosters a culture of responsibility through a well-crafted incentive structure. The majority of adoption efforts often falter due to a broken incentive system, partly stemming from the fragmented and hyper-specialised approach to responsible AI across its lifecycle. Policymakers and practitioners tend to focus solely on specific stages such as design, adoption, or maintenance, leading to disjointed strategies. Additionally, responsible AI initiatives often lack context-specificity, being framed as universal principles rather than tailored to the unique socio-cultural and environmental contexts of AI systems, hindering adoption incentives. A systems scaffold is key for establishing a holistic incentive structure aimed at promoting responsible AI adoption.

A systems scaffold supports the national responsible AI roadmap. Incorporating a systems scaffold into the development and operation of Australia's national responsible AI roadmap offers many benefits. It will provide a structure to support the government in crucial tasks such as assessing proposed interventions, allocating funding, and facilitating better coordination between organisations responsible for its design, development, and implementation. By breaking down activities conducted for responsible AI, delineating responsibilities, and defining KPIs, it facilitates the creation of a holistic action plan for Australian AI governance. This plan can then be tailored to align with the functions and duties of various teams, departments, and key public institutions in the AI lifecycle, from research and development (R&D) to data collection and post service delivery. Such alignment will improve AI-related decisions– e.g. assessment, selection, funding, management, communication, monitoring, and reporting.

A systems scaffold enables Australia to understand and evaluate its own position/role within the global context: a clear, comparative understanding of current responsible AI activities within and outside of Australia can provide nuanced and creative thinking in terms of how Australia can approach responsible AI - compared to our peers - in a way that aligns with our national values and interests. It would also help Australia to contribute more uniquely to global discussions as well.

This capability also empowers the Australian government to assess the impact of interventions carried out within Australia under the banner of responsible AI within a global context. It enables our quality of thinking and practice surrounding responsible AI to extend beyond Australia's political borders, potentially making a global impact. For instance, certain major AI companies or governments around the world may have lower thresholds for ethical and sustainability safeguards around AI design, development, and deployment. However, in Australia, by setting high standards and achieving excellence in adoption responsible AI, the country can establish itself as a benchmark for other governments worldwide seeking to adopt AI systems responsibly for the public good.

4. RECOMMENDATION AND NEXT STEPS

In light of the above discussion, we strongly recommend conducting thorough systems assessment(s) prior to developing national strategic interventions for responsible AI. This will ensure the development of a well-informed and holistic national policy that effectively aligns with the direction of our responsible AI initiative, meeting the needs, expectations, and values of the Australian people while advancing our national interests. These assessments may include:

- Assessing current capabilities across government, key industry sectors, and major education institutions is crucial to understand the default levels of responsible AI awareness, interest, and capabilities. This assessment helps identify areas of strength, gaps, and opportunities for improvement.
- Identifying and mapping key national and international stakeholders for engagement, including institutions, capacities, and ecosystem roles, enables effective collaboration and coordination in the responsible AI landscape.
- Identifying and mapping tools and frameworks currently available, along with any existing gaps, provide valuable insights for informed decision-making and strategy development in the realm of responsible AI.

Armed with these insights, we recommend the development of a national roadmap for responsible AI, including support for national programs such as funding schemes, education and training initiatives, and public communications programs. Our goal should be to enable the Australian government to leverage and harness the diverse skills, experiences, and interests within the Australian responsible AI ecosystem toward a clear and unified national vision. This vision not only recognises the diversity of solutions/interventions in this space but also their capacity to drive change towards responsible AI.

Here, we strongly recommend leveraging Australia's extensive capacity in science communication. Australian experts in science communication can play a crucial role in involving policymakers, AI practitioners, and the Australian public in the journey toward responsible AI. They can facilitate the creation of a shared vocabulary based on a systems approach, advancing AI communication across diverse domains, and encouraging cross-sector, cross-disciplinary conversation around about what constitutes responsible AI and how our nation should approach it. This open and collaborative process is vital for bridging disciplinary divides and fostering collaboration, ultimately facilitating the development and adoption of responsible AI practices through trust and legitimacy.

APPENDIX 1

Measures and initiatives developed by different actor/sectors to create positive change in AI management, achieving more responsible outcomes.

SECTOR/DESCRIPTION	EXAMPLES OF EXISTING INITIATIVES
Government Several governments have established the essential principles that underpin Responsible AI. Scientific research organisations are also helping the national government to develop operationalized guidelines for Responsible AI. OECD AI Policy Observatory reports there are more than 300 AI policy initiatives around the globe in this landscape.	• European Commission tasked an independent expert group, to develop an integrative framework for responsible and trustworthy AI (HLEG, 2019)
	• In Australia, the national science agency, CSIRO (Lu et al., 2022) uses the government's AI Ethics Principles to develop a Responsible AI Pattern Catalog for operationalizing responsible AI (from a software engineering perspective).
Industry Major AI companies have launched self-regulatory Responsible AI programs, by building tools and software to translate responsibility principles such as fairness, explainability, and accountability and use them across engineering groups and clients, as shown in de Laat (2021)'s list of software tools. The major industry actors tend to engage by developing tangible products to solve the problem.	• Microsoft and Google provide resources and recommended practices to build fairness, interpretability, privacy, and security into AI systems.
	• Fairness tools: Google (Facets, What-if-tool, Fairness Indicators); Microsoft (FairLearn); Facebook (Fairness Flow); IBM (AI Fairness 360 Toolkit); Salesforce (Einstein discovery tools).
	• Explainability tools: Amazon (SHAP); Microsoft (InterpretML); IBM (AI Explainability 360 Toolkit); FaceBook (Captum); McKensy (CausalNex)
	• Accountability tools: Google (Model cards); Microsoft (Datasheets); IBM (Fact sheets).
Academia In research, the notion of Responsible AI has attracted interest from fields as diverse as health, finance, urban studies, conservation science, marketing, and military affairs, to more specific cases such as COVID-19.	• Postgraduate coursework on Responsible AI (e.g., University of Queensland, UC Santa Cruz, Texas A&M University).
	• Curriculum design project (e.g., London New College of Humanities, 3Ai Institute at the Australian National University).

	• Interdisciplinary Research Center (e.g., Carnegie Mellon Responsible AI initiative, Cambridge Responsible AI research center, RAISE at the University of Washington).
Professional communities Professional communities and institutes offer guidance by publishing standards to describe technical specifications and procedures to develop Responsible AI systems. Certification process is another movement to enhance assurance. Independent institutions and a number of government agencies have established their own assurance mechanism to provide a seal of trust to the stakeholders involved. Consideration of broader implications that responsible AI has on other systems, such as approaches for managing risk as part of corporate digital responsibility and phased approaches to enabling global environmental sustainability.	• Working groups associated with ISO and IEEE have published guidelines; for example: IEEE both provides visionary documents on 'ethically aligned design' to show ethics in action, and also provides more detail technical guidance into components, workflows, protocol, and security requirements for machine learning in which a model is trained using encrypted data [IEEE 2830-2021].
	• Responsible AI Institute, based in the US, gives RAII certification to an AI system, which is designed, developed, and deployed in line with the OECD principles on creating AI systems.
	• The International Corporate Digital Responsibility Manifesto outlines seven principles for the practices and behaviours to help an organization be perceived as socially, economically, and environmentally responsible.

REFERENCES

Coeckelbergh, M. (2022). *The Political Philosophy of AI: An Introduction*: John Wiley & Sons.

de Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? *Philosophy & technology, 34*(4), 1135-1193.

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in artificial intelligence: A global study. *The University of Queensland & KPMG Australia: Brisbane, Australia*.

HLEG, A. (2019). Ethics guidelines for trustworthy AI. European Commission High-Level Expert Group on AI, April 8. In.

Lu, Q., Zhu, L., Xu, X., & Whittle, J. (2022). Responsible-AI-by-Design: a Pattern Collection for Designing Responsible AI Systems. *arXiv preprint arXiv:2203.00905*.

Nabavi, E. (2019). Why the huge growth in AI spells a big opportunity for transdisciplinary researchers. *Nature*, 429.

Nabavi, E., & Browne, C. (2023). Leverage zones in Responsible AI: towards a systems thinking conceptualization. *Humanities and Social Sciences Communications, 10*(1), 1-9.