# Submission in response to ‘Supporting Responsible AI discussion paper’ based on ADM+S Birmingham Workshop on Standards and Assurance for Trustworthy Data-Driven Technology

31 July 2023
Lead author: Henry Fraser
Contributing authors: Christine Parker, Fiona Haines, José-Miguel Bello y Villarino and Kimberlee Weatherall

## Questions

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

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

## Introduction

This submission responds to questions 5 and 13 in the discussion paper, focusing on standards and assurance, especially the approach to standards for AI contemplated in Europe’s draft AI Act. This submission outlines the European approach to standards and assurance for AI and evaluates the adaptability and desirability of such a regime for the Australian context.

We identify the arrangements where standards are likely to work best.

**When standards are likely to work**

*Regulatory discretion follows capabilities* - standards are likely to work better when the division of regulatory responsibilities between government, standards makers and other participants in regulation matches capabilities

*Technical assurance* - standards are likely to provide better assurance of technical features of AI systems than of socio-technical features

*Process and organisational assurance* - standards are likely to be better at providing assurance about the trustworthiness of processes and organisational arrangements, than about the overall question of whether an AI system is safe and responsible.

*Integrated assurance* - standards are most effective if integrated well with a wider regulatory ecosystem.

This submission also identifies some gaps in AI assurance infrastructure (not only in Europe but globally) that affect the desirability of the European approach, and the viability of standards in AI governance.

**Gaps in standards and assurance regimes for responsible AI**

*Expertise gap* - lack of expertise among standards bodies and participants in assurance to deal with socio-technical and public policy issues raised by AI governance

*Legitimacy gap* - lack of legitimacy on the part of standards bodies to determine consequential issues of public policy, such as how to determine when AI risk management is acceptable

*Inclusiveness gap* - inadequate representation of affected stakeholders in standards making

*Incentive and interest gap* - distorted incentives for participants in assurance regimes which may lead to box-checking, rather than truly safe and responsible AI

*Sustainability gap* - insufficient attention to ecological effects of AI, and overly quantitative, rather than qualitative, approaches to environmental impact

This submission makes the following recommendations, explained in more detail in the final section.

**Recommendations**

Recommendation 1: Do not over-rely on standards and assurance for AI governance

Recommendation 2: Australia does not have to be a fast follower of Europe on AI governance

Recommendation 3: Bridge the expertise gap in the assurance ecosystem (and the whole regulatory ecosystem) by developing multi-disciplinary AI governance expertise

Recommendation 4: Avoid a legitimacy gap in the assurance ecosystem by carefully considering the distribution of regulatory discretion

Recommendation 5: Reduce the representativeness gap in the assurance ecosystem by facilitating more inclusive participation in standards making and certification

Recommendation 6: Place more emphasis on the ecological effects of AI in AI governance

Recommendation 7: Develop detailed guidance on the socio-technical aspects of AI governance

## Authors and context of the submission

The authors are members of the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (ADM+S), which has also made a whole-of-centre submission addressing most of the questions posed by the discussion paper.

ADM+S and our partners at the University of Birmingham (UoB) held two separate, but thematically related workshops on standards and assurance for data-driven technologies at the University of Birmingham in the United Kingdom on consecutive days in May this year. This submission summarises the authors' impressions of key insights and questions from the workshops. As such it represents the expertise and views of the authors of this submission, rather than the views of the whole ADM+S centre, other workshop participants, or of our colleagues at UoB.

By way of further explanation, this submission draws upon two workshop discussions. The first workshop was hosted by Professor Karen Yeung and Dr Rotem Medzini from UoB as part of a project led by Yeung pursuant to the  European Lighthouse on Secure and Safe AI Network of  Excellence, funded by EU Horizon and UKR.[1] It was a closed workshop, convened under Chatham House rules, and attended by professionals in standards, certification and accreditation, from across Europe, with experience in assurance regimes for GDPR, medical devices and artificial intelligence. There were representatives from business, public health organisations, from national standards and accreditation bodies (such as BSI and UKAS), and from consumer organisations such as BEUC. The authors were invited participants.

The second workshop (organised by Prof Christine Parker and Dr Henry Fraser from ADM+S with the support of our colleagues at UoB) was an open international academic workshop with participants from a range of disciplines including law, computer science, political science, regulatory theory, criminology, science and technology studies, and platform governance.[2] Representatives from UK's Ofcom and members of international standards bodies such as ISO and IEEE  also attended.

## Standards and assurance in the EU AI Act

The most noteworthy AI governance initiative involving standards and assurance is Europe's draft 'AI Act'.[3] The European Commission issued a proposed text in April 2021. The Council of the European Union and European Parliament have each recently issued proposed amendments, with the final form of the Act currently being negotiated via 'trilogue', and likely to be finalised by the end of 2023.[4]

Along with a prohibition on certain uses of AI, the draft AI Act imposes a set of risk control requirements on "high risk AI systems", which pose risks to safety or fundamental rights (and, if the European Parliament's version is accepted, also systems which pose risks to health and the environment). These include requirements of accuracy, human oversight, quality assurance, documentation and logging, explainability and risk management.

Consistent with Europe's "New Legislative Framework", the Act contemplates that its requirements for "high risk AI systems" will be met through a "conformity assessment" - essentially a form of certification - against harmonised standards approved by European Standards Organisations.[5] Products must pass conformity assessment before entering the market.The default will be self-certification, although the Act also makes provision for certification by independent third parties.

[1] https://www.elsa-ai.eu/

[2] For more details, see video highlights <https://www.youtube.com/watch?v=Cg6XgPsKmPY>and event program <https://www.eventbrite.com.au/e/adms-university-of-birmingham-workshop-tickets-565687645977>

[3] *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative  Acts*, COM/2021/206 final, 21 April 2021.

[4] See, respectively, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts - General Approach* (Council of the European Union 2022) 2021/0106(COD) <https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf>; and *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* s P9_TA(2023)0236 <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html>

[5] See Art 40; Art 65(6)(b). Regarding the NLF, now under review, see European Commission, 'New Legislative Framework' (*Internal Market, Industry, Entrepreneurship and SMEs - European Commission*, 2021) <https://ec.europa.eu/growth/single-market/goods/new-legislative-framework_en> accessed 13 August 2021.

The New Legislative Framework (NLF) was designed to create a uniform approach to product safety for a wide range of products, from toys to boats to personal protective equipment. One of its hallmarks is that manufacturers or (sometimes) independent third party certifiers, known as “notified bodies”, undertake “conformity assessments” against certain “essential requirements” set out in the law in order to gain market access in the European Union. By contrast, other regimes for risky products (e.g. for pharmaceuticals) require the approval of a regulator before a product may be placed on the market. Another distinctive feature of the NLF is the role of harmonised standards (standards developed by European standards bodies and approved by the European Commission). Compliance with harmonised standards creates a “presumption of conformity” with legislative requirements.

In theory, the NLF creates an effective division of responsibilities. European Regulators give effect to public policy objectives with “essential requirements” in regulations or directives; standards bodies determine technical implementation through the development of standards; and manufacturers (or in the case of the AI Act, AI providers) who best understand the conditions on the ground, take primary responsibility for self assessment via conformity assessment against standards.

## When standards are likely to work

There is a role for standards and assurance in supporting responsible AI in Australia, and certain aspects of the approach proposed in Europe are relevant and adaptable to the Australian context.

*Regulatory discretion should align with capabilities* - The NLF is designed to allocate regulatory responsibilities to different participants in the regulatory ecosystem based on their capabilities. The spirit of this division is desirable and relevant for the Australian context. Standards and assurance are most likely to be effective in areas where standards makers and certifiers are most capable. As explained in more detail below, standards are less likely to be effective and legitimate if standards bodies and certifiers are expected to take on over-broad regulatory responsibilities that involve difficult and contested matters of public policy. These are matters that government, academia and civil society are better placed to address (although in an ideal world, standards making would become much more closely connected to, and influenced by, civil society).

*Technical assurance* - Standards and assurance are best adapted to support the technical aspects of safe and responsible AI. They are likely to be useful in developing and documenting good practice in relation to technical aspects of AI governance, including data governance, documentation and logging practices, algorithmic inspection and audit arrangements, training and testing, and establishing common metrics for accuracy and robustness. As explained in the section on challenges below, standards and assurance regimes - at least as currently constituted - are less well adapted to supporting socio-technical aspects of responsible AI.

*Process and organisational assurance* - Standards, certification, audit and other assurance practices may be useful in providing assurance that sensible processes have been followed in the development, deployment, and use of AI systems. For example, human rights impact assessments - which are likely to be standardised - may provide assurance that AI developers have at least considered human rights impacts in a systematic way, even if they do not necessarily ensure that human rights impacts are managed in the best possible way (especially since there will be disagreements about how to balance competing considerations). Standards and certification may also provide assurance that appropriate organisational measures are in place, with responsibilities for AI risks allocated to appropriately senior and qualified people within an organisation.[6]

[6] See e.g. NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) AI 100-1 (January 2023) <https://www.nist.gov/itl/ai-risk-management-framework>, p 8.

*Integrated assurance* - The effectiveness, desirability and adaptability of standards and assurance to AI and the Australian context will be influenced by other factors in the regulatory ecosystem such as: what legal obligations apply to various participants in the AI value chain and the assurance infrastructure (e.g. a duty of care for certifiers); the degree of oversight or involvement by regulators; the degree to which stakeholders can reach consensus about various aspects of AI governance; and appropriate allocation of different kinds of regulatory discretion as between government regulators, industry bodies, standards bodies, and others.[7]

## Adaptability and relevance of the European approach to the Australian context

While standards, set against an appropriate regulatory architecture, may support responsible AI, there are reasons to be cautious about adopting a European approach wholesale in Australia. Firstly, Europe has explicit, formal protection of human rights - for instance in the Charter of Fundamental Rights of the European Union. Australia has no such legal instrument. Standards and assurance regimes developed for the European context (or any other context where human rights are expressly protected by law) may assume an existing rights framework. Consequently, these regimes may fail to provide explicitly for the protection of certain rights and interests to the degree necessary to produce responsible and safe AI in Australia.

Secondly, Australia is not bound by Europe's new legislative framework. There is no pressing need to follow the European approach to AI regulation; but also no clear policy here on where standards bodies should fit into the governance of technology. This presents an opportunity. Australia is free to explore new ways to integrate standards and assurance into AI regulation and governance. However, in doing so, the government should be aware of several gaps and challenges associated with standards and assurance for responsible AI.

## Gaps in standards and assurance regimes for responsible AI

Many (though not all) participants in our workshops expressed serious doubts about whether an assurance regime for AI of the kind contemplated by Europe's AI Act - where regulators rely heavily on standards to determine the details of how safe and responsible AI is implemented- is adaptable or desirable for the governance of AI. We set out below some important challenges and problems for standards and assurance in AI governance and their implications.

*Expertise gap* - It is not clear that standards bodies have the experience or expertise to play the role contemplated by the European AI Act (or any significant role in tackling the normative, social and political aspects of AI governance).[8] Standards bodies, and other participants in assurance infrastructure (like accreditors and certifiers) generally have a technical, qualitative orientation,with engineers playing a prominent role. By contrast, assuring AI is safe and responsible in multiple dimensions - from human rights impact and environmental impacts to health and safety - involves social and political considerations. Experience with product safety, for instance, does not necessarily equip standards bodies to decide how accurate a criminal recidivism predictor must be, when risks of racial or other bias have been sufficiently mitigated, what kind of documentation ought to accompany it, or how its decisions ought to be explained. Socio-technical value judgments require experience with public policy, and may also call for expertise in a range of humanities and social sciences. Any

[7] See eg Colin Scott (2001) Analysing regulatory space: fragmented resources and institutional design. Public Law, 329-353; Fiona Haines and Christine Parker, Reconstituting the Contemporary Corporation Through Ecologically Responsive Regulation, *Company and Securities Law Journal*, 2023, 39(6), 316-331.

[8] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22 Computer Law Review International 97.

effective assurance regime for AI would need to correct this expertise gap, or be complemented by other regulatory regimes to address socio-technical aspects of AI governance..

*Legitimacy gap* - Standards bodies arguably lack the regulatory legitimacy to make judgements of the kind just described.[9] Why should it fall to technical standards bodies to decide what counts as a risk from an AI system, how much risk management is enough, or what kind of explanation of an AI decision with life-changing effects might be appropriate? In the first place, this degree of discretion is inconsistent with the division of responsibilities envisioned in the NLF. In any event (since the NLF is not relevant for Australia) there is at this stage no developed understanding or theory of whether, or under what conditions, it might be appropriate for technical standards bodies to exercise discretion in relation to non-technical issues that raise significant questions of public policy.

*Consensus gap* - Standards tend to work best when there is consensus about goals, and at least some degree of agreement and clarity about practical implementation.[10] Such consensus has still not crystallised for AI governance. One of the most striking impressions from the UoB industry workshop on standards was how uncertain many participants were about how to develop and certify against standards for AI, and how much of a work in progress standardisation for AI really is. There is a clear need for guidance by regulators, governance bodies and researchers with the requisite expertise and legitimacy to assist stakeholders in improving and cementing good practice.

*Resource and inclusiveness gap* - Standards making for AI must become more inclusive of stakeholders if it is to contribute effectively to safe and responsible AI.[11] A lack of representation in this process renders the problems of legitimacy and expertise described above more acute. The European Commission's standardisation request to European standards bodies and the European Parliament's proposed amendments to the AI Act both emphasise the need to promote inclusiveness in standards making.[12] However, barriers to participation tend to be practical rather than formal. In our workshops, representatives from standards bodies made it clear that joining a committee to work on a given standard (for example at ISO or IEEE) is generally straightforward. The problem is that large companies have the resources to support participation and networking by their representatives, while stakeholders such as civil society organisations do not.[13] As a consequence, commercial interests tend to have a disproportionate influence over the development and content of standards.

*Incentive and interest gaps* - As with all forms of regulation and governance, standards and assurance processes are susceptible to capture and misaligned incentives. Self-certifiers clearly have conflicts between their commercial interests, and the public interests supposed to be protected through the certification process.Third party certifiers also deal with potentially conflicting interests and incentives,

[9] Henry L Fraser and Jose-Miguel Bello y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough' (9 July 2023) <https://papers.ssrn.com/abstract=4516917> accessed 26 July 2023; Rotem Medzini and Karen Yeung, "Background paper: assurance regimes for data-informed services", Assurance Regimes for Data-Driven Services Workshop, (May 2023) University of Birmingham. See also Karen Yeung, "Operationalising Trustworthy AI Governance: Beyond Motherhood and Apple Pie?" ADM+S Symposium 2022 <https://youtu.be/5id15TlpacY>
[10] Julia Black, *Rules and Regulators* (Clarendon Press 1997)
[11] Christine Galvagna, 'Discussion Paper: Inclusive AI Governance' (Ada Lovelace Institute 2023) <https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/> accessed 15 May 2023.
[12] European Commission (2022) *AI Act: Draft Standardisation Request*, Annex II, accessed 14/4/23 <https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-COM-%E2%80%93-Draft-Standardisation-Request-5-December-2022.pdf>; *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* s P9_TA(2023)0236 <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html>, Amendments 103 and 104
[13] Ibid. See also Hans-W Micklitz, 'The Role of Standards in Future EU Digital Policy Legislation: A Consumer Perspective' (ANEC BEUC 2023) <https://www.beuc.eu/sites/default/files/publications/BEUC-X-2023-096_The_Role_of_Standards_in_Future_EU_Digital_Policy_Legislation.pdf> accessed 26 July 2023.

as we learned in the UoB closed industry workshop. Certifiers often owe duties of confidentiality to their clients, meaning they are not able to disclose risks that they detect which are outside the remit of their certification. In the worst case, narrow, formalistic standards and certification processes operate to paper over risks and bad practice: as occurred in the notorious Rana Plaza disaster.[14]

*Sustainability gap* - Environmental impacts have been something of an afterthought in AI governance. For example, references to environmental risk were a late inclusion in the European Parliament's proposed amendments to EU AI Act. One of the themes emerging strongly from our academic workshop was the importance of the physical realities of AI: the extent of environmental damage caused by the AI ecology and economy; and the degree to which this damage is obscured by AI hype and neglected in 'responsible AI' discourse.[15]

## Recommendations and open questions

In view of the analysis above, we offer several recommendations. These are accompanied by some key questions raised in our workshops, which will be important to consider as Australia's AI governance policy and regulation develops.

**Recommendation 1: Do not over-rely on standards and assurance for AI governance**

There are serious doubts about whether technical standards can really be the basis for a judgement that the whole complex system with all its inputs, impacts and human factors is safe and responsible.

**Recommendation 2: Australia does not have to be a fast follower of Europe on AI governance**

While the 'Brussels' effect is well known (e.g. the global impact of Europe's GDPR), Australia does not necessarily have to follow the European approach to AI regulation, where conformity with essential regulatory requirements is determined by compliance with standards. It is yet to be seen whether standards are up to the task set for them by Europe's draft AI Act.

**Recommendation 3: Bridge the expertise gap in the assurance ecosystem (and the whole regulatory ecosystem) by developing multi-disciplinary AI governance expertise**

Standards and assurance are likely to play *some* role in AI governance, even if not the extensive role contemplated in Europe. AI governance, by whatever means, is inherently multi-disciplinary and involves difficult socio-technical questions. There is an urgent need to develop the multi-disciplinary expertise that would be required to develop a safe and responsible governance of AI, not only in standardisation and assurance infrastructures, but throughout the regulatory ecosystem. Pathways to bridging the expertise gap include:

a. *Guidance and training* - Develop and implement guidance and training on fundamental rights, public health, environmental and other AI impacts for accreditation bodies and certifiers, and others involved in AI governance, as well as training on technical aspects of safe and responsible AI such as accuracy, robustness, data governance etc.
b. *Organisation measures* - Standards bodies, accreditors, certifiers and auditors should be encouraged, or required, to bring in experts from an appropriate range of disciplines.

[14] See, e.g. "More for show than safety: Certificates in the textile industry", European Centre for Constitutional and Human Rights <https://www.ecchr.eu/en/case/more-for-show-than-safety-certificates-in-the-textile-industry/>
[15] Simon Coghlan and Christine Parker, Harm to Nonhuman Animals from AI: a Systematic Account and Framework. Philos. Technol. 36, 25 (2023). https://doi.org/10.1007/s13347-023-00627-6; Sasha Luccioni, "The mounting human and environmental costs of generative AI", Ars Technica, 4/12/2023. Available at: https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/

c. *Multi-disciplinary partnerships between researchers, government and industry* - Universities and research centres such as our own are places where multiple disciplines, working deeply and widely, are under the same roof. Partnerships between universities and regulators, standards makers, accreditation bodies, certifiers, civil society and industry are necessary and valuable, and government should take steps to support these partnerships.

**Recommendation 4: Avoid a legitimacy gap in the assurance ecosystem by carefully considering the distribution of regulatory discretion**

How comfortable are Australians with relying on technical standards bodies to answer questions of public interest about rights, the environment etc.? Who has the legitimacy and expertise to set rules and policy in relation to AI risk acceptability, explanation, and other aspects of AI governance with significant public policy implications? If this combination of expertise and legitimacy does not yet exist in Australia's regulatory ecosystem, this may be a reason to establish a dedicated AI regulator. Further research and policy discussions are needed to develop a clearer sense of the role of standards bodies in highly charged policy decisions, and what complementary elements in a regulatory ecosystem are needed to ensure that socio-technical aspects of AI governance are performed with appropriate expertise and legitimacy.

**Recommendation 5: Reduce the representativeness gap in the assurance ecosystem by facilitating more inclusive participation in standards making and certification**

Representativeness and inclusiveness in standards making and certification could be enhanced by various means, including through government funding to assist civil society and academic participation in standards-making and assurance, and government funding of new, unconventional standards-making bodies.[16] There is value in a diverse standards and assurance ecosystem with competing standards and assurance regimes. Australia can invest both in the development of a range of standards and standards-makers locally, and in fostering Australian participation in global standard-setting for AI. Indeed, as standards making for AI is currently in an early stage, there is an opportunity to exercise influence by setting aside resources to support Australian AI experts' participation in standards making.

**Recommendation 6: Place more emphasis on the ecological effects of AI**

Standards and assurance may serve to increase the visibility of certain aspects of environmental damage caused throughout the AI value chain, but a comprehensive policy effort iwill be needed to manage the environmental impact of AI.[17]

**Recommendation 7: Develop detailed guidance on the socio-technical aspects of AI governance**

Government may not be best placed to provide detailed guidance on *technical* aspects of AI (as recognised by Europe's NLF approach). It is however better positioned than standards and assurance professionals to provide detailed guidance on how stakeholders should grapple with AI's

[16] Christine Galvagna, 'Discussion Paper: Inclusive AI Governance' (Ada Lovelace Institute 2023) <https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/> accessed 15 May 2023.

[17] See e.g. E Vinnari and M Vinnari M (2022) Making the invisibles visible: Including animals in sustainability (and) accounting. *Critical Perspectives on Accounting* 82. Special Issue: Covid and the Environment in Crisis: 102324. DOI: 10.1016/j.cpa.2021.102324; I Ali, PT Fukofuka and AK Narayan (2023) Critical reflections on sustainability reporting standard setting. *Sustainability Accounting, Management and Policy Journal* 14(4). Emerald Publishing Limited: 776–791. DOI: 10.1108/SAMPJ-01-2022-0054; Bogdana Rakova and Roel Dobbe (2023) Algorithms as Social-Ecological-Technological Systems: an Environmental Justice Lens on Algorithmic Audits. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, New York, NY, USA, 491 At: https://dl.acm.org/doi/abs/10.1145/3593013.3594014.

*socio-technical* aspects (such as dealing with trade offs between the rights and interests of different stakeholders). Where standards deal with difficult public policy questions the government may need to provide additional guidance to stakeholders.[18] Where a given standard or part of a standard does not meet Australian legal or ethical requirements, the government can specifically issue guidance to this effect (the European Commission does so with technical standards). If Australia establishes a dedicated AI regulator, guidance on difficult socio-technical questions involved in AI governance and commentary on technical standards and other forms self-regulation should be part of its remit.

[18] See e.g.Henry L Fraser and Jose-Miguel Bello y Villarino, “Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (Working Paper, September 30, 2021) <http://dx.doi.org/10.2139/ssrn.3960461>, part 5.