26 July 2023

Australian Government

Department of Industry, Science and Resources

Technology Strategy Branch

SAFE AND RESPONSIBLE AI IN AUSTRALIA: RESPONSE TO DISCUSSION PAPER

Thank you for the opportunity to offer responses to some of the questions raised in the discussion paper *Safe and Responsible AI in Australia* (June 2023).

We commend the department for being proactive in its consideration of an appropriate regulatory response to ensure the safe and responsible use of artificial intelligence. Much thinking has already gone into the discussion paper. We therefore limit our comments to questions 1, 3, 9, 11, and 14 – 20 of the document.

Our main contention is that public trust in the use of AI systems will only follow when one can objectively assure that the system adheres to the set principles and guidelines of appropriate behaviour. This is the case regardless of whether the regulatory strategy followed is voluntary or mandatory.

Guy Lupo is currently conducting PhD research at Swinburne University, under the supervision of Ass Prof Bao Quoc Vo, Prof Ryszard Kowalczyk and Ass Prof Natania Locke, with a view to develop an automated risk-based continuous assurance of AI systems' ethical posture. The intention is to develop the ability to assure that an AI system complies with a pre-determined code or set of objectives using automated, scalable systems. The responses below are specifically informed by this research.

Question	Response
1.	We found the definition of AI in the document to be rather ambiguous (something that should be avoided in a definition) and conflicting. For instance, this definition defines AI as "engineered systems ... without explicit programming," which implies that all AI systems are fully automated and autonomous. But the very next sentence in the same definition of AI states that "AI systems are designed to operate with varying levels of automation." The requirement to include only systems "without explicit programming" would also exclude many AI-related technologies such as logic programming, AI planning, etc. We also consider limiting AI systems to only "generate predictive outputs" makes this definition overly restrictive as AI systems have been used to generate creative outputs among many other applications.
3.	The European Union's proposed Artificial Intelligence Act (AIA) advocates for the development of an ecosystem of trust. This concept, operating adjunct to regulatory frameworks, is designed to foster cooperation, business growth, and innovation among trusted AI developers, deployers, and users. We propose that the Australian government evaluate the potential advantages of investing in and establishing the necessary infrastructure and programs, as well as promoting collaborations, to cultivate a similar ecosystem. This strategic move could serve

	as a catalyst for enhancing the adoption of AI, whilst ensuring its use aligns with ethical and responsible standards.
9.	Based on our proposed risk-based approach to autonomous ethics-based audit of AI/ADM systems, we consider several critical components/phase of AI lifecycle where transparency is essential including data (to train/test the AI systems) and validation of an AI system outputs regarding the identified and potential risks.
11.	While self-regulatory approaches are laudable and often preferred by industry, we are of the opinion that public trust will only follow if adopters of AI systems can show that their systems perform true to purpose. To this end, we recommend that government considers the adoption of assurance mechanisms that are supervised by a regulator. The result could take the form of a badge or a seal, visible on the platform through which a consumer interacts with the system, that confirms such assurance. The underlying assurance would be the same regardless of the functionality of the system, namely that the system adheres to a set of principles or guidelines committed to at the outset. The complexity of the assurance behind the scenes would depend on the risk factors associated with the particular system and its use.
14.	We fully support the risk-based approach for addressing potential AI risks. One benefit of this approach is that assurance may be customised based on the risk factors associated with a particular system. In other words, this approach is adaptable to different contexts and uses of AI systems.
15.	Navigating the challenges presented by a risk-based approach to AI regulation involves careful strategising and robust policy planning. A significant concern of this model is that it fails to offer a foundational standard that delineates "what good looks like". Consequently, it permits organisations to establish their own standards of compliance, a prevalent issue in all attestation-based frameworks. While this model provides a faster track to industry adoption by allowing organisations to accept, mitigate, and transfer risks according to their risk tolerance, it is inherently a longer road to achieving public trust, heavily contingent on the government's auditing capacity. Potential solutions to these challenges could draw from practices within the security industry, which employs a mature risk-based approach. This industry combines a basic standard (or hygiene level) with an attestation structure and third-party conformity testing. Furthermore, an investment in regulatory infrastructure for 'trust by design' could be instrumental, incorporating trust into organisational lifecycle, development, and internal auditing functions. This would enable scalability and lessen the dependency on government auditing. Fostering an ecosystem of trust can be an effective strategy. In such an ecosystem, trusted components complement each other, facilitating a quicker path to public trust without hampering industry adoption. It's essential to recognise that a balance between regulation, innovation, and public trust is key to the successful implementation and adoption of AI technologies.

16.	Since the risk framework of different systems and their contexts differ, we do not believe that a risk-based approach is less suitable for some developers, deployers or users. Instead, the complexity of the risk framework will differ depending on the nature of the system and its intended use. From a resources point of view, the automation of assurance could conceivably alleviate some of this concern. Moreover, if government assists with an independent seal of assurance (see question 11 above), users could take comfort in the reliability of the system regardless of the size of the deployer, thereby encouraging the adoption of AI.
17.	We agree with the elements listed in Attachment C, but would add that this would be an iterative process that must be continuously revised. This is especially so for impact assessments in new areas of deployment, where risk has not yet been well-defined. We would further add that 'humans in the loop' may be a significant barrier to adoption at scale. For this reason, we prefer that a model of automated assurance be developed. It must be possible to intervene in a system when its irresponsible use is detected. Here are some additional elements that might be considered as part of a risk-based approach: • Security – consider all risks, controls, mitigations and validation of AI system security; • Design & architecture – consider all risks relating to the design of AI system; • Evidence – establish a standard for AI system evidence recording and storing; • Data quality – consider all risks relating to the data, origin, and quality.
18.	Risk management and assurance of AI systems should be firmly rooted in, and leverage, existing governance, risk, and compliance practices that apply to general IT systems. The unique aspect of AI systems, which differentiates them from traditional IT systems, is their potential to cause harm to humans, and their significant dependence on data and algorithmic components. This exposes an entirely new vector of risks and, consequently, necessitates novel forms of mitigating controls. These considerations converge to form a new ethical dimension, seeking to illustrate the impact and outcomes on humans from a combination of systems, data, and other AI components. It's crucial, however, for organisations to avoid recreating or constructing an entirely new regime of risk management. Instead, they should build upon their existing IT and system risk processes. To this end, it would be beneficial for organisations to leverage existing AI-specific audit and assurance frameworks such as ECCOLA, Aequitas, FEMA-AI, Ethics-Based Audit (EBA), CapAI-A, and GAFAI. These frameworks can be seamlessly integrated into existing risk management processes and structures, enhancing them without necessitating wholesale replacement or upheaval. This approach allows for a more effective, efficient, and nuanced approach to managing the risks associated with AI systems.
19.	Large Language Models (LLMs) are integral components of an AI system, with their associated risks, controls, and mitigations not differing substantially from other model-related components. However, the distinctiveness of LLMs lies in their size and the fact that they

	cannot be downloaded, either due to commercial restrictions or sheer volume. Accordingly, LLMs introduce an additional set of risks stemming from their Software-as-a-Service (SaaS) nature, the cloud-based and outsourced nature of the model. These risks are primarily operational, closely associated with cybersecurity, data loss prevention, and general data privacy and protection. Another potential issue lies in the LLM service's capacity to assure their own security practices, supply chain integrity, and adherence to responsible use norms. A significant consideration is that LLM services generate responses based on data that is neither validated nor traceable. This creates potential liability issues for the consumer of the LLM and challenges in relying on the responses. Therefore, a risk-based approach is ideally suited for this scenario, as it enables the consumer organisation to define its policy thresholds and assess the service provider's compliance with its baseline risk appetite. In turn, the service provider can invest in transparency and align with the consumer organisation's risk practices, offering the necessary assurances on demand.
20.	We recommend a light touch mandatory approach, whereby the appropriate use of the suggested assurance mechanism would be verified by government recognition. Developers, deployers, and users that opt not to adopt the assurance of their system would not be allowed to use the government confirmation that their system is a responsible one. This approach must be scaled, similarly to the proposed approach in the EU, to mandatory compliance where the system poses a high risk to users or the public. We are in favour of adoption of a risk-based approach by both private and public sector organisations, as well as both deployers and developers. A risk-based approach cannot operate without the impact assessment of both deployers and developers, as some risks will present from the context of use (for instance, the business of the deployer) and others will present as a function of the development process.

Regards,

Ass Prof Natania Locke

Ass Prof Bao Quoc Vo

Mr Guy Lupo