4 August 2023

To: Department of Industry, Science, Energy and Resources (Australia)

RE: Safe and Responsible AI in Australia

The IEEE Standards Association (IEEE SA) welcomes the opportunity to provide its input to the Department of Industry, Science, Energy and Resources (Australia) and its report Safe and Responsible AI in Australia.

IEEE SA is a globally recognized standards-setting body within IEEE, the largest organization of technology professionals in the world. We develop consensus standards through an open process that engages industry and brings together a broad stakeholder community and comply with the WTO Principles for International Standardization.

IEEE SA, through its global community, has developed resources and standards globally recognized in the area of applied ethics and systems engineering and offers standards, training and education, certification programs, and more, to empower stakeholders designing, developing, and using Autonomous Intelligent Systems (AIS).

[It should also be noted that approved, published IEEE AI Ethics & Governance standards, some of which are referenced in the responses, are available for access free of charge at the IEEE GET Program here: <https://ieeexplore.ieee.org/browse/standards/get-program/page/series?id=93> ]

| **Question from Safe and Responsible AI in Australia** | | |
| --- | --- | --- |
| **Definitions** | | |
| 1 | Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why? |
|  | As the “[Supporting responsible AI: discussion paper](https://consult.industry.gov.au/supporting-responsible-ai)” notes there is no agreed upon definition for Artificial Intelligence, however IEEE would like to call attention to its work in [“Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”,](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf) as it does contain an accepted definition for AI and encompasses both ML, Generative AI models and all applications of automated decision making and recommendations.  An Autonomous/Intelligent System (A/IS) is a semi-autonomous or autonomous computer-controlled system programmed to carry out some tasks with or without limited human intervention capable of decision making by independent inference and successfully adapting to its context.  The [Glossary](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_glossary.pdf) that accompanies IEEE’s “Ethically Aligned Design 1st Edition” contains AI accepted terms and definitions and is a valuable reference source. The Glossary is an interdisciplinary resource for AI terms which may have meanings that are discipline specific: Ordinary Language, Computational Disciplines, Engineering, Economics & Social Sciences, Ethics & Philosophy, and International Law & Policy.  IEEE’s Ethically Aligned Design principles is the source document for [Australia’s AI Ethics Principles | Australia’s Artificial Intelligence Ethics Framework | Department of Industry, Science and Resources](https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles).  For the success of AI applications and systems, trust is crucial, which means that a definition must make sure that all potential users are well informed about these systems, especially those potential users with no expertise in the subject, including the engagement of direct and indirect stakeholders.  IEEE has a portfolio of standards associated with AI, covering such areas as adaptive instructional systems, age appropriate design, ethically aligned design and machine learning, that can contribute to a fuller understanding of AI and its applications. We note two below as example, but the full list of standards can be found at <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/>  IEEE [7010-2020](https://ieeexplore.ieee.org/document/9084219)TM. IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being  [IEEE 7001TM](https://standards.ieee.org/ieee/7001/6929/)- IEEE Standard for Transparency of Autonomous Systems |
| **Potential gaps in approaches** | | |
| 2 | What potential risks from AI are not covered by Australia’s existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks? |
|  | There are several potential AI risks that have been identified:    Misinformation, which may need additional measures such as a Code of Practice to ensure that the private sector prioritizes disinformation (e.g., as included in the [EU Action Plan against disinformation](https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation)), to improve civil society awareness, to increase press and media training through practical examples and their implications.  Another growing issue is one of algorithmic management in the workplace. Such systems may make unfair misattributions of employees’ intentions and character, ones which are very difficult to challenge. Consideration should be afforded as to what degree such systems can operate, especially for remote workers who may labor in domestic environments. Transparency should also be provided as to any decisions or predictions made by such systems, including the underlying predicates.  For example, we suggest that lessons of ‘Robodebt’ [Royal Commission into the Robodebt Scheme](https://robodebt.royalcommission.gov.au/) be considered for incorporation into Australia’s AI regulatory approach to use, procurement, governance, management and disposal of all autonomous and intelligent systems. As shown by this report, court outcomes and ongoing publicity, the risks have evolved into severe issues for both current and previous Governments, the public service, and citizens, both in terms of economic costs and for the wellbeing of these institutions.  In particular, the report’s recommendations regarding Automated Decision Making (Rec. 17.1 and 17.2)  17.1: ‘The Commonwealth should consider legislative reform to introduce a consistent legal framework in which automation in government services can operate….’ and  17.2: ‘The Commonwealth should consider establishing a body, or expanding an existing body, with the power to monitor and audit automate decision-making processes with regard to their technical aspects and their impact in respect of fairness, the avoiding of bias, and client usability’.  The paper also references a possible risk-based approach (Annex C). However, we note missing is a definition of what is considered a “risk”. It would be useful for the Australian DISR to provide some guidance of what are considered legal risks, commercial/economic risks and ethical risks, including providing understanding of the short-, medium- and long-term impacts of such risks, which should be mapped, managed, mitigated, and monitored.  IEEE suggests the following programs and standards be considered to help businesses and governments understand what risks are in light of AI ethics:   * [IEEE 7000™-2021](https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html?utm_source=ieeesa&utm_medium=aem&utm_campaign=ais-2021&_gl=1*p36uss*_ga*MTM0ODg2ODAxLjE2OTAzNDI2MDQ.*_ga_XDL2ME6570*MTY5MDM0NDc3Ni4xLjEuMTY5MDM0NjYyMy4yNy4wLjA.) integrates ethical and functional requirements to mitigate risk and increase innovation in systems engineering product design. * [Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf) * [Ethics Certification Program for Autonomous and Intelligent Systems](https://standards.ieee.org/industry-connections/ecpais/)(ECPAIS) creates specifications for certification and marking processes that advance transparency, accountability, and reduction in algorithmic bias in Autonomous and Intelligent Systems (AIS). * [IEEE CertifAIEdTM](https://engagestandards.ieee.org/ieeecertifaied.html?_gl=1*1095nml*_ga*MTM0ODg2ODAxLjE2OTAzNDI2MDQ.*_ga_XDL2ME6570*MTY5MDM0NDc3Ni4xLjEuMTY5MDM0NjI1OS4xMC4wLjA.)is a certification program for assessing ethics of Autonomous Intelligent Systems (AIS) to help protect, differentiate, and grow product adoption. |
| 3 | Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts. |
|  | AI technologies continue to advance and play an increasingly significant role in our lives.  AI systems should be developed and used in a manner that upholds ethical principles and safeguards the well-being and interests of individuals and societies. There is also a need to use regulatory and non-regulatory measures to raise awareness of AI, ethical use and awareness, combined with supporting AI digital literacy amongst legislators, public sector, private sector, and individuals / groups which is contextually relevant.  To that end, the following responsible measures need to be considered for implementation:   * Awareness and Intelligence of AI and its implications   The need to provide incentives to implement AI responsibly and within the relevant legal and non-regulatory mechanisms such as regulatory guidance, processes, frameworks and use of technical standards and assessment or testing methods like IEEE CertifAIEd’s certification within controlled environments such as regulatory sandboxes, and finally to provide guidance to industry actors through access to global and local expertise.  [IEEE’s AI Portfolio](https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/) is available for public access and improved understanding, to help organizations differentiate and build competencies and methodologies to achieve the outcome of a ‘trustworthy organization.’   * Use of assessment tools to identify the degree of readiness to adopt AI ethics principles.   Main benefits for adopting an AI readiness self-assessment include an updated identification of the conditions in which it is operating and interacting with stakeholders, assessing the threats and risks, seizing opportunities for learning and improvement, and addressing the needs for a future roadmap for action. Including the AI readiness framework can bring value to organizations as it will assist with organizations gaining a clear overview of their maturity and preparation for AI implementation or integration before investment or management decisions are taken.  The [IEEE CertifAIEd program can](https://engagestandards.ieee.org/ieeecertifaied.html?_gl=1*kajuk3*_ga*MTU5NDQyMTcwOC4xNjgyNTIxNzg5*_ga_XDL2ME6570*MTY5MDM4OTMxNC4xMTMuMS4xNjkwMzkwMDM4LjE1LjAuMA..) assist in determining readiness and compliance.   * Training programs for AI auditing / conformity assessment.   Certification relative to assessing ethics of Autonomous Intelligent Systems to help protect, differentiate, and grow product adoption.  Also see IEEE [7010-2020](https://ieeexplore.ieee.org/document/9084219)TM. Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being |
| 4 | Do you have suggestions on coordination of AI governance across government? |
| Implementation of a standards-based approach is extremely important to help establish a regulatory impetus for businesses to act as compliance for best practice and to increase interoperability on an international playing field.  Use of regulatory sandboxes are a safe environment to test out the standards for their usefulness and appropriateness to context, inviting the market to explore responsible innovation, for AI products to test for compliance or their societal impact, or for their standards applicability. The sandboxes can also point to any gaps in the regulatory environment.  For example, we suggest that there be clear Public Procurement guidelines across the government such as those provided for in [IEEE P3119TM](https://standards.ieee.org/ieee/3119/10729/) Standard for the Procurement of Artificial Intelligence and Automated Decision Systems.  We note that there needs to be greater stakeholder engagement across the government to truly understand the ethical risks and impacts on individuals / groups. |
| Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia. |
|  | Such practices as outlined above would help drive greater accountability, responsible governance practices, better assessment of ethical risk and impact, leading to better innovation and product/service improvement and trustworthiness of AI produced by organizations, resulting in adoption of responsible AI in Australia. |
| **Responses suitable for Australia** | | |
| 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? |
|  | The European Union (EU) is putting policies and legislation in place that will impact AI systems both within the EU and those that concern any EU entity.  These instruments presume conformity with requirements - whether standards, common specifications, conformity assessments, certificates, and EU declarations of conformity, which cover areas like the quality of data sets used to train, validate, and test the AI systems, technical documentation, record-keeping, transparency, human oversight, and robustness.  We posit that most of these requirements must be embedded in the design of the ‘high-risk AI system’ from the earliest stages of designing and developing the AI system. Even if the provider is not the designer or developer of the system, they still need to ensure that requirements are embedded in the system to achieve conformity status.  The EU AI Act also establishes a *presumption of compliance* with the requirements for high-risk AI systems, where a high-risk AI system is in conformity with relevant [harmonized standards](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R1025&from=EN).  We suggest review of suites of International Trustworthy AI standards offered such as  the EU Joint Research Report of 2023, [JRC Publications Repository - AI Watch: Artificial Intelligence Standardization Landscape Update (europa.eu)](https://publications.jrc.ec.europa.eu/repository/handle/JRC131155), which has analyzed IEEE standards on AI and their suitability to operationalize the obligations of AI providers of high-risk AI under the proposed EU AI regulation.  There are several countries who have adopted AI governance measures or are in the process of adopting regulatory guidance, examples to consider are:  **Spain**: It may be interesting to analyze the case of Spain. A [Spanish Artificial Intelligence Supervisory Agency (AESIA)](https://commission.europa.eu/projects/headquarters-spanish-agency-supervision-artificial-intelligence_en) was launched and the government will support a funding initiative ( Fond-ICO Next Tech) targeting cutting-edge start-ups, and has earmarked €500 million from the Recovery Plan to accelerate the development of Artificial Intelligence in Spain.  **Portugal:** also, through the Recovery Plan (EU funded), the government approved two major initiatives, gathering more with a total funding of nearly € 100 million.  **Germany**: [NFDI](https://www.nfdi.de/?lang=en) has been supporting large consortia, including one related to Data Science (DS) and AI. The work plan includes several activities being developed, such as:   * Community and Training: DS and AI training, skill development, and capacity building, as well as leveraging existing platforms and resources. * Research Knowledge Graphs: improving FAIRness of Data Science artifacts including research datasets, benchmarks, machine learning models and research software (code and executables). * Infrastructure and Services: to collect and share all input, called Digital Objects (DO) which is required to deliver quality-assured data analytics solutions. The registries and repositories keep track of releases of quality assessed data as well as Data Science solutions required for assessment and the resulting benchmark information. * Transfer and Application: to create a strong connection between the Data Science and AI sub-communities: (1) Natural language processing and language technology as well as Semantic Web, (2) biomedical research and clinical decision-making, (3) information sciences and (4) social sciences. * Interoperability and Cooperation: DS and AI involves a plethora of artifacts, e.g., datasets, models, ontologies, task definitions, code repositories, execution platforms, repositories, training materials, and so on. These artifacts are currently hidden in a number of platforms that manage the respective content. By making all digital artifacts available and interlinking them, NFDI4DS will foster interoperability, and collaboration between Data Science and AI platforms. |
| **Target areas** | | |
| 6 | Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ? |
|  | Similar approaches could be adopted but they must recognize the key differential between the contexts of public use of AI technologies and, and private sector use of AI technologies. It will also be important to document how each sector will be held accountable for its use of AI technologies.  Fundamentally it is about the harm that can be caused by uses of AI. In Public Sector uses of AI can be pervasive and due to the concentration of power through AI systems, the harms are greater. See RoboAdvice, SyRI and Dutch social welfare system cases as examples. This is however often jurisdictionally bound and may only impact a population.  Private Sector uses of AI can be ubiquitous crossing borders internationally and potentially causing harm across multiple populations and multiple sectors, including Public Sector procurement. Holding AI product/service providers accountable is harder across multiple jurisdictions, as no one country can do this alone. |
| 7 | How can the Australian Government further support responsible AI practices in its own agencies? |
|  | We suggest that the Australian Government review publication IEEE SA [Ethically Aligned Design for Business](https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead-for-business.pdf), which explains the cultural and behavioral change needed within organizations, and what it takes to enable AI ethics as a core organizational competency.  It provides an ‘AI readiness framework’ with steps to help facilitate a change in the organizational culture, as a proactive measure to build AI ethics into the governance structure of the organization. This would be a useful guide to support and build responsible AI practices in Australian Government Agencies. |
|  | In what circumstances are generic solutions to the risks of AI most valuable? |
|  |  |
| 8 | And in what circumstances are technology-specific solutions better? Please provide some examples. |
| From an ethical standpoint, different AI-enabled technologies have different potentials for algorithmic harm/erosion of fundamental/universal human rights.  For instance, most AI systems share a risk for certain structural ethical issues (i.e., privacy violations, opaque/biased decision making), but some have additional, often physical consequences (such as lethal decision-making in automated vehicles, or certain AI health applications).  Generic solutions would likely cover most of what matters ethically in this structural sense, but those technologies which also carry the potential for physical or lethal harm as a result of AI decision-making should also be fitted with sector-specific or technology-specific solutions. In other words, general solutions offer a way to cover most, but not all of what matters ethically, and it is important to close this gap. |
| 9 | Given the importance of transparency across the AI lifecycle, please share your thoughts on:  a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI? |
| Transparency in AI makes it possible to discover how and why a system made a particular decision. An AI system is considered to be transparent if it is *possible to discover why it behaves in a certain way*, for instance, why it made a particular decision.  A system is explainable if the way it behaves can be expressed in plain language understandable to non-experts.  To improve explainability to non-experts to improve public trust and confidence in AI, the general public should have a basic level of understanding of how AI systems work.  We suggest that public engagement include panel debates and high quality mass media with documentaries that present emerging and AI technologies and how they work in an interesting and understandable way.  Please note that accountability is very difficult without ethical transparency. Without accountability, it is most difficult to mitigate potential AI risks and to improve public trust and confidence in AI.  From the user perspective, it is important to ensure transparency concerning the presence of a user’s interaction with an AI system, i.e., in the use of AI-enabled public services, redress etc. It should be obvious to the lay user that they are interacting with an AI system, even if this AI system has impressive communication ability.  Transparency is a basic principle, contributing to trustworthiness of AI which means it will be critical in all stages of the AI lifecycle, from design up to the maintenance stage.  Also, governance models need to consider transparency, requiring that organizations, independently of the sector in which they operate, update, and accelerate their capabilities to face challenging and fast changing conditions that AI is imposing.  Transparency is particularly important for making sense of decision making, and the basis upon which a decision was made. These matters must be communicated in a manner which people are ready to receive, ideally with public awareness cartoons or diagrams to help explain complex AI phenomena. |
| b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented. |
|  | New governance and business models, including top management and stakeholders’ engagement, should integrate formal commitment towards [AI Ethics Principles](https://www.google.com/url?client=internal-element-cse&cx=006539740418318249752:f2h38l7gvis&q=https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf&sa=U&ved=2ahUKEwiz7Ljd2KuAAxXCmmoFHdrBDiMQFnoECAMQAQ&usg=AOvVaw3filvGcOrrl6CW4pGgVz-o), especially how, in practice, they anticipate AI impacts and which mechanisms are implemented to ensure transparency.  Review of [IEEE Standard 7001TM](https://standards.ieee.org/ieee/7001/6929/) will provide guidance on where and when appropriate levels of transparency in AI should be considered and potentially mandated. Use of a ‘System Transparency Assessment’ or STA, a method for assessing the transparency of an existing autonomous system, or use of a ‘System Transparency Specification’ or STS for specifying the transparency requirements of a system prior to its development or procurement may be considered, and when and where these are appropriate to use. |
| 10 | Do you have suggestions for:  a. Whether any high-risk AI applications or technologies should be banned completely? |
| We suggest that consideration needs to be made as to the level of risk of an AI system to determine what might be unacceptable.  A strong respect for mental privacy and freedom from manipulation may mean that some emergent applications should be banned or at least seriously curtailed. There is also the emerging field of neuro rights and BCIs (brain-computer interfaces) to consider when assessing risk. |
| b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts? |
|  | As a general ethical rule, choices on banned technologies should be society-specific (reflecting cultural values), rights respecting, and should mitigate the potential fallout of techno-moral solutionism (improving individual behavior via a paternalistic application of AI) even if the benefits of said technologies seem clear and seemingly desirable.  For example, in Europe, the AI Act is based on the identification of levels of risks. Unacceptable risks related to AI based solutions or systems that can configure manipulation, social scoring or targeting vulnerable groups are proposed to be banned. |
| 11 | What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI? |
|  | Data and AI literacy is often misunderstood as a set of technical skills, limited to data management and analysis and to the development and application of algorithms. However, data and AI literacy as a future skill of the 21st century serves to promote autonomy in a modern world shaped by data and its application as well as new technologies like AI and is therefore important for all people - not only for specialists.  Data and AI literacy therefore encompasses a set of cognitive, meta-cognitive, affective, and socio-emotional competencies, which are grounded in universal moral values and enable individuals to face the challenges of digital life and adapt to its demands.  There is also a need for a standardized framework to capture at least a minimum set of foundational and cross-cutting data literacy competencies relevant for an individual, organization or system. This will help to identify clear data literacy needs, support the effective targeting of policies and programs to enable data literacy, and provide a benchmark to assess the impact of such efforts.  IEEE [P7015TM](https://standards.ieee.org/ieee/7015/10688/) Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness, establishes an operational framework and associated capabilities for designing policy interventions, tracking their progress, and empirically evaluating their outcomes. The standard includes a common set of definitions, language, and understanding of data and AI literacy, skills, and readiness.  In addition, initiatives to improve the AI ethics competency within organizations, such as explained in IEEE SA [Ethically Aligned Design for Business](https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead-for-business.pdf) which identifies the cultural and behavioral change needed within organizations, and what it takes to enable AI ethics as a core organizational competency.  Use of ‘ethics by design’ methodologies, such as IEEE [7000-2021](https://standards.ieee.org/ieee/7000/6781/)TM, to deliberately assess and address stakeholder concerns when developing or procuring AI systems, and use of IEEE [7010-2020T](https://standards.ieee.org/ieee/7010/7718/)M to concurrently assess the qualitative/quantitative ‘well-being’ impact of the proposed or actual deployment of AI systems on individuals, communities and society - with assessments wider than just derived from economic impacts.  IEEE has introduced the [CertifAIEd](https://engagestandards.ieee.org/ieeecertifaied.html?_gl=1*10cupc6*_ga*MTM0ODg2ODAxLjE2OTAzNDI2MDQ.*_ga_XDL2ME6570*MTY5MDM4Nzk1NS4zLjEuMTY5MDM4ODgzOC4zNy4wLjA.) certification program to help instill trust between organizations and the consumer. IEEE CertifAIEd affirms an organization’s commitment to human values, dignity, and well-being, and to respecting, protecting, and preserving fundamental human rights. It conveys an organization’s capability to fulfill applicable transparency, accountability, reduction of algorithmic bias and privacy requirements, fosters trust, and facilitates the adoption and use of AI solutions. It also helps to enhance confidence in public and private enterprises that wish to realize the benefits of AI ethics certification in the absence of, or as a complement to, broadly accepted and enforced regulations for AI, while mitigating risks, liabilities and adverse impacts on their reputation and market share. |
| **Implications and infrastructure** | | |
| 12 | How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia’s tech sector and our trade and exports with other countries? |
|  | Banning high-risk activities in line with protection of established human rights can also bolster normative convergence in global AI governance. This will have beneficial network effects on the Australian tech sector/ economy insofar as incentives for human rights-respecting technologies will be high, including the potential of lowering cost barriers for export of AUS technologies, i.e., due to the need for product differentiation for different regulatory environments, etc.  We posit that banning AI technologies may deter importers of such goods from trying to introduce them into Australia, even by stealth (i.e. embedded within another product). |
| 13 | What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks? |
|  | Consideration might be given to extending [Infosec Registered Assessors Program (IRAP) | Cyber.gov.au](https://www.cyber.gov.au/resources-business-and-government/assessment-and-evaluation-programs/infosec-registered-assessors-program) assessment program to encompass IEEE’s CertifAIEd program for Registered Assessors |
| **Risk-based approaches** | | |
| 14 | Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach? |
|  | A ‘risk-based approach’ should be used, in conjunction with an ‘impact based approach’, to assess potential and actual impacts on individuals, communities and society in accordance with qualitative and quantitative well-being criteria, in accordance with IEEE 7010 – 2020TM.  Note that in IEEE [7000-2021](https://standards.ieee.org/ieee/7000/6781/)TM, the risk-based approach concerns risk to ‘human values’ like fairness, inclusiveness, justice, autonomy, etc., and not just consideration of financial risk or impacts. These approaches also recommend continuous and iterative assessment of both risk and impact, prior to, during and after deployment of AI systems. |
| 15 | What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome? |
|  | Limitations include use of a standard risk-based approach in corporate governance which may involve only considering financial risk and which may only be reviewed intermittently (i.e., bi-annually). AI systems need to be continuously assessed and monitored for both risk (to human values) as well as the 12 domains of well-being impact, as described in IEEE 7010 – 2020TM.  Additionally, a conventional risk-based approach may not directly involve all stakeholders in the assessments. In order to be ethically aligned, all stakeholders need to be involved.  There is a need to consider risks to values (i.e., other than financial impact or physical harm) as it requires careful analysis which many firms/institutions likely lack the resources to conduct themselves.  Please see IEEE C[ertifAIEd](https://engagestandards.ieee.org/ieeecertifaied.html?_gl=1*1fv1z66*_ga*MTU5NDQyMTcwOC4xNjgyNTIxNzg5*_ga_XDL2ME6570*MTY5MDM5NDA4OS4xMTUuMS4xNjkwMzk0MDk0LjU1LjAuMA..) as it creates specifications for certification and marking processes that advance transparency, accountability, and reduction in algorithmic bias in Autonomous and Intelligent Systems.. |
| 16 | Is a risk-based approach better suited to some sectors, AI applications or organizations than others based on organization size, AI maturity and resources? |
|  | We posit that a risk-based approval may be particularly suitable for any sector or vertical involving public or individual safety, privacy, legal rights, human rights, health, and wellbeing. |
| 17 | What elements should be in a risk-based approach for addressing potential AI risks? |
| Development of a catalog of AI system risks and AI risks management practices is needed to address potential AI risks. |
| Do you support the elements presented in Attachment C? |
|  | Impact Assessments should be holistic, and derived from evaluating the 12 domains of well-being, as described in IEEE 7010-2020 TM.  We suggest a full examination of the [IEEE 7001 TM](https://standards.ieee.org/ieee/7001/6929/) standard for appropriate transparency requirements and assessments for each type of stakeholder, particularly to enable them to seek redress or contestability of decisions made by the AI system.  Human in the Loop and oversight assessments should be considered as outcomes to use of transparency standards as recommended by IEEE 7001TM and accountability considerations, in order to demonstrate the human operator's capacity to challenge the system. (This is also a component of CertifAIEd TM Accountability criteria).  Explainability is important to build public trust and should be based on definitions within IEEE 7001 TM.  Training and Education should involve appropriate and sufficient ‘AI Literacy’, including AI ethics readiness framework as described in ‘Ethically Aligned Design for Business’. With AI Literacy, organizations can differentiate and build competencies and methodologies to become ‘trustworthy organizations’, by promoting and incentivizing responsible innovation practices.  Monitoring and documentation is an essential component of assessment and audit practice for developing and maintaining a responsible and trustworthy AI system, and likewise essential for certification in IEEE CertifAIEd TM criteria. |
| 18 | How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication? |
|  | The AI risk-based approach of IEEE 7000-2021 TM assumes use of an organization’s own risk-based frameworks, assessment frameworks and management processes, thus reducing potential duplication. |
| 19 | How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)? |
|  | The risks of such systems are amplified due to an ability for unanticipated functions to lurk within models until elicited via a particular prompt sequence. This is especially the “case” where models may gain unexpected new capabilities as they increase in scale, which is very difficult to predict in advance.  Anticipation of the risk to human values (including environmental risks) would be facilitated by using IEEE 7000-2021 ™, by including representation of all stakeholders involved (including future generations), and mitigation to these human values during the design process. Use of IEEE 7001 TM would help anticipate the requirements for transparency of the input training datasets, particularly IP, copyright, and privacy risks as well as transparency to the public and users as to how and why recommendations and suggestions are produced and ethical decisions were made during development and design, and use of IEEE 7010-2020 TM would anticipate the impacts across 12 domains of well-being to the individual, communities and to societies.  Use of [IEEE 7002 TM](https://standards.ieee.org/ieee/7002/6898/) to improve and strengthen the frameworks and processes of an organization in dealing with privacy using a risk and mitigation-based approach throughout the AI development lifecycle from inception to retirement of the system. |
| 20 | Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? |
| There is a need for both voluntary self-regulation and for regulatory guidance regarding compliance to be established due to the potentially high risks and consequences involved with AI systems. |
| And should it apply to: |
| Both approaches can be used. Voluntary standards developed by global standards bodies like the IEEE can be referenced by legislation.    Voluntary standards are a form of ‘soft law’ that are more adaptive and responsive to change, managing AI and associated risk than legislation is able to adapt to technological changes. |
| a. public or private organizations or both? |
| An AI system, in its entire life cycle, brings potential technical and organizational risks affecting it that must be identified, documented, and accessible. A similar approach should apply to both sectors. The origin and reasons of these potential risks, responsible person, measures to be taken, and actual situation upgrading, also must be identified, and documented.  Periodically tests and checks models for ethical issues and bias must be run.  To prevent attacks on the AI System with the aim of inferred in data and information, the following measures should be taken:   * Analysis of the potential attacks with an evaluation of the possible risks, * Countermeasures to mitigate the risks, * Penetration tests have to be conducted.     Both public and private organizations.  Public organizations are accountable to their constituents and to the public, and need to demonstrate this, both in terms of wise and accountable investment of funds in AI systems, and for an alignment of ethical values such as fairness, justice, environmental responsibility, alignment with human rights as well as legality of decision making and recommendations generated by AI systems.  Private organizations are accountable to their owners or shareholders but also accountable to their customers, clients, and their supply chains. |
| b. developers or deployers or both? |
|  | Both, for developers and deployers. Accountability for a risk-based and impact-based approach needs to be consistent across the full range of the supply and value chain, from inception to procurement, integration, operation, and retirement. As an example, please see IEEE 7001 TM on Ethical Privacy which describes organizational frameworks and development across the entire AI lifecycle in regard to personal data, and from the IEEE CertifAIEd TM criteria suites, which take into account risk evaluation for all actors. |

We look forward to further discussions with the Department of Industry, Science, Energy and Resources (Australia) regarding the report on Safe and Responsible AI in Australia. If you have questions, please do not hesitate to contact Karen Mulberry at [k.mulberry@ieee.org](mailto:k.mulberry@ieee.org).