Supporting responsible AI: government discussion paper

Index:

- Introduction- 3
  - About XXX - 4
  - XXX's responsible AI - 4
  - What it does- 5
  - Risk detection in action- 5
  - Successfully regulating AI-5
- Recommendations- 6
- Definitions- 8
- Potential gaps in approaches- 8
- Responses suitable for Australia- 11
- Implications and infrastructure- 15
- Risk-based approaches- 16

INTRODUCTION

XXX is pleased to provide a response to the Commonwealth Department of Industry, Science and Resources Technology Strategy Branch, regarding its *Supporting Responsible AI: Discussion Paper* (discussion paper).

XXX is a powerful software platform that specialises in data discovery, investigation, and cybersecurity. Wi th advanced data processing capabilities, our software enables organisations to efficiently analyse, search, and extract valuable insights from vast and complex datasets, including unstructured and structured data.

XXX has developed software with comprehensive features and a user-friendly interface, which has increased the adoption of our platform across the public and private sectors. Our experience in helping customers in law enforcement, regulatory, corporate and legal organisations to uncover critical information, accelerate investigations, and mitigate potential risks, has provided us with an understanding of the benefits and potential risks of Artificial Intelligence (AI).

XXX is proud to be a sovereign company headquartered in Australia, employing over 100 highly skilled experts within Australia and 430 people globally. We are committed to fostering the development of Artificial Intelligence technology that will benefit society – which we refer to as AI for Good – whilst extending the opportunity to grow innovation in this field both at home and overseas. As an Australian business operating globally, XXX brings a unique perspective to the practicalities of global AI developments in global markets and how they can and should be most effectively applied to the domestic environment.

We readily agree that current and future regulation must address head-on the myriad concerns associated with AI technology, which have the potential to cause serious social harm, including:

- malicious abuse (disinformation/misinformation, deep fakes, fraud); and
- unintentional misuse (overreliance on 'black box' solutions and massive generic datasets that are prone to bias, data privacy and copyright violations, as well as adverse environmental impacts).

That said, if managed carefully we believe the ultimate benefit that AI technologies will deliver to humanity far outweighs the substantive challenges. AI's potential as a powerful transformative force that delivers good in the world must be thoughtfully explored and supported. For example, AI can reduce the instance of human error, can take on high-risk activities such as deep sea or space exploration, and can take on mundane and monotonous tasks without experiencing fatigue, which in a high skills-based country like Australia, can enhance national productivity and enable employers to diversify and broaden roles. AI can reduce human error or take on tasks that are impractical, impossible, dangerous or highly inefficient.

- Risk Analysis: Identify anomalies, patterns or relationships in text data across massive datasets that would be impractical or impossible for humans, for law enforcement, public safety, and corporate organisations to dynamically manage - with application across use cases as varied as

the identification of insider threats, investigations in to fraud rings and child trafficking, and earlier identification of suicide ideation.

- Information Governance: Identify personal, confidential or sensitive information to support compliance regulations, protect the privacy of citizens and organisations, and support cyber security, records retention management, and general data management efficiencies.
- Investigations & eDiscovery: AI can sift through massive datasets at speed and scale impossible for humans, to assist with and accelerate large consequential investigations and litigation efforts.
- Healthcare: Using image analysis for breast cancer detection four years before traditional methods, or the analysis of data generated by personal devices to alert people or healthcare professionals of potential health issues and risks.
- Exploration: AI enables us to use technology to explore our oceans, deep space, or other places too dangerous or inaccessible for humans.

About XXX

XXX AI is designed to perform automated data discovery and text analysis using natural language processing (NLP) and machine learning. More specifically, our AI leverages a blend of supervised[1] and u nsupervised[2] learning techniques, which allows us to take advantage of AI's scale, speed and accuracy w hile critically also maintaining a human-in-the-loop approach to address nuance, bias, or inaccuracies.

XXX AI features an easy-to-use "no-code" interface, designed to empower non-technical subject matter experts (SMEs) with the ability to build, optimise, test, and validate models. Additionally, the software critically provides dynamic and intuitive insights into how our results are generated, which enables improved risk management as well as real time feedback on the impact of any adjustments made to the models. Each element has been designed to be understood by non-technical professionals without any need for software programming, AI engineering, or data science expertise.

XXX AI serves as an enablement capability that infuses automated intelligence into XXX's suite of data processing and data intelligence solutions. However, it is important to note that our AI-enabled offerings have been purpose-built to keep humans in control at every stage of the process, including the options to restrict or withhold the AI model creation/editing features, or even to disable the AI capability completely, in the rare instances that this is required or desired.

What It Does

Given the unrelenting and exponential growth of data across business, governments, and society in general, AI is no longer a luxury but rather an essentialenabler of organisations' ability to operate effectively in the modern age. XXX AI is a user-centric utility that has been uniquely designed to make t his possible by automating the analysis of textual data at scale, for the purposes of data discovery, classification, entity extraction, pattern detection, and prioritisation. In this way, our AI dramatically reduces the amount of gruelling, time-intensive reading often required in our today's working environment, which enables customers to focus on the most cerebral and nuanced interpretation of the data - to help inform, optimise and accelerate (but not replace) their decisions and actions. The

[1] Supervised learning (or supervised machine learning) is a subcategory of machine learning and AI, defined by its use of labeled datasets to train algorithms that classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

[2] Unsupervised learning (or unsupervised machine learning) uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

technology is most effectively used in data-heavy environments that involve the analysis of thousands and sometimes millions of documents for a range of markets and use cases in legal, finance, government, law enforcement, and more.

XXX's responsible AI

XXX's AI is supported by three core guiding principles that underpin our Responsible AI development efforts:

- Explainability: Opening the 'black box' to ensure that we transparently expose the training data (data used in the creation of the models), while also providing useful insights and reasons (defensibility and repeatability) behind our results that can be easily understood and evaluated by non-technical people. By ensuring that users of our AI-powered solutions have this critical level of insight and oversight, XXX A I empowers users with the level of awareness and understanding that enables them to maintain full con trol of the AI's usage and output..

- Accessibility: Removing the 'ivory tower', putting AI into the hands of the domain experts who are typically closest to the data and the problems they are trying to solve, rather than relying on IT staff and software engineers. This not only helps to demystify and democratise this sophisticated technology, but also reduces time-consuming, error-prone handoffs and communication breakdowns that often take place between knowledge workers and technical teams.

- Specificity: In contrast to what XXX views as a concerning over-use of ever-increasing large language models (LLMs) to drive generative AI, we take a more thoughtful, deliberative, and targeted approach. Our AI is designed to be deployed in customers' secure environments and run on affordable central processing unit (CPU) based systems. Furthermore, rather than trying to be all things to all people which is a feature of some of the largest LLMs, XXX customers can quickly adjust templated models, or build new models, to fit their specific requirements and targeted use cases. Once again, the humans are in control, and the AI works to support their specific needs.

Risk-detection in action

As an example, XXX's AI technology has been used to automatically analyse social media posts as an early warning detection mechanism, to alert of potential suicidal ideation among military veterans in. This alerting mechanism demonstrated that structured natural language understanding (NLU) models could be trained by non-technical professionals to reduce the 'noise' (false positives) within the data. The reality is the model used has been more successful than highly trained suicide experts from Harvard University, given the AI's ability to be always-on and always-alert once programmed to monitor risks among veterans.

Successfully regulating AI

We believe that this current moment in AI's evolution represents an historical turning point in societal advancement where the application of AI is emerging from the shadows of science fiction into the mainstream. And as with all human innovation over the last century (from industrialisation, aviation, the invention of the combustion engine, modern healthcare, renewable energy etc.), a thoughtful and deliberate approach to some degree of acceptable risk is necessary for advancement.

As a technologically advanced nation, Australia can and should play a leading role in helping manage the risks and defining the many potential benefits AI can offer. But with this advancement comes great responsibility, and for this reason XXX believes that AI's evolution must be thoughtfully harnessed and aligned with the highest ethical standards to drive trust and empower humanity, not threaten it.

From both a regulatory and industry standpoint, if the correct approach to balancing regulation and encouraging innovation is employed, Australia has the potential to usher in a new era of digital innovation to driven by an empowered workforce that is technological powerhouse.

XXX's response to this discussion paper has been framed by the belief that AI's evolution must be thoughtfully harnessed and aligned with the highest ethical standards to drive trust and empower humanity, not threaten it. Future regulatory guardrails should ideally be developed with support from industry and ensure a balance whereby trusted outcomes don't diminish innovation.

Recommendations:

1. That the Australian Government establish a working group between relevant government agencies and Industry partners to facilitate real-time responses to issues arising from the increasing adoption of AI.
    a. The Working Group on AI should comprise of both industry experts as well as representatives from agencies such as the Australian Cyber Security Centre (ACSC) at the Australian Signals Directorate, the Cyber and Infrastructure Security Group at the Department of Home Affairs, the Office of National Intelligence and within enforcement agencies as recommended by the Attorney General's Department.
    b. The working group should be accessible to relevant Members and Senators within the Australian Parliament, with reporting of transparency, new trends advancements and potential issues submitted to Parliamentarians at regular intervals.
    c. As an additional element to the working group's mandate, consideration should be given to whether this group should also have a wider role in contributing to the evaluation of potential *benefits* of AI – and the strategic roadmap for delivering trusted AI as a genuine asset for the Australian economy.

2. The creation of a voluntary industry code.
    a. In a similar manner to the code currently in place to address misinformation and disinformation by social media platforms by the Digital Industry Group (DIGI).
    b. The code would be reviewed by the AI working group on a set, periodic basis, and work closely with relevant Commonwealth agencies to ensure broad industry compliance.

3. Reducing knowledge gaps across Australian society about the applications of AI technology (education / learning campaign to promote public trust in AI, including working with state governments to include in future school curriculum)
    a. XXX would be pleased to work closely with appropriate Australian Government departments and agencies to develop and implement a knowledge building campaign across vulnerable or targeted members of Australian society.
    b. Given the likely interplay between AI used for malicious purposes through frequently used digital platforms, (from news websites to social media), XXX would be amenable to discussing platform-based knowledge building campaigns, to equip users of those platforms with the tools to identify, report and avoid potential AI-based traps.

4. Improving interstate knowledge sharing capabilities and law enforcement collaboration across other allied markets (including Interpol or the United States' Federal Bureau of Investigation).
    a. The nature of technology-based attacks is borderless, and as the geopolitical fallout associated with the conflict in Ukraine has demonstrated, can equally present national security risks and opportunities.
    b. XXX would recommend the Government seek to share information with other trusted allies through departmental knowledge sharing, regulators, agencies, universities and research houses.
    c. Given the breadth and depth of its global customer base, XXX believes it is uniquely positioned to assist where needed to advance this initiative.

DEFINITIONS

1. **Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why? Potential gaps in approaches**

   XXX concurs with the definitions currently outlined within the discussion paper.

POTENTIAL GAPS IN APPROACHES

1. **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?**

   Regulatory and enforcement solutions regarding AI must address the purposeful and harmful intention, by malicious actors seeking to exploit Australians online.

   XXX identifies the following as potential risks within the Australian market resulting from an unregulated approach to innovation in AI:

   - Privacy breaches: There is a risk of excessive data collection beyond what is necessary
   - Algorithmic bias: untrained (or 'under-trained') data has the potential to result in tools that propagate bias
   - Exclusionary impact: the potential for the benefits and risks of AI to be unevenly applied across the community, needs to be very carefully considered and managed.
   - Job displacement: As AI and automation continue to advance, some industries and job roles are at a higher risk of being automated and whilst this transition is somewhat inevitable, it should be thoughtfully managed
   - Cybersecurity threats: Deep-learning training is compromised with intentional malicious information
   - Malevolent exploitation of AI technologies: Severe malicious use could potentially destabilise society
   - Environmental Impacts: Language models are described as "large" based on the number of values or parameters used to build them. Some of these LLMs have hundreds of billions of parameters and use a lot of energy and water to train them. For example, GPT3 took 1.287 gigawatt hours or about as much electricity to power 120 U.S. homes for a year, and 700,000 liters of clean freshwater,
   - Educational Impacts: Over-use of generative AI (and other AI capabilities around Natural Language Processing (NLP), problem solving and code generation) in schools could materially impact the analytical, critical thinking, research, and creative skills and capabilities of younger children. Meanwhile these important attributes could also suffer atrophy for older students and even young professionals who have been raised as "digital natives".

   AI however can also be used as an instrument for significant societal good, and its innovation and further application should be encouraged by regulators and enforcement agencies in Australia. For example, AI can produce significant benefits to lawmakers, regulators, healthcare practitioners, and enforcement agencies through its forensic capabilities, as well as researchers or rescuers undertaking high risk activities such as deep-sea diving or space exploration.

To mitigate these risks, XXX recommends that regulatory actions should involve establishing strong, accurate and reliable privacy data protection and privacy guidelines that reflect the same pragmatic approach being applied in the EU. The agreed guardrails that channel AI's development should be transparent, accessible and specific.

2. 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.

Given AI is dramatically fast moving and a comparatively recent and still-evolving technology, XXX's position is that any legislative or regulatory actions need to be balanced by significant industry participation.

XXX recommends the implementation of:

- A voluntary industry code, to be overseen in partnership with the Australian Government, its departments and agencies as appropriate.
- A working group with annual reviews between industry actors, academia, DIGI and its members, the Department of Industry, Science and Resources, the Department of Home Affairs, the Attorney General's Department, and including the Cyber Security Commissioner and the head of the Department of Home Affairs' Deputy Secretary of the Cyber and Infrastructure Security Group.
- Biannual reviews with a trusted government partner, such as the Australian Strategic Policy Institute (ASPI), part-funded by the Australian Government and industry to review the social applications of AI technologies.
- Broad knowledge-building campaigns across Australia to upskill individuals on the opportunities and risks within AI.
    - The limited understanding of AI applications across our society is of significant concern to companies like XXX.
    - Limited understanding, misunderstanding or significant knowledge gaps across sections of the Australian community create opportunities for malicious actors to use AI for negative or criminal outcomes.

3. Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia. Responses suitable for Australia

AI's rapid advancement represents an important opportunity for the Australian Government to drive the development and deployment of AI for the benefit of all Australians.

Principled coordination of AI governance across government can help ensure consistent approaches and facilitate the responsible development and adoption of AI in Australia.

A coordination mechanism could be implemented through:

- Establishment of a whole-of-government coordination branch housed within the Department of Prime Minister and Cabinet (PM&C), to ensure that all departments and

agencies adopt a common shared language and approach towards the use and policy frameworks relating to AI.

- Establishment of a permanent Joint Parliamentary Committee with responsibility of oversight of the Australian Government's implementation of AI, including but not limited to regulatory and enforcement mechanisms.
- Establishment of a working group: A dedicated entity can coordinate AI-related policies, regulations, and guidelines across different government departments and agencies.
- Sharing knowledge and best practices: Facilitating knowledge exchange and collaboration among government entities can help identify emerging risks, address regulatory gaps, and promote consistent standards.

RESPONSES SUITABLE FOR AUSTRALIA

1. 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? Target areas

XXX believes there is much to learn from the regulatory models being implemented by the EU (a 'rights based' regulatory approach) and United States (a 'market based' regulatory approach), given the divergence in both jurisdictions' approaches. While both models will favour a risk-based approach, the reality is their enforcement agencies, mechanisms and existing legislation governing digital and data processes and security, critical infrastructure and overall threat landscape and capability account for the different regulatory models.

XXX believes an Australian working group examining both models, in consultation with experts in existing regulation within Australian Government departments and aided by third-party experts in the Australian Strategic Policy Institute, would be best placed to provide the most holistic review on both models. This would allow for an expert Australian perspective on which legislative or regulatory or enforcement mechanisms would best work within existing frameworks, without creating additional difficulties through unnecessary red tape for Australian industry.

TARGET AREAS

1. Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

While different risk assessments may be legitimately applied to AI in the context of private and public sector use, we would encourage a harmonised approach where those risks are the same.

In this context, the EU AI Act focuses more heavily on a tiered approach to the risks associated with the AI application itself – categorising risk based on the intended use or purpose of the AI along the following lines:

a. Prohibited AI use – eg subliminal manipulation
b. High-risk AI systems – eg critical infrastructure
c. Limited-risk AI systems – detect emotions, generate content
d. Minimal-risk AI systems – not covered above.

2. How can the Australian Government further support responsible AI practices in its own agencies?

In a recent report by CSIRO on helping Australian businesses navigate responsible AI principles, the steps to implement Australian Government's AI ethics principles are highly beneficial to support responsible AI practices and should be taken into consideration. These include:

- Establishing internal guidelines and policies: Developing clear guidelines and policies for AI use within government agencies can ensure adherence to ethical principles, transparency, and accountability.
- Providing training and resources: Offering training programs and resources on AI ethics, responsible AI development, and risk management can empower government employees to make informed decisions and implement best practices.

- Encouraging collaboration and knowledge sharing: Facilitating communication and collaboration among different government agencies can promote the sharing of experiences, lessons learned, and best practices in AI governance.
- Conducting audits and assessments: Regular audits and assessments of AI systems used by government agencies can ensure compliance with ethical guidelines, data protection regulations, and algorithmic fairness principles.
- Engaging with external stakeholders: Seeking input from external experts, researchers, and civil society organizations can provide valuable insights and diverse perspectives on responsible AI practices.

3. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

- In the case of no-risk and low-risk applications and use of AI, more generic approaches can and should be applied, including public education efforts, establishing business and govt best practices, and broad legislative guidelines. For example, it would benefit society to be informed about what AI is, how it can be used, how to avoid improper use and where to seek guidance and support. This may include providing insights on the differences of the various kinds of AI, including ChatGPT guidelines such as "never upload private information to an application like ChatGPT or Bard" and "be mindful that the output from these systems may infringe upon other's copyrighted works."
- Meanwhile, for mid-high-risk or banned uses, more targeted, technology-specific approaches would be highly recommended. This may include:
    - The official classification of AI systems to differentiate them from each other, i.e., Generative AI, Natural Language Generation (NLG), Transformer-based AI, Pre-Transformer-based, etc. The classification will offer a way to assess the capabilities and differentiated risks and opportunities associated with each, which will avoid painting AI with a single brush.
    - Given Generative AI and rapidly growing LLMs, there is a distinct need to address the risks and challenges of these solutions as a distinct group. Concerns here abound and need focused attention and meaningful solutions including:
        - Watermarking of output to offer validation of how the writing was "authored"
        - Copyright and trademark infringement issues surrounding the unauthorised use of the works of thousands of creators.
        - Misinformation/Disinformation/Deep Fakes
        - Citing sources for material produced (visual, audio-based, and textual)
        - Warning labels that give businesses and the public awareness of the potential risk and challenges of usage
        - Guidance for businesses, government and law enforcement that work in particularly high-risk segments of society including legal, finance, insurance, and healthcare.

More broadly, some more specific risks and AI technology solutions we have encountered include:

- Ensuring data privacy: Implementing robust privacy protection measures, such as anonymisation techniques or strict access controls, can address privacy risks associated with AI systems across various applications.
- Mitigating algorithmic bias: Developing standardised architectural methods to identify and reduce bias in AI algorithms can be applied across different sectors, ensuring fair outcomes in decision-making processes.
- Enhanced forensic analysis: AI-enabled forensic technologies and processes can significantly reduce the time it takes law enforcement agencies to complete forensic analysis, increase the average number of cases completed per analyst, and expedite the completion of criminal cases within months versus years using traditional methods.
- Fraud investigation: Protecting Australia's consumers and businesses against online scams requires investigations into large volumes of digital evidence that can be confidently expedited using AI-enabled approach.
- Responding to Data Breaches: AI-powered text mining automates the identification of PII/PHI and other sensitive information, reducing time spent manually reviewing documents by 600 percent or more by minimising false positives, while surfacing items that could have missed by human error. This enables organisations to accelerate the sending of notifications to at risk individuals and organisations, and remain in alignment with compliance requirements.
- Insider Threat Triage: AI solutions can comb through large volumes of documents to spot risky content that helps validate whether an investigation of an individual is warranted, enabling teams to respond quickly to minimise potential damage, whilst also sparring unnecessary investigations of valuable staff.
- Controlled Unclassified Information (CUI) Identification: Sophisticated AI solutions can be leveraged to discern nuances in documents, traditionally relegated to careful human review, to automate the screening of data at varying degrees of classification/sensitivity, including CUI.
- Data Discovery: Generic forms of AI can be leveraged successfully to spot interesting, helpful or previously unknown patterns in large datasets a wide range of traditionally manual tasks such as content clustering/grouping for data organisation/management, Redundant, Obsolete, and Trivial (ROT) remediation, records retention management, research and risk analysis of datasets.
- Public or Personal Safety: Automated analysis of electronic communications or social media posts can help to identify things like risky or criminal behaviour, hate speech, sexual harassment, depression, excessive drug use, or suicide ideation.

4. Given the importance of transparency across the AI lifecycle, please share your thoughts on:
    a. where and when will transparency be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

    Transparency should be prioritised in the following scenarios:
    - Critical decision-making: When AI systems are used to make decisions that significantly impact individuals' lives, such as in criminal justice, finance, or healthcare, transparency is crucial to ensure accountability and fairness.

- Data-driven decisions: Transparency is valuable when AI algorithms rely on large datasets to make predictions or recommendations, as it helps identify potential biases, errors, or discriminatory outcomes.
- Understandability (or 'explainability') of AI decisions: In contexts where human oversight and intervention are necessary, transparency enables humans to understand and interpret AI-generated results, enhancing trust and facilitating error correction.

b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

In our recommendations, XXX has noted that a sustainable model of transparency reporting to relevant parliamentary and regulatory bodies should be implemented.

To facilitate the industry-wide transparency reporting which could be undertaken by a working group, individual companies using, deploying or otherwise interplaying with AI, should be tasked with maintaining annual reporting and record keeping to assist the working group in providing the most up to date information.

Mandating transparency requirements across the private and public sectors can be implemented by:

- Requiring documentation: Organisations could be mandated to maintain comprehensive documentation of their AI systems, including data sources, model architectures, and decision-making processes, allowing external audits and assessments.
- Understandable standards (directed at 'explainability'): Regulators can establish guidelines or standards to ensure that AI algorithms provide explanations or justifications for their decisions in a manner that is understandable to humans.
- Algorithmic impact assessments: Requiring organisations to conduct assessments of the potential impact of their AI systems on individuals' rights, social values, and well-being can enhance transparency and accountability.

5. Do you have suggestions for:
   a. Whether any high-risk AI applications or technologies should be banned completely?

AI is still in its relative infancy, and it is therefore difficult to define what high-risk iterations of the technology constitute. However, areas that should be considered high-risk and therefore banned include:

- AI systems that enable or facilitate human rights violations, discrimination, or harmful propaganda. This is a key area of attention under the EU AI Act.
- AI technologies with significant safety risks that cannot be effectively mitigated.

b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

Incorporating this into the working group's review to the Australian Government after a two-year period, within a fixed-term reporting period, the following criteria and

requirements should be considered for identifying AI applications or technologies that warrant banning, along with their specific contexts:

- Potential for severe harm: Assessing the potential risks and negative consequences associated with the AI application or technology, especially in terms of safety, privacy, security, or societal impact.
- Lack of ethical considerations: Evaluating whether the AI application or technology aligns with ethical principles, such as fairness, transparency, accountability, and respect for human rights.
- Public interest: Considering the broader societal implications and public opinion surrounding the AI application or technology, particularly in cases where its deployment might significantly affect individuals or communities.

6. What initiatives or government actions can increase public trust in AI deployment to encourage more people to use AI? Safe and responsible AI in Australia consult.industry.gov.au/supporting-responsible-ai 35 Implications and infrastructure

- Transparent and understandable AI systems: Promoting the development and adoption of AI technologies that provide clear explanations of their decision-making processes, ensuring accountability and reducing opacity.

- Robust data protection and privacy regulations: Implementing strong regulations and safeguards to protect individuals' data, ensuring proper consent, and preventing unauthorised access or misuse.
- Independent auditing and assessments: Encouraging third-party audits and assessments of AI systems to verify their compliance with ethical guidelines, fairness, and safety standards.
- Public awareness and education: Conducting public campaigns, workshops, and educational programs to inform the public about AI technologies, their benefits, and potential risks, fostering informed discussions and understanding. This could leverage the results of deliberate and targeted programs which can be used to benchmark and demonstrate improved public service efficiencies from AI – such as reduced human error, reduced public service response times, improved and more inclusive hiring practices.
- Collaborative governance: Engaging citizens, civil society organisations, and industry stakeholders in policy discussions and decision-making processes, allowing for diverse perspectives and building public trust through inclusive governance.

IMPLICATIONS AND INFRASTRUCTURE

1. 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?

XXX understands the challenge governments face in balancing technological innovation, trade and investment priorities, and securing the welfare and freedoms of its citizens. The impacts can vary depending on the specific bans and contexts, but may include:

- Technological innovation: Banning certain high-risk activities may restrict innovation in the development and deployment of AI technologies, potentially impeding the growth of the tech sector and limiting Australia's competitiveness in global markets.
- Trade and international relations: Bans on specific AI technologies or applications could impact international collaborations and trade agreements, as other countries may have differing regulatory approaches or rely on such technologies for their own operations.
- Ethical reputation: Banning high-risk activities associated with ethical concerns can enhance Australia's reputation as a responsible and ethical player in the global tech industry, potentially attracting partners and investments aligned with such values.

2. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks? Risk-based approaches

XXX proposes the following changes:

- Standards and certification frameworks: Developing or updating standards and certification programs specific to AI technologies, ensuring compliance with ethical guidelines, safety standards, and best practices.
- Regulatory oversight and enforcement: Strengthening regulatory bodies and processes responsible for monitoring and enforcing AI-related regulations, ensuring that conformity assessments are conducted and addressing non-compliance appropriately.
- Collaborative partnerships: Enhancing collaboration between government agencies, industry, and research institutions to develop conformity assessment methodologies and assurance frameworks specific to AI technologies.

RISK-BASED APPROACHES

1. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

Governing AI through a holistic risk-based approach provides regulators and enforcement agencies w ith the most practical means of addressing potential AI risks in the short to medium term and is the a pproach XXX recommends the Australian Government aim to adopt in concert with the recommendations outlined at the beginning of our submission.

2. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

The main benefits of a risk-based approach include:

- Focus on high-risk areas: Identifying and addressing the most significant risks associated with AI technologies, providing targeted mitigation measures and regulatory scrutiny where they are most needed.

- Flexibility and adaptability: Allowing regulations and governance measures to evolve as AI technologies advance and new risks emerge, ensuring continued relevance and effectiveness.
- Efficiency and resource allocation: Optimising the allocation of regulatory resources and efforts by prioritising areas with the highest potential risks, minimising unnecessary regulatory burdens on low-risk applications.

Limitations of a risk-based approach can include:

- A core limitation of a purely risk-based approach is the high opportunity costs associated with the lost potential benefits and opportunities associated with a particular AI. That is, by focusing solely on the risks, many good ideas and solutions to a given challenge may be lost in the process by virtue of the risk they pose in a statistically narrow set of circumstances. However, this limitation can be avoided by considering opportunity costs as one of the risk factors analysed.
- Subjectivity and interpretation: Assessing and quantifying risks can involve subjective judgments, requiring careful consideration of potential biases and transparency in decision-making.
- Emerging risks and uncertainties: Identifying and addressing novel or unforeseen risks can be challenging within a risk-based approach, as they may not fit into existing risk frameworks or regulatory categories.

3. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

   At this stage of its development, AI is still overall a comparatively nascent technology, the implications and applications of which are still to be fully understood. Given their resources to respond to emerging challenges and developments, large companies with their more extensive resources and assets, have to date shown an appetite to respond relatively to quickly to significant risks.

   XXX is aware the Australian Government has already considered the likelihood of challenges faced by organisations of different size and scale as part of the work it undertook to develop the Critical Infrastructure Acts and regulations, and in appointing a Cyber Security Coordinator. The byplay between AI and other critical digital communications technologies and services will likely require a response of scale from regulators to support sole traders, smaller organisations, or organisations deploying AI to address their business challenges.

   It is likely that within a few years of all actors achieving a significantly greater understanding of the opportunities and challenges present in AI, legislative or regulatory actions will be able to better govern the technological capability in a comprehensive manner.

4. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

- Establishing standardised frameworks and methodologies: To assess and quantify risks associated with AI applications, considering factors such as data quality, model performance, potential harms, and mitigating measures.
- Compliance requirements: Defining clear compliance requirements for organisations based on risk levels, ensuring they adopt appropriate risk mitigation strategies, such as testing, auditing, or third-party assessments.
- Monitoring and reporting mechanisms: Implementing mechanisms to monitor the ongoing performance and risks of AI systems, requiring organisations to report incidents, algorithmic updates, and compliance measures.
- Enforcement and sanctions: Outlining enforcement measures and potential sanctions for non-compliance with risk-based regulations, providing incentives for organisations to prioritise risk mitigation.

5. 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 last 12 months has seen AI move from being a conceptual emerging risk on the Chief Risk Officer's future radar to a fast-moving risk which requires active management today.

Given the uncertainty of the technology and its applications, it can be tempting to create a bespoke AI risk framework to assess and manage risks. We believe it is best to integrate AI risk management into your existing risk management framework. For example, it can be helpful to break down this topic into more manageable components:

- Strategic Risks: risks that AI technology may impact the short, medium- and long-term objectives of the business. For example, what is the risk that AI technology may disrupt the industry, impact the value of your technology intellectual property, or change consumer habits. Strategic risks are best tackled as part of the strategy setting and refresh processes.
- Operational & Compliance Risks: risks associated with using AI tools and technologies as part of business processes. For example, cyber and information security risks or intellectual property ownership risks. Operational and compliance risks can be considered as part of functional or departmental level risk assessments.
- Product Risks: risks associated with selling or embedding AI technologies into your own products. For example, product functionality, ethical, bias, and black box risks. These risks are best addressed as part of existing new products and change risk management processes.

As with all good risk management, the objective is to identify and understand risks and uncertainties to make informed risk management judgments and decisions. Being clear on ownership and accountability is a crucial building block. Assigning ownership for AI risk to a specific person or Committee can help ensure that it has the focus it needs, avoid potential duplication of efforts and ensure that AI risks and opportunities can be navigated responsibly.

6. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

A risk-based approach can be applied to general-purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs), by considering potential risks associated with their deployment and use. Risk assessments could involve examining factors such as the models' training data, potential biases, interpretability, the potential harm caused by incorrect outputs, and the context of deployment. Mitigation strategies might include external audits, fairness checks, or user feedback loops to monitor and address risks associated with the use of such systems.

For guidance and context, we suggest referencing and exploring the EU's AI Act, with particular emphasis on their four-tiered risk system:

- Banned Uses: subliminal manipulation; social scoring by public authorities; real-time biometric ID systems for law enforcement.
- High Risk: critical infrastructure safety; worker management; essential services (e.g., credit scoring); law enforcement
- Limited Risk: interact with humans, detect emotions, or generate or manipulate content that are otherwise not high-risk or prohibited uses.
- Minimal Risk: include all other AI systems that do not meet these criteria.

7. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:
    a. public or private organisations or both?
    b. developers or deployers or both?

We have first-hand experience across the globe of the various approaches taken by our customers to the responsible application of AI. As a starting point, we believe that appropriate consideration should be given to mandating obligations through thoughtful, principles-based regulations.

The approach of principles-based regulation will be important to enable Australia's legislative expectations to accommodate the exponential speed and changes that will continue to be a feature of AI for the foreseeable period.

In the initial phase – while regulators and the industry are developing a deeper understanding of the potential benefits and risks of AI - it will be important to ensure the right balance is struck between innovation and regulation. We have seen this managed effectively across customers and industries, via an emphasis on an 'if-not-why-not' approach to compliance - where organisations report if they are complying with the new obligations and if not, why so. This might come in the form of the if-not-why-not compliance with an agreed industry code, the framing of which could occur in the relevant legislation.

The learnings from this initial stage, can then be sourced to determine whether a more strident or technically specific ('black-letter') regulatory approach should be pursued in phase 2.

As part of this initial approach (and potentially beyond), we would also encourage ongoing government investment in programs to deliver organisational certification evidencing (e.g.) compliance with prescribed standards. This would encourage industry-driven leadership and create an appropriate opportunity for organisations that take responsible AI seriously, to effectively signal to the market.

We believe this approach should be taken to both the public and private sectors. And should be tailored but applicable to both developers and deployers.