IBM Government and Regulatory Affairs

Consultation Team
Safe and Responsible AI in Australia: Discussion Paper
Department of Industry, Science and Resources
Email: consult.industry.gov.au/supporting-responsible-ai
26 July 2023

Dear Team

IBM Submission to the 'Safe and Responsible AI in Australia: Discussion Paper'

Thank you for the opportunity to comment on the above Discussion Paper. IBM is one of the world's most enduring and innovative technology companies. Operating for over a century, and in Australia for over 90 years, IBM integrates technology and expertise, providing infrastructure, software (including market-leading Red Hat) and consulting services for clients as they pursue the digital transformation of the world's mission-critical businesses. IBM has also been at the forefront of AI technologies since the late 1950's and has developed a comprehensive approach to ensuring that trust and responsibility are at the heart of its development and deployment of AI for both IBM and its clients.

At the outset, IBM would like to make some general remarks about artificial intelligence (AI) and then address some specific questions outlined in the Discussion Paper.

IBM's Approach to AI

While IBM is not a consumer-facing company, we play a critical role in helping business clients use AI to make their supply chains more efficient, modernize electricity grids, and secure financial networks from fraud. IBM's suite of AI tools, called 'IBM Watson' after the AI system that won the US television quiz show 'Jeopardy!' more than a decade ago, is widely used by enterprise customers worldwide. Just recently we announced a new set of enhancements, called 'WatsonX', designed to make AI even more useful to our clients in their business operations.

IBM has strived for more than a century to bring powerful new technologies like artificial intelligence into the world responsibly, and with clear purpose. We follow long-held principles of trust and transparency that make clear the role of AI is to augment, not replace, human expertise and judgement.

It's often said that innovation moves too fast for government to keep up. But while AI and recent developments in foundational models have given rise to greater public attention on AI, the time for government to play its proper role has not passed us by. This period of focused public attention on AI is precisely the time to define and build the right guardrails to protect people and their interests. Our recommendations in relation to those guardrails are outlined below.

The Concept of 'Precision Regulation' and Five AI Policy Pillars

AI is just a tool, and tools can serve different purposes. For example, a wrench can be used to assemble a desk at home or construct a commercial aeroplane, yet the rules governing those two end products are not primarily based on the design or the features of the wrench – they are based on its use. Accordingly, we urge the Australian Government to adopt a "precision regulation" approach to AI. This means establishing rules to govern the deployment of AI in specific use-cases, not regulating the technology itself. A precision regulation approach strikes an appropriate balance between protecting people from potential harms and preserving an environment where innovation can flourish. In the context of AI, this approach is based on the adoption of five key principles as outlined below:

IBM's Five AI Policy Pillars:

1. **Designate a lead AI ethics official**. Providers and owners of AI should designate a person responsible for trustworthy AI, such as a lead AI ethics official. This person would be accountable for internal guidance and compliance mechanisms, such as an AI Ethics Board, that oversee risk assessments and harm mitigation strategies. As the complexity and potential impact of AI systems increases, so too must the accountability embraced by different organizations providing various functions in the AI lifecycle. A market environment that prioritizes the adoption of lead AI ethics officials, or other designated individuals, to oversee and manage this increasing complexity could help to mitigate risks and improve public acceptance and trust of these systems, while also driving organisations' commitment to the responsible development, deployment, and overall stewardship of this important technology.

2. **Different rules for different risks.** All entities providing or owning an AI system should conduct an initial high-level assessment of the technology's potential for harm. As noted previously, such assessments should be based on the intended use-case applications, expected end-users, how reliant the end-users would likely be on the technology, and the level of automation. Once initial risk is determined, a more in-depth and detailed assessment should be undertaken for higher-risk applications. In certain low-risk situations, a less in-depth appraisal would likely suffice. For those high-risk use-cases, the assessment processes should be documented in detail, be auditable, and retained for a minimum period of time.

3. **Be transparent about using AI**. Transparency breeds trust; and the best way to promote transparency is through disclosure. Unlike other transparency proposals, this approach does not entail companies revealing source code or other forms of trade secrets or IP. Instead it focuses on making the purpose of an AI system clear to consumers and businesses. Such disclosures, like other policy imperatives here, should be reasonably linked to the potential risk and harm to individuals. As such, low-risk and benign applications of AI may not require the same type or level of disclosure that higher-risk use-cases might require.

4. **Ensure your AI is explainable.** Any AI system on the market that is making determinations or recommendations with potentially significant implications for individuals should be able to explain and contextualize how and why it arrived at a particular conclusion. To achieve that, it is necessary for organizations to maintain audit trails surrounding their input and training data. Owners and operators of these systems should also make available – as appropriate and in a context that the relevant end-user can understand – documentation that detail essential information for consumers to be aware of, such as confidence measures, levels of procedural regularity, and error analysis.

   In this regard, IBM has adopted the use of **AI Factsheets** (a similar concept to nutrition information labels but for AI) to help clients and partners better understand the framework, operation and performance of the AI models that we create. Further information on IBM Fact Sheets is outlined below in response to question 9.

5. **Ensure AI is tested for bias.** All organizations in the AI developmental lifecycle have some level of shared responsibility in ensuring the AI systems they design and deploy are fair and secure. This requires testing for fairness, bias, robustness and security, and taking remedial actions as needed, both before deployment and after it is operationalized. Owners should also be responsible for ensuring use of their AI systems is aligned with laws regarding anti-discrimination, public safety, privacy, financial disclosure, consumer protection, employment, and other relevant legislation depending on the use-case.

   For many use-cases, owners should continually monitor, or retest, the AI models after the product is released to identify and mitigate against any machine-learning resulting in unintended outcomes. Policies should create an environment that incentivizes both providers and owners to do such testing well. In IBM's view this can be achieved without creating a set of new and AI-specific regulatory requirements, but rather by adhering to a set of agreed-upon definitions, best practices, recommendations and globally developed standards appropriate to the use of AI in each particular industry or use in a manner that promotes compliance.

Question 1. Do you agree with the definitions in this Discussion Paper

The definitions outlined in the discussion paper are a very good starting point for a public dialogue on this important issue.

Question 2. What potential risk from AI are not covered by Australia's existing regulatory approaches? Do you have any suggestions for possible regulatory action to mitigate these risks?

The Discussion paper provides a good overview of the current and widely recognised risk categories that AI raises. In addition, the recent Report of the Human Technology Institute at the University of Technology, Sydney (UTS) on the *'State of AI Governance in Australia'* provides a useful inventory of laws and regulations that already apply in Australia in relation to the use of AI. In a chapter devoted to current legal obligations of Australian organisations using AI, the HTI Report outlines a range of existing laws, largely technology neutral, that may currently apply in the areas of privacy and data use, consumer protection, cyber security, work health and safety, anti-discrimination, legal duties of care and negligence law. The Report states;

> *"Unfortunately, the extent and nature of existing legal applications applying to AI systems are not well understood by corporate leaders responsible for developing and deploying AI systems today. Despite increasing enforcement activity by regulators, HTI research reveals that corporate leaders tend to see the lack of AI-specific regulation as indicative of an 'AI Wild West'.*

The fact that corporate leaders may not fully appreciate that current laws do apply to their use of AI is no reason to introduce broad-based AI specific laws, particularly given that a technology neutral approach has in the main served Australia well. That is not to say that some amendments, particularly to definitions in existing laws, may be required in specific circumstances to ensure that AI applications are covered in particular sectors. The critical issue seems to be a clear need to provide education to corporate leaders and organisations regarding the best way to establish a governance framework for address their use of AI, and outline their legal obligations (as they exist today or may be amended in the future) in relation to their particular use of AI.

Question 3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia?

IBM supports the program being undertaken by the National AI Centre, an entity within the CSIRO, which aims to bring together partners from government, industry and the research sector to boost exploration and adoption of AI in Australia. It has established the Responsible AI Network (RAIN) which is a cross-ecosystem collaboration aimed at uplifting the practice of responsible AI across Australia's the commercial sector.

In addition to the NAIC, the RAIN comprises the Australian Industry Group (Ai Group), Australian Information Industry Association (AIIA), CEDA, CSIRO's Data61, Gradient Institute, Standards Australia, The Ethics Centre, The Human Technology Institute at UTS, and the Tech Council of Australia. RAIN aims to provide curated advice and best practice guidance within the six actionable pillars of Law, Standards, Principles, Governance, Leadership and Technology in order to promote the responsible use of AI in Australia. The NAIC and its RAIN partners is also considering strategies to help inform and educate corporate leaders about the best ways to implement an appropriate governance system for their organisations use of AI, including liaising with the Australian Institute of Company Directors.

IBM strongly endorses these types of strategies as important non-regulatory initiatives that support and help drive responsible AI practices in Australia.

Question 4. Do you have suggestions on the coordination of AI Governance across Government?

IBM recommends that the governance regimes recommended by Federal, State and Territory governments need to be as consistent as possible and thereby avoid unnecessary and additional duplication and costs to organisations doing business across the country and at different levels.

At present, the high-level principles adopted and recommended by various governments and industry in Australia are largely aligned with the principles adopted by responsible companies including IBM which promotes a trust and transparency approach to AI. However, as more detailed governance regimes emerge, it will be important to ensure that areas of significant divergence do not arise. The National Cabinet, supported by the National AI Centre, may be a good vehicle for ensuring dialogue and discussion to promote consistency of AI Governance recommendations across the country.

Question 5. Are there any governance measures being undertaken or considered by other countries that are relevant, adaptable and desireable for Australia.

Many countries around the world are looking at governance issues that arise in relation to AI. The European Union, for example, is looking at a comprehensive and centralised regulatory regime. In contrast, Singapore and Japan have proposed soft law frameworks that rely on collaboratively developing rules and regulations with industry and others, while leveraging existing laws and regulatory authorities/agencies to balance the benefits of innovation against potential harms to individuals. The National Institute of Standards and Technology (NIST) in the US has developed an AI Risk Management Framework as part of a strategy to create definitions, benchmarks, frameworks and standards in this area.

IBM supports these co-regulatory approaches with industry, and also encourages government to incentivize developers and deployers to voluntarily embrace globally recognised standards, as they develop.

The UK is also developing a regulatory regime that is risk based and context-specific. The UK's 'context-specific' approach also empowers regulators to weigh the risk of using AI against the costs of minimising opportunities to do so. In other words, the potential costs of *not* using AI is also a factor to be considered in the risk-based approach. To achieve the level of 'context-specificity' required, the UK is proposing to empower existing UK regulators to apply these principles to their particular industry rather than develop a 'mega-AI regulator'. IBM sees merit in this approach.

Question 6. Should different approaches apply to public and private sector use of AI technologies?

IBM supports a regulatory regime that is based on 'levels of risk' and the 'specific use-case' of AI. If this approach is adopted, there is likely no need to differentiate as a matter of principle between the public and private sector. Having said that, there may be some exceptions required, given the unique missions and characteristics of public sector entities versus private sector counterparts.

Question 7. How can the Australian Government further support responsible AI practices in its own agencies?

As discussed above, many governments around the world are examining ways to promote responsible AI practices. Closer to home, the NSW Government has adopted a sensible approach to drive responsible and ethical AI practices that are mandatory for NSW Government agencies in relation to AI. These are outlined in the NSW Government's AI Strategy and associated ethical principles which can be found at:

https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-strategy

https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy/mandatory-ethical-principles

IBM recommends consideration of the NSW Government approach.

Question 8. In what circumstances are generic solutions to the risks of AI most valuable? In what circumstances are technology-specific solutions better?

The Australian Government has already adopted its '8 AI Ethics Principles', which outline a generic approach to the challenges of AI and IBM supports this approach at a high level.

However, in terms of specific applications, IBM supports a regulatory regime that is based on 'levels of risk' and the 'specific use-case' of AI. These specific use cases may be in the areas of medicine, human resources, financial services, public and workplace safety and many others, and the level of regulation that is appropriate may vary widely. For example, the technology used in driverless cars on public roads may be similar to the technology used in an autonomous robot stacking boxes in a warehouse, but most would agree the former requires a far higher level of regulatory oversight than the latter.

Question 9. Please share your thoughts on the importance of transparency across the AI lifecycle:

9A) Where and when is transparency most critical and valuable to mitigate potential AI risks and to improve public trust and confidence?

At the outset, IBM believes that transparency is fundamental to building trust and confidence in the use of AI and the best way to promote transparency is through disclosure. Transparency in AI should focus on making the purpose of an AI system clear to consumers and businesses. Such disclosures, in line with our other recommendations, should be reasonably linked to the potential risk and harm to individuals. As such, low-risk and benign applications of AI may not require the same level of disclosure as higher risk use cases should reasonably require.

IBM Fact Sheets

AI transparency refers to the disclosure of information related to the development and/or deployment of AI systems. Examples of this information include: what data is collected? how it will be used and stored? who has access to it? What are the test results for accuracy, robustness and bias? Importantly, transparency does not, and should not, entail companies revealing source code or other forms of trade secrets or IP, but instead it focuses on making the purpose and the properties of an AI system clear to users. Different users will need different types of information. Likewise, different AI applications or use cases will implicate different information needs. In order to help developers think about what, and how, to disclose relevant information, IBM has developed the concept of an AI FactSheet.

IBM's AI FactSheets are not meant to explain every technical detail or disclose proprietary information, such as source code or other trade secrets, about an algorithm. Rather, the goal is to promote human decision-making in the use, development, and deployment of AI systems, while also accelerating developers' education on AI ethics and their broader adoption of the concepts of transparency and documentation.

We also recommend that the Australian Government should work with industry and other partners to:

- Strengthen mechanisms for global coordination on AI transparency-enabling best practices;
- Use multistakeholder environments to drive consensus around clear, consistent and global standards, along with best practices for AI transparency by documentation; and
- Explain and educate regulators on how AI transparency tools can help them better meet their goals of protecting consumers and citizens.

9B) Should transparency requirements be mandated across the public and private sectors?

Depending on the specific transparency obligations proposed, it may make sense to differentiate obligations for private sector actors versus the public sector – especially accounting for the sensitivities related to the protection of trade secrets. However, both the public and private sectors should both be held to a risk-based approach to governing AI, drawing distinctions between AI systems based on use-cases and applications.

Question 10

A): Do you have any suggestions for whether any high-risk AI applications or technologies should be banned completely?

Since early 2020 IBM has firmly opposed the use of any technology, including facial recognition technology for mass surveillance, racial profiling, violations of basic human rights and freedoms, or any purpose which is not consistent with IBM own values and Principles of Trust and Transparency. IBM urges the Australian Government to adopt a similar position in regard to the use of AI for these purposes.

IBM's leadership position on the issue of facial recognition is part of a larger commitment to advancing the social and public policy dialogue about the potential impacts of advanced technologies. Its leadership in the field of AI Ethics in particular is well known. IBM played a key role in shaping and were one of the first two signatories to the Vatican's Rome Call for AI Ethics, we partnered with the University of Notre Dame to establish a first-of-its-kind research lab dedicated to establishing best practices in technology ethics, and IBM continually provides expertise and guidance to help policy makers grappling with questions posed by emerging technologies. We believe that technology can have a positive impact on society, but only if its deployed responsibly.

Consistent with our support for what we call ‘Precision Regulation’ IBM urges Governments around the world to place the tightest restrictions be placed on end uses and end users that pose the greatest risk of societal harm. In practical terms, we believe that controls on facial recognition should:

- focus on facial recognition technologies that employ “1-to-many” matching end uses, the type of facial recognition system most likely to be used in mass surveillance systems, racial profiling or other human rights violations. These systems are distinct from “1 to 1” facial matching systems, such as those that might unlock your phone or allow you to board an airplane – in those cases, facial recognition is verifying that a consenting person is who they say they are. But in a “1-to-many” application, a system can, for example, pick a face out of crowd by matching one image against a database of many others;
- limit the export of “1 to many” systems by controlling exports of both the high-resolution cameras used to collect data and the software algorithms used to analyize and match that data against a database of images; and
- restrict access to online image databases that can be used to train “1 to many” facial recognition systems, where explicit consent of the individual in the image for its use may be unclear or non-existent.

Question 11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI.

See response to question 3 above.

Question 12: How would banning high-risk activites (like social-scoring or facial recognition technologies in certain circumstances) impact Australia’s tech sector and our trade and exports with other countries.

Developing technologies for use-cases that are contrary to internationally recognised human rights, as discussed in our response to Question 10, is not an industry that we believe is in Australia’s best interests to cultivate, nor to support other countries to develop. Accordingly, we urge the Australian Government to limit the ability of certain foreign governments to obtain the large-scale computing components required to implement an integrated facial recognition system.

We also urge it to consider the recent human rights record of a potential country of export and, where appropriate, place the strictest controls on export of facial recognition technology, especially “1 to many” matching systems, on countries with a history of human rights violations or misuse of such technology.

Further, in partnership with Australia’s allies, we recommend that the Australian Government use mechanisms, such as the Waasenaar Agreement, to limit the ability of repressive regimes to simply obtain these technologies from other countries where no such controls are in place.

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

IBM supports the work that is being done by Standards Australia and other international, industry led standards bodies urges the Australian Government to continue to fully participate in these global processes to ensure that the Australian AI industry is fully aligned and can interoperate with global technology developments in this area. Ideally the Australian Government should promote the use of these international standards across the Australian AI eco-system.

Risk Based Approaches

Question 14. Do you support a Risk-Based approach to AI?

IBM supports a risk based approach to AI, as outlined in other parts of this submission.

Question 15. What do you see as the main benefits or limitations of a Risk-based approach to AI?

A risk-based approach to AI regulation allows for a balanced and proportionate response to the use of AI, based on its potential impact. For simple example, the impact and risks involved in using an AI solution based on facial recognition to unlock a mobile phone are comparatively minor and therefore this use of AI should not be regulated. In contrast, using an AI solution based on facial recognition to arrest a person suspected of committing a crime can be extremely significant and therefore should be very strictly regulated.

For higher-risk AI use-cases, IBM believes that companies should be required to conduct impact assessments showing how their systems perform against tests for bias and other ways that they could potentially impact the public, and attest that they have done so. Additionally, bias testing and mitigation should also be performed in a robust and transparent manner for certain high-risk AI systems, such as law enforcement use-cases. These high-risk AI systems should also be continually monitored and re-tested by the entities that have deployed them.

IBM recognizes that certain AI use-cases raise particularly high levels of concern. Law enforcement investigations and credit applications are two often-cited examples. By following the risk-based, use-case specific approach at the core of precision regulation, Governments can mitigate the potential risks of AI without stifling its use in a way that dampens innovation or risks cutting off from the trillions of dollars of economic growth that AI is predicted to unlock.

Question 16: Is a risk -based approach better suited to some sectors, AI application s or organisations than others, based on organisation size, AI maturity and resources?

IBM believes that the risk-based approach, utilising the principles outlined under in other parts of this submission, is the best way to ensure consistency across the Australian economy. The key issue will be to ensure that organisations understand their obligations, regardless of their size, AI maturity, and resources.

In IBM's view there are many potential industry associations and community partners who will be well positioned to assist the Australian Government in the practical implementation of a risk-based approach to managing AI. Indeed, informing and educating organisations about their responsibilities in developing or deploying AI is somewhat aligned to the journey that Australian governments and organisations are currently undertaking in relation to informing and educating Australian business and the public on Cyber Security. In other words, implementation will require an ongoing partnership between the Australian Government, industry and the community.

Question 17: What elements should be in a risk-based approach and do you support the elements in Attachment C?

As discussed elsewhere in this submission, IBM believes that a risk-based approach should include the elements that can deal appropriately with the issues of

- Transparency
- Bias and Fairness
- Explainability
- Robustness
- Privacy
- Human Oversight

These are broadly aligned with the elements that are outlined in attachment C.

Question 18: How can an AI risk-based approach be incorporated into existing assessment framework or risk management processes?

IBM has proposed that both public and private entities using AI should be required to have strong internal governance processes. These include, among other things:

- Designating a lead AI ethics official responsible for developing an organization's trustworthy AI strategy;

- Establishing an AI Ethics Board (or similar) to serve as a centralized clearing house to help guide implementation and oversight of that strategy. Not every AI system presents the same risk, which requires any AIA to take account of the different intended context, reliance of the end-user(s) on the output, and degree of automation employed.; and

- The use of AI Impact Assessments.

By way of practical example, the IBM AI Ethics Leader and the AI Ethics Board play a critical role in overseeing our internal AI governance process, creating reasonable internal guardrails to ensure we introduce technology into the world in a responsible and safe manner. The IBM AI Ethics Board in particular was central in IBM's decision to sunset our general purpose facial recognition and analysis products, considering the risk posed by the technology and the societal debate around its use.

The IBM AI Ethics Board, along with a global community of AI Ethics focal points and advocates, reviews technology use-cases, promotes best practices, conducts internal education, and leads our participation with stakeholder groups worldwide. In short, it is a mechanism by which IBM holds itself, and its staff accountable to its values, and our commitments to the ethical development and deployment of technology.

AI Impact Assessments

AI Impact Assessments can be a very useful and practical tool. In our view, it is reasonable to require deployers of AI systems to make an initial assessment and disclosure about whether an AI system they propose to implement is high-risk or not, focusing on the impact of the decisions being supported by the AI and the degree to which informed human oversight is being provided. The IBM Fact Sheets contain much of this type of information specific to IBM applicaitons.

The concept of AI Impact Assessment is a process that describe the potential impact of an AI system, as well as document related information relevant, such as the intended purpose, data set provenance, and performance data. AI Impact Assessments are designed to aid internal actors seeking to make decisions within the broader governance established by risk management frameworks. In other words, risk management frameworks define the boundaries of what is permitted in an AI system's lifecycle, while impact assessments provide information individuals can use to assist decision-making within those broader constraints.

IBM believes that all entities providing or owning an AI system should conduct an initial high-level assessment of the technology's potential for harm. Such assessments should be based on the intended use-case application(s), end-user(s), how reliant the end-user would be on the technology, and the level of automation. Once initial risk is determined, a more in-depth and detailed assessment should be undertaken for higher-risk applications. In certain low-risk situations, a more cursory appraisal would usually suffice. For those high-risk use-cases, the assessment processes should be documented in detail, be auditable, and retained for a minimum period of time.

Of course, it is important to recognize that while impact assessments can potentially help in identifying concerns posed by high-risk AI applications, they are not necessarily effective for managing all risks presented by AI. Rather, impact assessments should be viewed as one tool among many for managing those risks. Additionally, the use of impact assessments in risk-management should not automatically imply the particular AI use case in question is high risk.

We would also clarify that, to avoid over-regulation, the definition of high-risk and any regulatory requirements flowing from that should not apply to AI systems during their research and development, but only to those deployed or placed on the market, given that any risks will only occur during operational use and not in the research and development phase.

What do AI Impact Assessments (AIAs) assess?

In IBM's view AIA's should assess the following;

- **Transparency:** AIAs should clearly include descriptions of the intended purpose of a given AI system, as well as detailing – to the extent practicable – the information necessary for users to make reasonably-informed decisions on the use of the system. The appropriate degree of transparency here should be linked to the level of risk presented by the use of the AI system. They should not require companies to disclose source code or other trade secrets or IP. Additionally, developers should be maintaining records and documentation related to training data and other inputs, confidence measures, error analysis, and other relevant details.
- **Explainability:** In certain contexts, AIAs may need to provide certain details on how a system arrives at its predictions or influences decision-making processes to users. However, it is important to note that explanations will vary by audience and use, and therefore not all AI systems should require the same type or level of explanations.
- **Robustness.** AI systems should be capable of handling exceptional conditions if and as they arise, such as input abnormalities and adversarial attacks. AIAs should have supporting guidance on how developers can improve robustness to detect and mitigate cybersecurity risks and promote users' confidence in an AI system's model and outputs.
- **Privacy.** To ensure users' data is properly protected, an AIA needs to assess the provenance, security, and other privacy preserving aspects of data sets used by AI models. Separate data privacy impact assessments may also help contribute to a more holistic AIA process by helping to describe and reduce the risks associated with selection and use of data sets to reduce risks of privacy breaches, including but not limited to anonymizing all personal data proposed to be used in a data set.
- **Bias Mitigation:** Both providers and owners of AI systems have a responsibility to ensure their technologies are fair. When considering best practices for overseeing testing and mitigation of bias in an AI system, an AIA should account for the different roles of various actors within the larger AI developmental lifecycle to effectively allocate each responsibility to the party positioned to address it. For example, owners may need to ensure their AI systems align with anti-discrimination laws and implement processes for ongoing

monitoring and maintenance, while providers may need to assess more fundamental features of the training data sets used to train early-stage AI models.

In regard to the role of Government and AI Impact Assessments (AIAs), IBM recommends that policymakers should focus on prioritizing rules and regulations that establish baseline expectations of transparency in the *process* of developing and conducting AIAs, rather than prescriptive disclosure requirements, which could present a host of unintended consequences, such as revealing competitive secrets. In other words, policymakers should prioritize more general requirements for the use of testing and AIAs for high-risk AI systems and avoid prescriptive approaches on *how* to test.

Policymakers should also note the the value of organizations using existing accountability mechanisms to promote the use of AIAs and constrain how their AI systems are used by third parties. For example, reputational incentives have a strong influence on how organizations develop and deploy these technologies and when coupled with the use of contractual clauses and soft law arrangements, can help promote responsible development and deployment.

Question 19: How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or mulitmodel foundation models (MFMs)?

There is no reason why a risk-based approach to AI proposed in other parts of this submission should not also apply to large language models (LLMs) and multimodal foundation models (MFMs).

The explosion of generative AI systems in recent months has also caused some to call for a deviation from a risk-based approach and instead focus on regulating the technology itself, rather than its application. We believe that this would be a serious error, unnecessarily hindering innovation and limiting the benefits the technology can provide. A risk-based approach ensures that guardrails for AI apply to any application, even as new, potentially unforeseeable developments in the technology occur, and that those responsible for causing harm are held to account.

Question 20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool, or be mandated through regulation?

A) A risk-based approach can be voluntary, self-regulatory, or mandated. In IBM's view, a regulatory approach can be formalised to promote compliance provided that it reflects the principles outlined earlier in this submission to mitigate the potential risks without stifling innovation and dampening the opportunities for economic growth that AI offers the economy. This can apply to both public and private organisations.

B) It is critical that any AI regulatory regime make a clear distinction between developers and deployers.

Developers design, code, or produce AI models or systems, usually for downstream use by other deployers. Deployers use AI systems to, for example, aid in decision-making directly tied to end-users (i.e., consumers),. These differences are critical in terms of developing appropriate regulatory obligations. An extract is outlined below. In regard to AIA's, IBM recommends as follows;

Developers conducting design evaluations should document, as appropriate:

- The potential for consequential decisions resulting from use of the AI model or system to disproportionately impact people on the basis of protected characteristics, to the extent feasible, and the steps taken to mitigate the risk of such harm occurring;

- Known limitations of the AI model or system;
- An overview of the data used to train the AI model; and
- Metrics used to evaluate performance and bias.

Deployers conducting impact assessments should document, as appropriate:

- Purpose and intended use cases specific to the sector of operation of the deployer;
- The potential for consequential decisions resulting from use of the AI system to disproportionately impact people on the basis of protected characteristics, to the extent feasible, and the steps taken to mitigate the risk of such harm occurring;
- Information about data used to customize the system after purchase, if applicable, and real-world data inputs and outputs;
- Metrics for evaluating the AI system's performance, if applicable;
- Transparency measures; and
- Post-deployment monitoring and user safeguards, if applicable.

Attached at Annexure A is a general framework for dealing with the different obligations that should fall on developers and deployers that IBM developed with one of its industry associations, the Business Software Alliance.

In conclusion, IBM would like to thank you for the opportunity to contribute to developing safe and responsible AI in Australia. If you would like to discuss further or have any questions in relation to this submission, please contact me via email at kaaren@au1.ibm.com.

Yours sincerely

Kaaren Koomen AM
Director, Government & Regulatory Affairs
IBM Australia and New Zealand

Annexure A: Model AI Impact Assessment Legislation Framework

Section 1: Key Terms and Concepts

Consequential Decision:
A Consequential Decision is a determination made by a Deployer that has a legal or similarly significant effect on an individual.

Legal or Similarly Significant Effect:
A legal or similarly significant effect is a decision that determines an individual's eligibility for and results in the provision or denial of housing, employment, credit, education, access to places of public accommodation, healthcare, or insurance.

Consequential Artificial Intelligence Decision System (CAIDS):
Consequential Artificial Intelligence Decision Systems are machine-based systems or services that utilize machine learning, artificial intelligence, or similar techniques to the extent they provide an output that is not predetermined and have been specifically developed, or an AI system specifically modified, with the intended purpose of making or materially supporting Consequential Decisions.

Developers:
Any entity that designs, codes, or produces a CAIDS, or modifies an AI system with the intended purpose of making a Consequential Decision, whether for internal use or for use by third parties

Deployers:
Any entity that uses a CAIDS to make Consequential Decisions
Determining whether a covered entity is acting as a Developer or Deployer is a fact-based determination that depends upon the context in which the CAIDS operates.

Section 2: Prohibited Practices

A Deployer shall not use a CAIDS in a manner that violates state or federal laws prohibiting unlawful discrimination against individuals.
Nothing in this section shall limit a Developer or Deployer from processing data for legitimate internal testing for the purpose of preventing unlawful discrimination, rebalancing data sets to address disparities that are identified, or otherwise evaluating the effectiveness of a CAIDS.

Section 3: Deployer Impact Assessment Requirement

Deployers shall implement and maintain a risk management program that establishes the policies, processes, and personnel that will be used to identify, mitigate, and document risks arising from the deployment of a CAIDS. In implementing the risk management program, a Deployer shall take into consideration (a) its size and complexity; (b) the nature and scope of the CAIDS, including its intended use; (c) the sensitivity and volume of data processed in connection with the CAIDS; and (d) cost of available tools.
Deployers shall be required to perform an impact assessment prior to deploying a CAIDS and annually thereafter. If there are material changes made to the purpose for which a CAIDS is used or the type of data it receives, or if a demonstrable and material difference in how the CAIDS performs is detected, a new impact assessment shall be conducted.
In performing an Impact Assessment, Deployers shall maintain documentation for a reasonable time period in light of the intended use regarding:

1. *The purpose of the CAIDS and its intended use cases, deployment context, and benefits;*

2. *The extent to which the use of the CAIDS once it is deployed is consistent with or varies from the Developer's description of intended uses;*
3. *The potential for denials of housing, employment, credit, education, access to places of accommodation, healthcare, or insurance that could result from use of the CAIDS, whether those harms may disproportionately impact people on the basis of protected characteristics, and the steps taken to mitigate the risk of such harm occurring;*
   4. *A description of data that will be processed as inputs by the CAIDS once deployed and a description of the outputs produced by the CAIDS;*
   5. *If applicable, an overview of the type of data used to customize the CAIDS;*
   6. *Metrics for evaluating the CAIDS's performance and known limitations;*
   7. *Transparency measures, including information identifying to individuals when a CAIDS is in use; and*
   8. *Post-deployment monitoring and user safeguards, including a description of the oversight process in place to address issues as they arise.*

Section 4: Developer Obligations

(1) *Developers shall make available to a Deployer the information reasonably necessary for the Deployer to comply with its requirement to perform an impact assessment, including documentation regarding a CAIDS's capabilities, known limitations, and guidelines for intended use. Nothing in this Act shall require the disclosure of trade secrets or other confidential information.*

(2) *Developers shall implement and maintain a risk management program that establishes the policies, processes, training procedures, and personnel that will be used to identify, mitigate, and document risks arising from the development of a CAIDS, including conducting a design evaluation pursuant to paragraph (3). In implementing the risk management program, a Developer shall take into consideration (a) its size and complexity; (b) the nature and scope of the CAIDS, including its intended end use; (c) the sensitivity and volume of data used to train the CAIDS; and (d) cost of available tools.*

(3) *Developers shall conduct a design evaluation for a CAIDS. The design evaluation shall consider information relevant to the potential for unlawful bias in connection with the intended end use of the CAIDS. Developers shall, as appropriate given their role in the development of the CAIDS, maintain documentation for a reasonable time period in light of the intended use regarding:*

- **a.** *The purpose of the CAIDS and its intended end use cases, features, and benefits;*
- **b.** *The potential for denials of housing, employment, credit, education, access to places of accommodation, healthcare, or insurance that could result from use of the CAIDS, whether those harms may disproportionately impact people on the basis of protected characteristics, and the steps taken to mitigate the risk of such harm occurring;*
- **c.** *Known limitations of the CAIDS, including factors affecting performance;*
- **d.** *An overview of the type of data used to train the CAIDS and how the data was collected and processed; and*
- **e.** *Metrics for how the CAIDS's performance was evaluated prior to sale.*

Section 5: Enforcement and Oversight Mechanisms

Developers and Deployers shall publicly certify that they are in compliance with their obligations under this Act. Developers and Deployers may use an impact assessment or design evaluation conducted in accordance with other laws or regulations if such assessment or evaluation is reasonably similar in scope.

If the Federal Trade Commission obtains credible information or evidence that a Deployer or Developer has violated the Act, the Commission may issue a Civil Investigative Demand that requires the Deployer or Developer to disclose to the Commission the contents of a relevent impact assessment or design evaluation.

Impact assessments and design evaluations shall be confidential and exempt from public inspection and copying under the Freedom of Information Act, 5 U.S.C. § 552. The disclosure of an impact assessment or design evaluation pursuant to a request from the Commission shall not constitute a waiver of attorney-client privilege or work product protection with respect to the assessment or evaluation and any information contained therein.

Section 6: Pre-emption

The provisions of this Act shall preempt any law, rule, regulation, or other requirement of any State or political subdivision of a State to the extent the purpose, application or operation of the law, rule, regulation or other requirement relates to what constitutes an unlawful use of a CAIDS or the performance of an impact assessment or design evaluation, or the equivalent thereof, of a CAIDS.