Capgemini invent

AHRC RESPONSE PAPER

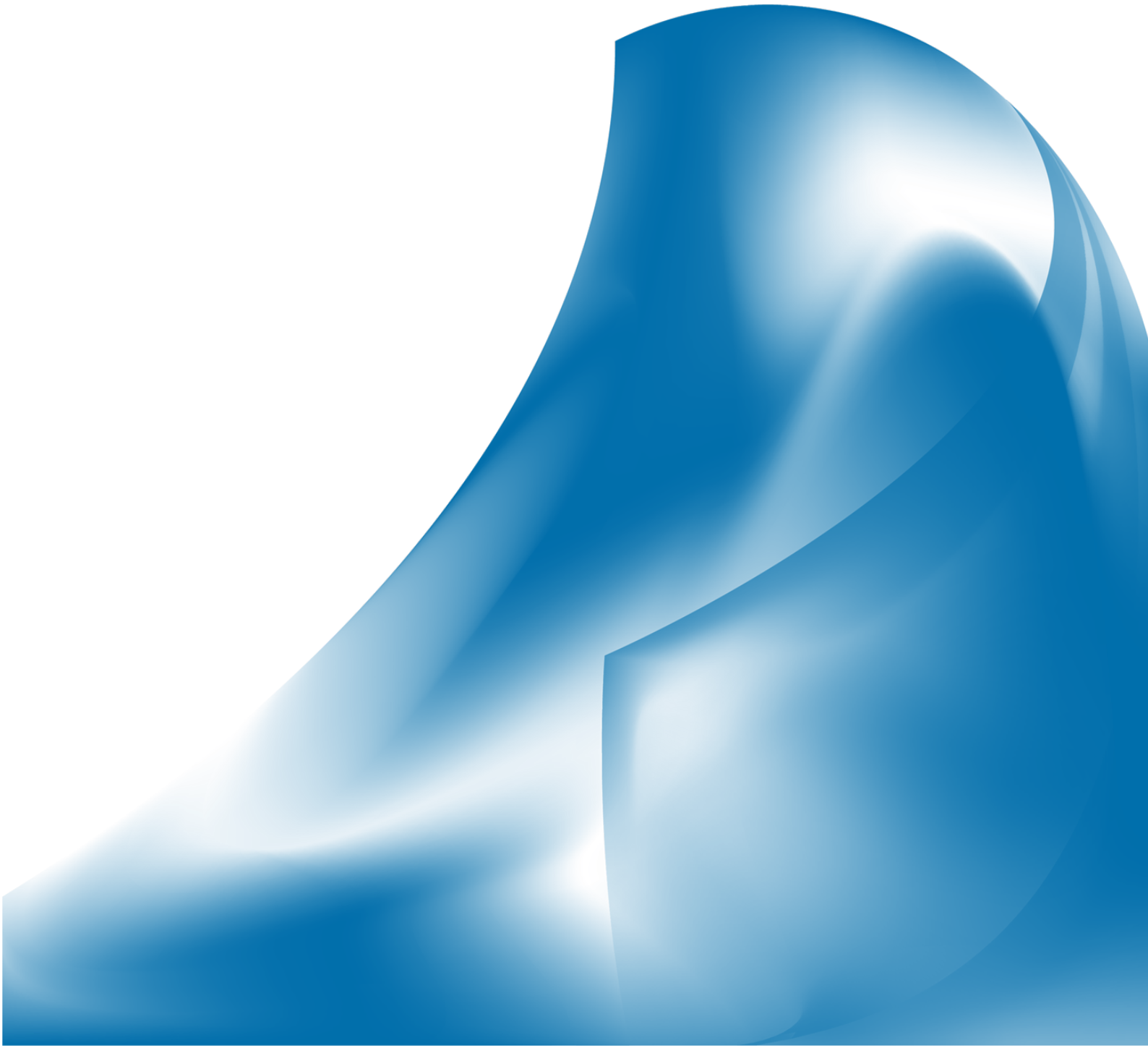

Table of Contents

EXECUTIVE SUMMARY

Capgemini is a global leader in consulting, technology services and digital transformations. Our teams innovate to address a breadth of client opportunities in the world of cloud and digital. Capgemini Invent combines strategy, technology, data science and creative design to solve complex business and technology challenges. We are responding to the submission paper by the AHRC with deep rooted experience in the area of Artificial Intelligence from a technical and functional viewpoint, however, the contributors to this response also have prior expertise in areas of law and policy reform.

Capgemini has been recognised consecutively for the last 8 years by the Ethisphere Institute as one of the World's Most Ethical Companies. The Ethisphere Institute is the global leader in defining and advancing the standards of ethical business practices that fuel corporate character, marketplace trust, and business success.

The Capgemini Research Institute recently published a research paper on Ethics in AI to which it found that "51% of executives consider that it is important to ensure that AI systems are ethical and transparent. Organizations are also taking concrete actions when ethical issues are raised. As the figure below shows, more than two in five executives report to have abandoned an AI system altogether when an ethical issue had been raised." (https://www.capgemini.com/au-en/wp-content/uploads/sites/9/2019/07/CRI-AI-in-Ethics_web-1.pdf)

We believe the future leaders combine the mind, the heart and the gut to become game changers: embedding a profit for purpose approach with a flair for disruption. As a globally renowned technology company, we have the ambition and means to contribute to helping solve major societal questions through our practice of Invent for Society. Invent for Society aims to tackle social impact with our clients in areas such as digital inclusion, poverty prevention and waste reduction. One of the main areas of focus for Invent for Society is that of Trust in an Intelligent World which aims at making the most of data and Artificial Intelligence, whilst also reinforcing digital human rights and trust. You can find more on Capgemini's Invent for Society here https://www.capgemini.com/au-en/service/invent/invent-for-society/.

Our response below focusses on areas in which our contributors are experts. Our response includes answers to questions A, E, F and our educated thoughts on proposals 5, 7 and 8.

1. RESPONSES

1.1. PROPOSAL 5 AND QUESTION A

Proposal 5: *The Australian Government should introduce legislation to require that an individual is informed where AI is materially used in a decision that has a legal, or similarly significant, effect on the individual's rights.*

Question A: *The Commission's proposed definition of 'AI-informed decision making' has the following two elements: there must be a decision that has a legal, or similarly significant, effect for an individual; and AI must have materially assisted in the process of making the decision.*

Is the Commission's definition of 'AI-informed decision making' appropriate for the purposes of regulation to protect human rights and other key goals?

In response to the proposal and related question on "AI-informed decision making" Capgemini would like to draw attention to the current status of AI usage, and look forward to the applicability of the proposal within the future usage environment.

Good decision-making utilizes diverse sources of information. With the increase in IT, networked infrastructure and automation, individuals have seen an increase both in the amount of information available for the decision-making process, as well as their ability to process this information. The increase in prevalence of this technology has seen a change from discrete systems supporting decisions to integrated systems of systems which together form a global resource in which the individual functions.

The framing of the question and proposal for AI "materially assisting" in the decision-making process reflects the current density of implementation where AI systems are isolated specialists, operating as islands within a broader set of deterministic systems. Capgemini view this as a transitory phase, with an increase in the number of AI-type technologies involved in the processing that eventually leads to an end-customer experience. Examples may include the use of AI-technology in data gathering, including where data is gathered by non-AI systems in an environment where the user is nudged or affected by AI or chains of AI systems (e.g. video extraction, voice extraction, NLP, profiling, next-best-decision). This scenario where AI takes on a multitude of roles, some trivial in execution but with the capability to materially affect outcome, points to a future where the question of "if" is no longer truly relevant.

Secondly, a challenge arises with the use of the term "AI" here. The intent of this point is not to discuss what qualifies as AI technology, but to recognize that the assumed benefit of the proposal does not restrict itself to only AI technology, nor does the harm it presumably intends to address. Automated decisions based on human-optimised linear regression models present the same perils for the impacted individual, with many of the same risks of misuse, lack of transparency, etc.

Finally, and following the previous point, it is unclear why a distinction would be made for AI technology. We recommend instead that language be considered that reinforces the following principles:

- That where a decision is made, the decision-making organization should be able to explain how that decision was arrived at, regardless of whether the information processing was performed by a human, a deterministic system or a non-deterministic system.
- That this requirement further requires the organisations making decisions to have accountable parties for each step of that decision-making process or has contractual agreements in place to provide the same.

Capgemini does note that disclosure is likely desirable in situations where the nature of the system or individual with whom a person is interacting could not reasonable be inferred. This would apply for AI systems masquerading as human (chat-bots), AI systems or backends which would reasonably be assumed as simple deterministic systems (recommendation systems), as well humans involved in interactions which would reasonably be assumed as only with a machine (e.g. personal assistants).

1.2. PROPOSAL 7 AND PROPOSAL 8

Proposal 7: *The Australian Government should introduce legislation regarding the explainability of AI-informed decision making. This legislation should make clear that, if an individual would have been entitled to an explanation of the decision were it not made using AI, the individual should be able to demand:*

a) a non-technical explanation of the AI-informed decision, which would be comprehensible by a lay person, and

b) a technical explanation of the AI-informed decision that can be assessed and validated by a person with relevant technical expertise.

In each case, the explanation should contain the reasons for the decision, such that it would enable an individual, or a person with relevant technical expertise, to understand the basis of the decision and any grounds on which it should be challenged.

Proposal 8: *Where an AI-informed decision-making system does not produce reasonable explanations for its decisions, that system should not be deployed in any context where decisions could infringe the human rights of individuals.*

Capgemini recognises the overarching obligation and need to provide explanations and reasons for AI-informed decisions, particularly where those decisions may adversely affect individuals.

It is not clear on the level of granularity or detail required to be enough for a non-technical vs technical explanation which may create several risks:

1. The non-technical explanation becomes a 'summary' of business logic that can be repeated irrespective of the decision being made. The risk is that the 'non-technical explanation' becomes a repeatable business logic that would be provided to all applicants who receive a decision. As such, applicants would struggle to 'challenge' a decision using this explanation, as the applicant would no longer be challenging the decision, but challenging the logic or function used to arrive at that decision. Businesses & decision-makers may struggle to provide a non-technical explanation that pertains specifically to an individual.
2. Complex AI models involving deep learning and neural networks can be extremely difficult to provide explanations beyond a summary of business logic. The risk of non-compliance is high given the current state of AI products in the market and the

complex logic occurring. It may not always be possible to provide a valuable non-technical explanation in some cases and will always defer to technical explanations.

3. The explanation becomes 'too detailed' and the distinction between technical and non-technical becomes unclear.
4. Within data being assessed, certain features of business logic may be given preference above others as part of the application. For example, a loan application may assess history, school, location etc. If an applicant gets rejected and asks for reasons, the 'explanation' may provide the business logic and credit scoring against several criteria. If the applicant is rejected because they live in an area with a poor credit score, the applicant may struggle to successfully challenge the 'logic', particularly because the logic is not specific to their application but is a generic set of business rules.
5. We see two prominent risks if a non-technical explanation is the provision of business logic:
    a) People presented only with business logic would only be able to challenge that decision by challenging the business logic in its entirety. Many of these logics are based on industry standards or common practices.
    b) The lay person would not be able to verify the business logic has been successfully applied without a full technical explanation. In most scenarios, this would require the lay person to have access to a person with the appropriate technical skillsets to interpret the technical explanation. Most applicants won't have access to this knowledge.

"The better view, among experts involved in this area appears to be that it is almost always possible to design an AI-informed decision-making system so that it provides a reasonable (albeit not perfect) explanation of the basis for the decisions or recommendations it generates." (Human Rights and Technology Discussion Paper 2019 97)

Whilst acknowledging explanations can be given for any AI-informed decision, consideration should be given to the cost of providing explanations in complex, neural network systems. There are scenarios where a technical explanation may involve significant efforts, from multiple experts to assess, identify and sort through data sets to provide a sufficient explanation for an individual decision being made. In these cases, the 'reasonableness' of the explanation provided will vary greatly and risks devolving into unsatisfactory explanations due to the complexity of the system.

Considerations:

1. Establishing an independent watchdog or additional standards and guidance to assist enterprise to be compliant with this requirement.
2. Templates or sample explanations may also assist industry in setting up correct processes to provide explanations for AI-informed decisions.

'...Australian courts already have powers to hear matters involving commercially sensitive evidence. Therefore, there is nothing that should preclude a court from receiving and assessing this sensitive evidence (such as an algorithm), with safeguards that prevent its broad publication.' (Human Rights and Technology Discussion Paper 2019 100)

The suggestion that Australian Courts have powers to hear commercially sensitive evidence may not be a feasible approach, particularly given the multi-nationals involved in developing and implementing AI-based products. Many of the most advanced, complex AI tools are produced by large multinational software companies. There are also complex service agreements and intellectual property rights to consider between service providers & consultancies to implement and use these products. Further guidance is needed as to how courts will enforce compliance and navigate these agreements to align with the rule of law.

1. The technical paper notes that Australian Courts have powers to hear matters involving commercially sensitive evidence. Most applicants don't have means to apply in Court for provision or Orders to receive sensitive evidence (i.e. Welfare applicants).
2. Low-cost bodies or watchdogs may need to be setup to enable low-cost means for applicants to view sensitive evidence. These bodies will need strict controls in place to ensure compliance and leakage of sensitive information. This requirement will be difficult to manage because usually stricter controls will increase cost and may force applicants to lodge proceedings in Courts. This will inadvertently create economic imbalance favouring enterprise over individual rights.
3. Applicants may need low-cost access to technical expertise to assess and validate technical explanations. This cost is unlikely going to be feasible for most applicants (particularly given the shortage of skillset available in the Australian market).
4. Additionally, many AI products leverage third parties during implementation and ongoing services. Consideration should be given to scenarios where the decision-maker relies on third party software or products. These scenarios provide several risks:
    a. Risk of contractual breach (due to IP or non-disclosure agreements)
    b. Unclear who is accountable for compliance
    c. Difficult to enforce compliance (i.e. multinational service providers – noting that some multinational agreements may not be subject to Australian law despite the engagement of Australian entities).

New legislation should also be clear on the distinction between private or public bodies.

Public bodies (particularly regulatory bodies) usually have higher obligations under administrative law principles to provide reasons for a decision (i.e. when making determinations pursuant to legislation). For example, a determination that an individual is not entitled to Centrelink benefits. Private bodies usually do not have the same rigorous obligations to provide reasons, particularly given the nature of the decisions being made. For example, a decision to reject an application for a mortgage through a private bank.

Ordinarily, decisions made by private bodies are less likely to have an impact on the human rights of individuals. As such, the obligations imposed, particularly the requirement to provide detailed explanations should be distinguished between private and public entities. The impact of not-distinguishing between these bodies is increased compliance costs, stifling of innovation and investment and is likely to deter foreign activity in the Australian economy.

1.3. QUESTION E

Question E: *In relation to the proposed human rights impact assessment tool in Proposal 14:*

a) When and how should it be deployed?

b) Should completion of a human rights impact assessment be mandatory, or incentivised in other ways?

c) What should the consequences be if the assessment indicates a high risk of human rights impact?

d) How should a human rights impact assessment be applied to AI-informed decision-making systems developed overseas?

Capgemini recognises that risk management is a useful tool in avoiding unintended outcomes and harm to stakeholders from the use of Artificial Intelligence technologies. Of importance in considering the use of risk management is the tendency for Artificial Intelligence technology to better perform for groups for whom large amounts of data is available, and so is naturally challenged to provide equitable treatment to marginalised members of society, minority groups and those for whom there will be less data (those already disadvantaged by limited technology adoption). It is therefore of vital importance that these voices are well considered in any risk exercise. Capgemini therefore strongly discourages any risk assessment exercise that provides viewpoints of risk only from majority groups, internal groups or a limited set of stakeholders, thereby compounding risk of disadvantage with underrepresentation in understanding of unintended outcomes.

With respect to timing of the risk assessment, it is first important to consider that decisions which affect the user's, or other stakeholder's, experience can emerge at several junctures within the lifecycle of strategy, product or capability development. Additionally, it should be recognised that from the perspective of the company, not all damage from realised risk will be from litigation or non-compliance. Organisations will be aware that unintended outcomes can lead to brand damage, loss of customer trust & loyalty, damage to positive culture of workforce, etc. By addressing risks to these aspects of a company's health, a risk management approach can be utilised to not only demonstrate compliance, but to manage the other inevitable risks that accompany a bold strategy.

It is for this reason that Capgemini suggests a framework, for example that specified as part of our Ethical AI offering, that addresses risk from strategy through to the edge of the organisation and to in-life use by customers. By recognising that the nature of the technology, and the way it is utilised, inevitably diversifies the location of risk emergence the company can ensure the correct accountabilities exist in their organisation and that the workforce is empowered and supported.

We would recommend:

- An exercise to understand the inherent, macro-level risks that accompany a company's AI strategy, including the broad trade-offs to the company's other strategic goals, the company's vision and purpose. This exercise should recognise risks to regulatory obligations, as well as to other key concerns such as brand, customer trust, workforce motivation, etc.
- Regular, scheduled, programme-level assessments and reviews across the enterprise architecture, including product development, supplier management, process management, project/programme management, development, data management, human resources, customer engagement and product/service support.
- Key stage gates in projects involving the development of Artificial Intelligence technologies.
- Model Risk Management for in-life models, alongside regular product and service performance reviews.

By approaching the risk management exercise as an enabler for producing products and services that enthuse customers, rather than a compliance hurdle, we anticipate more positivity in uptake. Any regulation should therefore be accompanied by assistance to leverage a risk management framework for positive outcomes.

1.4. QUESTION F

Question F: *What should be the key features of a regulatory sandbox to test AI-informed decision-making systems for compliance with human rights? In particular:*

a) what should be the scope of operation of the regulatory sandbox, including criteria for eligibility to participate and the types of system that would be covered?

b) what areas of regulation should it cover e.g., human rights or other areas as well?

c) what controls or criteria should be in place prior to a product being admitted to the regulatory sandbox?

d) what protections or incentives should support participation?

e) what body or bodies should run the regulatory sandbox?

f) how could the regulatory sandbox draw on the expertise of relevant regulatory and oversight bodies, civil society and industry?

g) how should it balance competing imperatives e.g., transparency and protection of trade secrets?

h) how should the regulatory sandbox be evaluated?

Capgemini has extensive experience in this area through our work assisting in financial services environments. The design and management of a sandbox should consider the following.

Functioning

With regards to the process of utilizing the regulatory sandbox, Capgemini considers that it is important to first identify the benefits for the parties involved.

In forming our recommendations, we consider the following benefits as targeted outcomes of the sandbox. For the regulator a sandbox allows:

- Planning and notice of activities to certify a product or line of products as compliant allowing smoothing of demand on resources
- Increased insight and familiarity with various applications and use-cases, allowing efficiencies in focusing effort
- Identifying compliance issues and concerns early in the development process, avoiding additional effort in technical examination
- Transparently testing, refining and sharing best practices for ethics in AI and technology

For the organization utilizing the sandbox:

- Early notification of issues or regulatory concerns, preventing investment in developing ultimately unsuccessful products
- Guidance from subject matter experts in compliance to co-create products that meet business need and remain compliant
- Avoidance of costly product withdraw, brand-damage, litigation and other impacts of releasing products ultimately deemed as non-compliant

For users, customers and other societal stakeholders

- Avoidance, as much is possible, of detrimental outcomes from the use of non-compliant products
- Increased trust in products that have been sandboxed in cooperation with regulatory SMEs.
- Exposure of new emerging innovation, small businesses and other parts of society that would have a critical role to play in the future of Australia but have less power than more established players in society.

With these goals in mind Capgemini recommend a co-design approach during sandbox activities between regulatory authorities and utilizing companies wherever possible to jointly uncover concerns and issues as early in the development process as possible. It is our experience that early identification of regulatory concerns during the software lifecycle is likely to be an easier exercise, preserving resources in both parties, as well as steering development by the utilising company before investment in non-compliant product.

We recommend that prior to any product utilising a regulatory sandbox, that a formal exercise is undertaken to identify:

- the relevant risks inherent in the aims of the product. These risks should cover at a minimum human rights and other related obligations to the individual.
    - Capgemini notes that not all risk arising is strictly litigable by regulation. This would be an opportunity to address other ethical concerns that may lead to brand damage, damage to strategic intent, staff retention, recruiting, etc.
- to what level the company believes they have an obligation or duty to constrain these risks (including measurable boundaries and standards which should be met or superseded to meet this)
- the proposed mitigations where the risks cannot be addressed fully to meet a proposed threshold.

We recommend that this assessment and response be the entry point for discussions on use of the sandbox environment, setting out both the expectations of the product, but also allowing alignment on whether these expectations provide adherence. These thresholds, metrics and the required mitigating functionality would form part of the requirements for the technical product. Any mitigations not directly implemented as part of the technical product would form part of the general product (e.g. channels for query, redress, etc.)

Capgemini also recommend that as part of this exercise a suitable stakeholder identification and engagement exercise is performed to both inform the understanding into the risk discussions, but also to aid in the creation of realistic, representative and inclusive test cases for the sandbox environment. The stakeholder exercise should include both intended users, as well as representation from other impacted and affected groups.

Based on the outcomes of the testing within the sandbox and the type of application (e.g. newer applications could require more monitoring), a mechanism for on-going supervision and monitoring, perhaps via the appropriate industry regulator, is recommended. This would allow for the right Just-In-Time support for high risk, or unproven use cases that may have ethical concerns.

Technical specification of the sandbox environment

Capgemini recommends the environment be designed for portability, leverage open architecture approaches and cater to multiple key groups (developers, project managers, innovators / inventors and the general public). The sand box must adhere to baseline Ethical architecture principles like data integrity, access control, auditing, code Genuity, trust, confidentiality, privacy, authenticity, attack prevention, federated trust etc. The sandbox environment must be architected to support several features:

1. Hosting - centrally host new AI and technology innovations
2. Security – Allow zero trust platform where environment access, integration & data is fully secured.
3. Local testing – allowing a copy of a lightweight sandbox to be portable and useable for testing by developers within their own environments and organisations
4. Data as a service – refer and introduce sources of data to ensuring completeness of testing
5. Inventory of use cases – list of scenarios that have successfully graduated from sandbox testing
6. Sandbox clinics – regular collaboration and support mechanisms for users of the sandbox
7. Connectivity – provide for multiple ways to connect and instantiate the sandbox, ensuring isolation of testing / projects from each other.
8. Auditing – Captures all the access & software related activity on the environment.
9. Ease of Deployment – provides single click product deployment on sandpit environment.

Ongoing monitoring and dynamic systems

Many of the systems classified as "Artificial Intelligence" are non-deterministic in that they are not rules-based, and perform inferences based on previously encountered data. In addition, many systems dynamically learn using data encountered. Both these traits dictate that systems tested within the sandbox cannot be fully assured of behaving in a compliant manner outside of the confines of the environment. It is the nature of the technology that it is often impossible to exhaustively test. In addition, many of the technologies do not display linear responses to variance in input, making it difficult or impossible to set "edge-cases" to represent the limits of the responses likely from a system.

With this in mind Capgemini recommend any use of the sandbox environment to be accompanied by a robust framework to measure the ongoing compliance to risk-boundaries set at specification of the system, recognizing that non-compliance may also occur due to a changes within the environment of the system as well as changes to the system itself. The parameters of this framework should form part of the exit criteria for the sandbox environment.

Capgemini recommends consideration of a Model Risk Management approach, with reference to similar approaches suggested by Financial Service Regulators for traditional financial models. For example, the Federal Reserve System's Supervisory Guidance on Model Risk Management (SR 11-7).

On funding, adoption and usage

With respect to funding and engagement, there are a few models to explore:

1. Create a set of flagship open-source projects and run them with the support of "founding sponsors"
2. Establish public-private partnerships or consortiums to further define the appropriate model – e.g. sandbox for critical infrastructure would have different needs compared to customer experience related scenarios.
3. A de-centralised approach to sandbox testing involving the industry regulators driving the funding, approach and usage (e.g. APRA for the banking and insurance sectors).

Capgemini also view the sandbox as a sell-point for investment into the Australian economy. The provision of a trusted, mature sandbox environment provides a cost-minimised and lower-risk onramp for overseas technology providers to enter the Australian market. By expediting the availability of technology within the Australian market, and lowering barriers to entry, Australia would be able to both uphold their values and simultaneously enable frictionless cooperation with technology leaders.

2. CLOSING STATEMENT

In closing, whilst the AHRC has provided significant thought and strong guidance on the topic of ethical considerations with technology, Capgemini has written this response to support the clear implementation of policy in this space.

Capgemini continues to innovate and lead the conversation in this area. We also offer our clients a robust Ethical AI framework, accompanied by implementation of practical interventions, to assist companies embracing a Principles-Based approach to Artificial Intelligence development and use. Historically, Capgemini has played a significant role in shaping government policy and private-public sector collaboration in Europe. One of our ambitions from 2020 is to play a similar role in Australia and we have several initiatives on the go to play a greater role in solving some of Australia's challenges – e.g. productivity, aging population, growth of urban communities, fostering innovation etc.

We are glad to have this opportunity to provide input to the AHRC's submission and look forward to tackling ethical issues in AI in cooperation.

If you have any questions, please reach out to Matthew Newman (matthew.newman@capgemini.com) or Amanda Hajj (amanda.hajj@capgemini.com)

3. APPENDIX

Matthew Newman – Capgemini Invent, Innovation and Strategy, Associate Director

Matthew is a world-leading expert on the business application of AI Ethics. Backed by 20 years' experience in many of the world's most prestigious organisations, he offers complete experience in the adoption of a principles-based approach to AI strategy.

Duncan Cameron - Capgemini Invent, Future of Technology, Lead Consultant

Duncan is a qualified lawyer and consultant working across a variety of emerging technologies & innovation platforms. He was previously engaged as an advisor for several blockchain projects responsible for regulatory compliance, strategy for VC funding, operations and risk management. He now primarily works with large enterprise on innovation and commercial management.

Vivek Singh – Capgemini Invent, Future of Technology, Director

Vivek is currently working as Account Chief Architect in Financial Services. He has over 20 years of IT Experience including over 10 years in Senior Leadership roles in Strategic Product Planning, Architecture and Product Delivery.

Amanda Hajj – Capgemini Invent, People and Organisation, Consultant

Amanda has a passion for supporting organisations through large scale changes and transformations. Amanda often works in complex business environments with a focus on employee experience. Amanda has extensive experience in the realm of public policy and has worked closely with government officials to enact policy change.

Kirit Kundu – Capgemini Invent, Future of Technology, Senior Director

Kirit has spent the last 20 years solving complex business problems for clients, primarily within financial services. Over the last decade, Kirit has primarily undertaken advisory roles such as Product Lead, Product Owner or Chief Architect for large banking and insurance transformation programs, to mobilise business led change, deliver on buy side and sell side M&A deals and improve effectiveness of IT organisations.

Capgemini invent

ABOUT CAPGEMINI

A global leader in consulting, technology services and digital transformation, Capgemini is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms. Building on its strong 50-year heritage and deep industry-specific expertise, Capgemini enables organizations to realize their business ambitions through an array of services from strategy to operations. Capgemini is driven by the conviction that the business value of technology comes from and through people. It is a multicultural company of almost 220,000 team members in over 40 countries. The Group reported 2016 global revenues of EUR 14.1 billion.

Learn more about us at www.capgemini.com

People matter, results count.