Canberra, Australia
London, United Kingdom

+1300 996 905
info@castlepoint.systems
www.castlepoint.systems

Castlepoint.

Supporting Responsible AI Submission 20230804

CASTLEPOINT – SUPPORTING RESPONSIBLE AI

Contents

Who are we?

Castlepoint Systems™ is a regtech solution for information and records governance.

Our Data Castle™ model oversees and regulates all information in a network boundary.

Opportunities

AI can minimize harm

"Artificial Intelligence and automation is the only way to address the increasing risk of non-compliance that comes with a growing population, growing public service, and escalating regulatory and risk environment – but it must be explainable and able to be contested."

We can rethink what 'harm' is

"To support responsible AI practices across government agencies, those agencies need to know and recognise that the decisions they make with data can have real impacts on real people. And therefore, any use of AI for those decisions is inherently high risk. They must understand and internalise the first point before they will be able to put the second point into practice"

Challenges

We need objectivity

Risk should be mapped against human outcomes, before organisational ones. If an organisation could cause harm to an individual or social group… [the AI] must be explainable and contestable. It is the size and type of the harm that matters, not the size and type of the company"

We need to ensure risk alignment

"Risk-based approaches to any compliance can result in very different appetites across different agencies, and affect the overall governance posture in Departments. Having different risk cultures creates significant issues for governance alignment…. Risk-based approaches are likely to create perverse incentives to minimize risk levels, and subsequently dilute risk treatments"

Our recommendations

The NAA should take a formal position on the level of risk posed by ADM for records governance, and require agencies to apply commensurate AI oversight and control

Include definitions of 'black box' and 'white box' AI in any policy guidance or regulation, as the differences between the two are fundamental to the discussion of Ethical AI and explainability

Boards and individual Directors should be accountable for failures in Ethical AI compliance and any breaches or resulting harm

Standardise risk and impact models across government and industry to reduce subjectivity and perverse incentives

What do we do?

Castlepoint is a regtech product that reads, registers, and classifies every record in a network using Artificial Intelligence, and then determines and applies all relevant regulatory rules (retention, privacy, handling and security) to the data. It makes compliant information management invisible to users, and works across any system, without technical impacts.

We are an Australian company, based in Canberra, ACT. We launched our product in 2018. Our clients include Commonwealth and State government, industry, and universities.

Our vision is to change the way the world manages information, so that people, communities, and companies are safer and smarter.

Why do we do it?

Every organisation subject to regulation needs to know where its high value and high risk information is, and what rules apply to its management. They need to know who uses the information, what they do with it, and what events happen that trigger compliance actions. They need to be able to find relevant information across all systems, so that it can be used, reused, shared, and protected.

It's not only data that is growing. Regulatory obligations are increasing as well. Artificial Intelligence and automation is the only way to address the increasing risk of non-compliance that comes with a growing population, growing public service, and escalating regulatory and risk environment – but it must be explainable and able to be contested.

Significant harm has befallen individuals and communities whose information is in the custody of government or corporate stakeholders, and continues to occur regularly. It is no longer acceptable to allow sensitive or important data about citizens to be lost, spilled, or misused because it is too challenging to manage. AI has made it possible to be responsible with data for the first tie in the digital era. But AI is also introducing a new kind of risk, for a new kind of harm, which could move us out of the frying pan and into the fire.

What do we think?

Government needs to consider the depth and breadth of harm that can arise from information mismanagement, and recognize that information management represents a high risk use case for artificial intelligence.

We believe that Responsible or Ethical AI for information management must be mandated, and must apply to all organizations who hold data about citizens, and/or make decisions using that data which could detrimentally affect those citizens.

Responses

Castlepoint appreciates the opportunity to respond to the questions raised in the Supporting responsible AI: discussion paper as part of the consultation process.

Definitions

1.1 Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

We agree with the definitions but recommend also including definitions of 'black box' and 'white box' AI in any policy guidance or regulation, as the differences between the two are fundamental to the discussion of Ethical AI and explainability. Note that these may also be referred to as 'closed box' and 'open box' AI respectively, and this may be more appropriate and inclusive terminology to adopt.

Potential gaps in approaches

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

The National Archives of Australia is not represented in existing policy and guidance listed in the paper. The NAA does have some guidance on use of emerging technologies, and refers government records managers to the Industry AI Ethics Framework, but currently does not take a position on explainability for AI or other ADM with regards to information management.

AI for records management has been adopted across most federal government portfolios in Australia as of the date of this submission. Agencies are using both explainable and non-explainable AI in the governance of records, including deciding what records are preserved, what records are destroyed, and what records are protected. These types of decisions are irreversible once applied, and can cause unintended consequences and harm.

While the *Archives Act* is and should remain technology-neutral, the NAA should take a formal position on the level of risk posed by ADM for records governance, and require agencies to apply commensurate oversight and control of any AI or other ADM they adopt for these purposes.

Note that there are records retention rules in many Acts and Regulations that apply to non-government entities. These corporate organisations are also increasingly adopting AI for records sentencing and disposition, and they should also be required to ensure that harm from AI-driven decision making is mitigated. Over-retention, under-retention, and misclassification of information they hold can cause harm to their customers, often at significant scale. Any use of AI to make these decisions must be explainable and contestable by those affected persons and social groups.

The Protective Security Policy Framework (PSPF), administered by AGD, is currently silent on Artificial Intelligence and ethics. The use of AI for information classification should be addressed in the PSPF, in Policies 8, 9, 10, and/or 11. Cybersecurity is not only ancillary to use of AI, it is core. Artificial Intelligence is currently being used in Australian government and corporates to support decisions about what information is classified, sensitive, or otherwise high risk, and which information is not (and can therefore be shared or disseminated). If the decision is wrong in any of these cases, individuals and communities can be denied access to records they should be able to see, or conversely, information that should be protected can be spilled.

2.3 Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

Education and communication are an important part of supporting change from a bottom-up perspective. From the top-down, making Boards and individual Directors accountable for failures in compliance can demonstrably improve adoption of good governance.

2.4 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.

It is going to be important to define the risk criteria at a government level. Risk-based approaches to any compliance can result in very different appetites across different agencies, and affect the overall governance posture in Departments. Having different risk cultures creates significant issues for governance alignment in Machinery of Government changes.

Reliance on the ANAO or NAA to audit compliance, or a formal Inquiry, is usually a retrospective activity which follows harm that has already been realized. At the point that systems are processes are already embedded, it becomes hard for agencies to respond to the recommendations. For example, after the Comrie and Palmer reports into the unlawful treatment of Vivian Solon and Cornelia Rau, the Immigration Department embarked on a more than three quarter of a billion-dollar business transformation program over five years called Systems for People. But a capability review in 2012 found that the Systems for People program was not successful. The report pointed out that Immigration was still at high risk of another unlawful deportation or detention because of ineffective information governance, despite two formal Inquiries and significant funding to address the issue. It is very hard to 'cure' improper governance and non-compliance; it must be prevented.

Only a holistic coordination of compliance regulations can help prevent compliance breaches.

Responses suitable for Australia

3.1 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?

The global environment is evolving, but all advanced economies are reflecting the same essential principles and expectations for ethical AI. The Australian government should continue to monitor the

application of these principles, particularly how they are communicated and enforced. Lessons from other economies can be applied to Australia, through a lens of our corporate culture and customs.

Target areas

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

No – the intention of ethical AI is to protect the rights, entitlements, and safety of citizens. Both governments and corporates hold information about citizens, the misuse or mismanagement of which can cause significant harm.

In the same model as the Security of Critical Infrastructure Act, some private sector verticals can be considered higher risk for use of AI and ADM, and therefore requiring higher levels of oversight. The focus here should not be based solely on national security per SOCI, but also on individual safety (including psychological safety).

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

It is important first for the Australian Government to recognize the potential for harm to individuals that has not previously been high-profile: mismanagement of their records.

Under-retention of records affected the Windrush generation, Caribbean immigrants to the UK who had all their landing card records destroyed by the UK Home Office records team to save space, resulting in many being unlawfully deported. Keeping records for too short a time has affected the chances of reparations for the Stolen Generations, Maralinga atomic testing survivors, and child sexual abuse victim survivors in private and government institutions in Australia. Destruction of records is an example of an impact that is 'irreversible and perpetual'.

Over-retention of records has been newsworthy recently: Optus and Medibank spilled sensitive records that their previous customers did not expect them to still be holding. When the Australian National University was breached in 2018, 19 years of staff and student records were taken by a foreign government, 60% of which had been over-retained by the records team and should have already been destroyed. This breach was also irreversible.

And inability to discover information causes significant harm. The UK National Common Intelligence Application database held around 400 pieces of missed information about the man who would become the Manchester Arena terrorist bomber, killing 22 people, and injuring more than 800, most of them children. In April 2022 Medicare scrapped backlogged medical record requests for the previous six months, affecting around 2,400 people, who may need those records for insurance claims or to access the National Redress Scheme for child abuse, also saying their systems were just too hard to search. These outcomes caused very high impact on those individuals.

These harms have been suffered on the individual level, and have never been effectively addressed at the national level. The ANAO has been issuing reports highlighting ineffective Commonwealth records management since 2002, and the most recent ANAO report on records management has shown that data governance is actually worse than ever, with 93% of Commonwealth records unmanaged.

AI like ours has been developed in the last five years specifically to address this problem, and reduce the risk of harm – and it has been rapidly adopted to provide ADM for records protection, retention, and destruction decisions across government and corporates. The only way to manage information at scale, and start to protect records and the people they are about, is with AI. But there is a significant risk that this AI adoption introduces unexpected or unintended consequences of its own, and subsequently, more harm. AI empowers us to finally start making decisions about what to do with peoples' records at scale. If those decisions are wrong, closed box AI will not allow any recourse.

To support responsible AI practices across government agencies, those agencies need to know and recognise that the decisions they make with data can have real impacts on real people. And therefore, *any* use of AI for those decisions is inherently high risk. They must understand and internalise the first point before they will be able to put the second point into practice.

4.3 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.

We do not have a comment on this question at this time.

4.4 Given the importance of transparency across the AI lifecycle, please share your thoughts on

- where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?
- mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

Robodebt is the most recent and high-profile example of the importance of transparency for recourse, even though it was not an AI solution. All algorithms must be contestable, AI or otherwise. Incorrect algorithms gave rise to the Horizon postmaster scandal in the UK in the 1990s, but the harm from those algorithms was multiplied by the inability of the falsely accused to argue against the algorithm. Transparency must be mandated. The challenge will be to regulate the more subjective aspects of transparency – what is a sufficiently 'meaningful explanation'? The regulations may require a Man on the Clapham Omnibus test to set a precedent.

4.5 Do you have suggestions for:

- whether any high-risk AI applications or technologies should be banned completely?
- criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

We do not have a comment on this question at this time.

4.6 What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

We do not believe the public will avoid using AI. The market will drive adoption, whether the AI capabilities of commercial and government solutions are disclosed or not.

Implications and infrastructure

5.1 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?

There is always a tradeoff between human rights and optimized commercial outcomes. Australia should not consider undermining human rights for any trade advantage.

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

We do not have a comment on this question at this time.

Risk-based approaches

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

All decisions should be informed by risk. What is in question is the standardization of the mapping of that risk. 'Risk-based' approaches that allow organizations to self-assess their risk levels subjectively are likely to create perverse incentives to minimize risk levels, and subsequently dilute risk treatments.

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

Risk should be determined based on the kind of harm that can arise, in a similar model to the standardized PSPF Business Impact Levels tools used by the Federal Government. This mitigates the risk of risk minimization as it provides more objective parameters on which to determine real risk.

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

As AI ethics is about protecting citizens, the standardized approach should be the same, no matter who is using the AI. Risk should be mapped against human outcomes, before organisational ones. If an organisation could cause harm to an individual or social group as a result of ADM or AI, that ADM and AI must be explainable and contestable. It is the size and type of the harm that matters, not the size and type of the company.

6.4 What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C? (Attachment C is found on page 40 of the discussion paper)

We agree with the use of Impact Assessments, following a similar structure to the PSPF BILs, but also aligned with the OAIC Privacy Impact Assessment framework and tool.

6.5 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 government must be careful not to create feedback loops between different regulations, which point only to each other and do not provide material guidance.

There should be one source of truth on AI Ethics regulation, which should be addressed, and further contextualized, in all legislation relevant to citizen outcomes.

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

Closed box AI cannot be used for regulatory purposes until such time as it can be made explainable. Universal principles on Ethical AI require that

- Algorithmic-based decisions which affect individuals cannot be made without human oversight
- Algorithmic or automated decisions are explainable where they have a detrimental impact
- The explanations behind how the decision was reached are meaningful, useful, and valid.

This is not currently possible with LLMs, neural nets, or supervised Machine Learning. Relying on records of the training data originally used for these types of AI is not sufficient to the purpose, as they evolve constantly without human oversight.

We believe these types of AI have an important role to play, and a powerful utility for government and corporations. But they must not be used where the outcomes of their decisions can harm individuals.

6.7 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:

- public or private organisations or both?
- developers or deployers or both?

We believe that Responsible or Ethical AI must be mandated, and must apply to all organizations who hold data about citizens, and/or make decisions using that data which could detrimentally affect those citizens.