Palantir

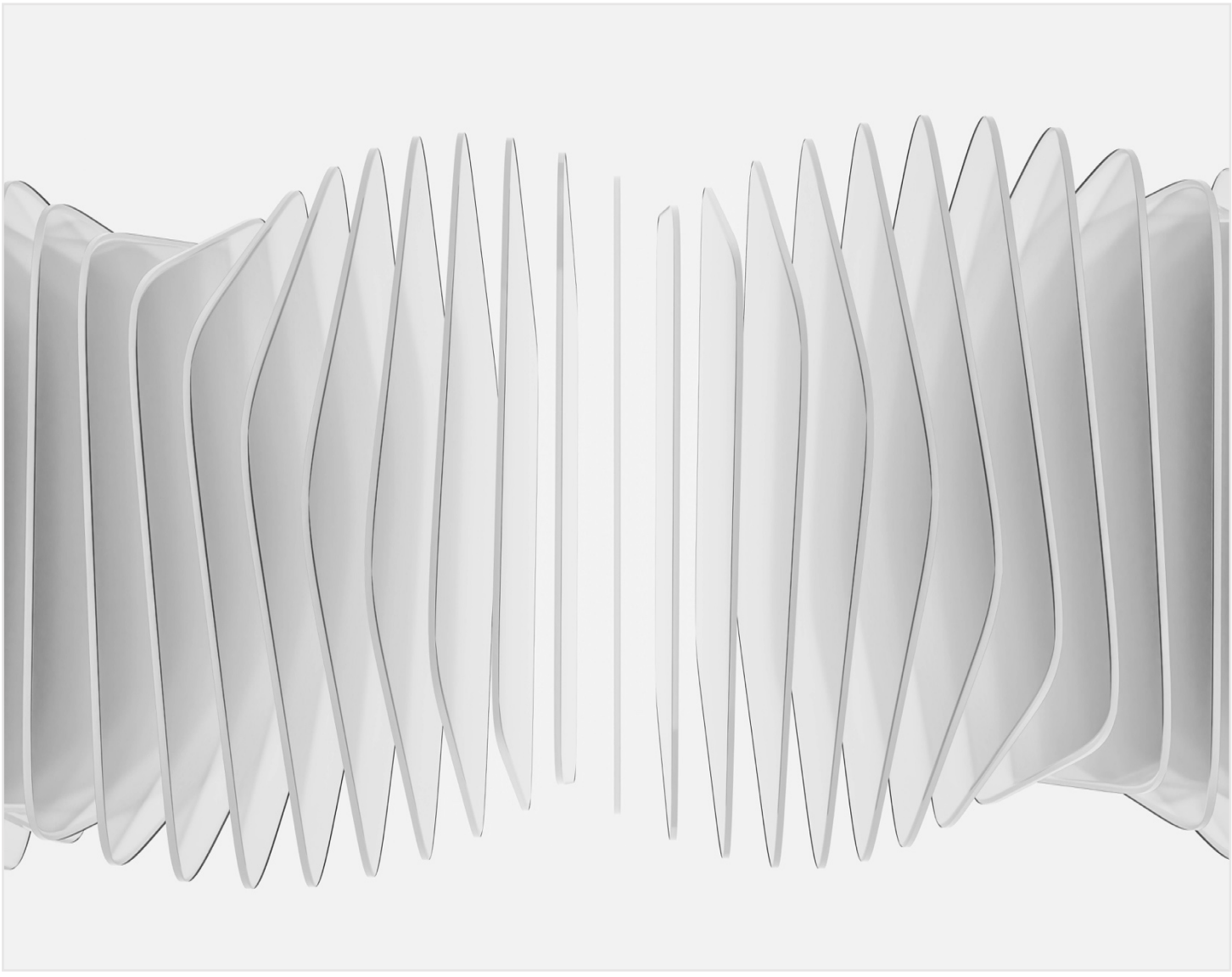
Palantir Australia Response to Supporting Responsible AI Discussion Paper

Department of Industry, Science and Resources

July, 2023

palantir.com

Contact:
alampert@palantir.com



Overview of Palantir Australia's Response

Palantir Australia commends the Department of Industry, Science and Resources for taking this step to better understanding and support safe and responsible AI Practices, building on work already being done across various parts of the Australian Government. We welcome ongoing engagement with the Department as we keep iterating towards appropriate AI regulation and policy.

Our responses are informed by Palantir's global experience and lessons learned from refining a practice of responsible and trustworthy AI development and deployment.

We have been guided by a few key design principles in our response. These include:

- We recognise that AI systems are highly dynamic, and technology is continuously evolving. Any regulation needs to be similarly dynamic and responsive to these changes. This will involve building in feedback loops of varying lengths - from short term in days or weeks through to longer-term measured in years.
- We believe that AI is most effective and defensible when employed to assist and enhance human execution and decision-making rather than to replace it.
- We see benefit in Australia adopting international norms and standards by default, and creating differing standards and approaches only by exception where truly necessary.
- We see the need to regulate end uses of AI rather than the technology itself.
- Notwithstanding the tension with the need to regulate end uses of AI, where possible, we welcome a broad-based approach to regulation. This may take the form of an over-arching Act, and then regulation relevant to specific end uses. We see benefit in minimising the boundaries between different pieces of regulation and legislation, to reduce the amount of time and regulatory effort that goes to evaluating boundary conditions to determine which regulations a deployed AI system is subject to. [0] Where context-specific regulation is required, we suggest keeping definitions and processes as common as possible.
- We see the enforcement of law to be at least as important as the law itself. Any AI regulation will require a well-funded regulator to be able to appropriately undertake enforcement.
- We consider voluntary commitments, pledges, and self-regulation will likely be insufficient to protect against the potential harms of AI [1]
- We believe that AI Ethics Principles are also insufficient to protect against the potential harms of AI, and that observations drawn from operationally-thoughtful and responsibly-constructed experiments in suitably constrained conditions are a more effective approach to understanding and responding to real-world risks.

As noted, we believe that AI is most effective and defensible when employed to assist and enhance human execution and decision-making rather than to replace it. Especially for applications that carry significant impacts on individuals' livelihoods, human rights, and well-being, the limits of AI must be acknowledged to help determine the right level of human intervention to ensure moral agency and culpability. As should be the case with the use of all technology, the impact of AI should be in elevating humanity, not in undermining, exploiting, endangering, or replacing it.

Our responses below are further grounded in Palantir's principled operational approach to AI, which places an emphasis on the contextual applications and practical challenges of AI (which are distinct from more theoretical or performative articulations of AI and its challenges). This framing is further expanded upon in our approach to AI ethics [2], in which we recognise at the outset that AI "does not exist in vacuum, but rather is inextricably tied to its contexts of application, its operational uses, and the full data operating environment that surrounds the much narrower AI components that tend to dominate the discourse of AI Ethics." [3]

[0] Examples of such boundary conditions that consume regulatory attention from the Commonwealth Privacy Act include assessing what is Personal Information, or whether a Small Business Exemption applies.

[1] We note and broadly agree with Australian eSafety Commissioner Julie Inman Grant's recent reflection that "Frankly, I don't think AI pledges are going to work." (Source: https://www.afr.com/technology/be-sceptical-when-big-tech-promises-to-self-regulate-ai-esafety-boss-20230725-p5dr08, 2023-07-25)

[2] https://www.palantir.com/pcl/palantir-ai-ethics/

[3] https://blog.palantir.com/the-efficacy-and-ethics-of-ai-must-move-beyond-the-performative-to-the-operational-1792e933b34

Response to Discussion Paper Questions

Definitions

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

Palantir Australia identifies a lack of precision and consistency in the definitions in the discussion paper. The definitions presented unfortunately conflate AI technology categories (or model types) with their applications. We recommend more clearly differentiating types of AI from the applications thereof. Keeping these concepts distinct provides a better framework for contextual considerations of both risks and benefits of AI use. The distinction between technology and application is important, since any regulation needs to be attuned to the use cases of applications in specific domains, rather than addressed to any generic technology or model itself. In the discussion paper, Large Language Models (LLMs), Multimodal Foundation Models (MfMs) and Automated Decision Making (ADM) systems are all presented as “applications” of AI “technologies” - this seems confused and misleading, as these are themselves broadly applicable classes of technologies that can be used within many different application contexts. Even with "general purpose AI", as Large Language Models (LLMs) are often framed, the evaluation of risks and benefits is highly contextual, and will depend greatly upon the intended use cases and applications.

Additionally, no mention or thought appears to be given to the software or interfaces that exist around any AI model - which are critical to how humans interact with such systems, and therefore have a significant bearing on the likelihood of risks manifesting as harms. We suggest adding consideration of these ancillary components, which include the hardware and software interfaces which control AI applications, as well as the technical controls which govern the use of AI applications (e.g., application access controls, user interface and interaction design decisions). Thoughtfully designed and implemented ancillary components have significant potential to mitigate risks and enhance benefits realised through AI.

Potential gaps in approaches

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?

Given recent media and public attention given to long-term, existential risks that may arise from the use and development of AI, we suggest it may be helpful to distinguish between such highly uncertain **existential risks,** and contrast those against **concrete harms** that are real, specific, and either have actually occurred - even if in a different context of use - or are readily foreseeable. Much more regulatory attention should be given to addressing and avoiding **concrete harms** than highly speculative and controversial **existential risks.** Calls to address the possible existential risk of super-intelligent AI risk both impeding the development of beneficial uses of AI, as well as distracting researchers, regulators, policy makers, and technologists from the importance of addressing nearer-term risks of more likely harm. These harms will vary depending on how and where AI systems are deployed, and are best addressed by regulation that accounts for different context of use.

We also suggest that any regulation would benefit from a clearer articulation of the space of potential harms and risk. The discussion paper coverage of Challenges (p 7-9) focuses mostly on issues of bias, ignoring other significant categories of harm and risk. These are far from theoretical - AI systems are already in wide, and rapidly growing, use across all manner of industries and problems spaces, and they can and do fail in many different ways across and within these sectors leading to real harms.

Within this space of concrete harms, we suggest more prominent consideration of provisions to address the inherent brittleness of AI technologies that is often ignored, misunderstood, or even purposefully misrepresented for commercial, programmatic, or publicity benefits and to the detriment of the institutions reliant on these technologies.

For example, Machine Learning (ML) techniques, as one prominent category of AI technology, often produce the suggestion to lay observers — by virtue of the very title — that ML systems are continuously learning and adapting to novel situations and parameters in real-time in ways that make them resilient to ever-changing application environments. The reality of the technology is very often far from that. ML describes the technique for building models that, once trained, lock in a set of parameters that remain fixed until the model is updated based on new data, model features, optimisation parameters, etc. This process of version control is a function of the fact the ML models, like other technologies, should in strict terms be expected to perform with some degree of predictability based on tested and validated specifications, and to only be permitted to operate within well understood boundaries. It is a requirement of addressing the inherent brittleness — i.e., the tendency of models be anti-resilient to changes in data, environment, application context, etc. — of ML models.

Despite this, a prevailing assumption — based on misconceptions of AI as human-like intelligence or Machine Learning as active human-like learning — is that these technologies are truly resilient, adaptable, and transferable to virtually any new environment. One treatment is to ensure that, at least for higher-risk AI systems, that multi-disciplinary teams of technologists, social scientists (including anthropologists and psychologists), and policymakers work together to advance the responsible design and use of technology, while safeguarding against risks of harms, and focus on building fault tolerance and resilience into technology systems.

In terms of addressing these risks through regulation, we recommend measures including:

- Ensuring that risk identification and mitigation for AI systems cover the entirety of fully-integrated, end-to-end AI systems, not just component AI tools and models.
- Ensuring that accountability and risk mitigation for AI systems are applied throughout the entire system lifecycle - with a specific focus on clarity of any AI model applicability and on-going monitoring, maintenance, and reporting.

We also note that accountability mechanisms are likely to be insufficient or even counter-productive unless there is clarity about who is ultimately responsible for the actions of an AI system. In the complex ecosystem of various AI developers, providers, integrators, operators, and users, it can be unclear who is ultimately accountable for the negative implications or potential harms of AI system use. Regulation needs to provide clear guidance for what is expected from each of the different types of actors in the AI ecosystem. Regulation needs to capture and emphasise responsibilities that lie with technology researchers, manufacturers, marketers, and developers for not just short-term testing and early deployment outcomes, but also long-term maintenance assurances and product liabilities for the sustained delivery of marketed results.

In addressing the above risks, a rigorous regime of *ex-ante* measures should be incorporated, complemented by *ex-post* measures such as audits. Such an approach can help build better trustworthiness into AI systems from the outset. Accountability measures need to be built into the entire process of AI development and deployment - starting as early as the collection or aggregation of training data by model developers - to help prevent potentially flawed, erroneous, biased, or otherwise unrepresentative or unsuitable data from becoming embedded into AI systems. Such *ex-ante* measures can ensure there is sufficient transparency of the workings of AI systems, as well as consideration given by-design to concerns around privacy, safety, robustness, and fairness of resulting AI systems. Absent such measures, it can be difficult to identify and address issues of bias, discrimination, and other ethical concerns in a timely manner.

Accountability measures also need to extend into the entire AI system, including supporting infrastructure. Without well-functioning system architecture that provides granular access controls, logging, monitoring for abuse etc., AI systems can directly produce, or indirectly facilitate, unintended consequences that cannot be mitigated *ex-post*.

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.

As we have publicly remarked [0], Palantir believes that meaningful progress towards controlling AI and harnessing it for human flourishing does not come from pausing experiments, but rather - and perhaps somewhat counter-intuitively - from leaning into the fielding of operationally-thoughtful and responsibly-constructed experiments in suitably constrained conditions that force us to identify and confront the real challenges of technologies *in situ*.

This operational "field-to-learn" approach to AI deployment provides one template for ways that governments can - through funding or other convening authorities - provide frameworks that enable both technical innovation *and* legal and ethically accountable boundary setting. It does so by better exposing technologists, ethicists, policy-makers, social scientists, domain experts, and AI users to the specific challenges of AI deployment and use, as opposed to more theoretical musings that, while interesting, are often untethered from the reality of both the technology and the operational setting. Those benchmark realities are essential components of constituting AI regulation that actually work - that is, that address the practical and real challenges of AI technologies in real-world use. Of course, the field-to-learn experimentation we are suggesting should be closely monitored and governed by principles that ensure safety, accountability, and especially transparency, even before these principles are codified into AI regulation.

[0] Courtney Bowman, Appearance at UK House of Lords Committee on AI In Weapon Systems, Palantir Blog (2023), https://blog.palantir.com/appearance-at-uk-house-of-lords-committee-on-ai-in-weapon-systems-2354862a6641

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.

One challenge is that it takes time for issues to emerge and rise up through the regulatory system (complaints need to be made, issues investigated, etc.). There needs to be proactive seeking out of detailed case studies that can be learned from - including both successful and problematic deployments of AI systems, in real-world contexts. Forums that explore these case studies need to include

relevant regulators as key stakeholders, along with input from representatives of relevant and diverse stakeholder groups - e.g., Aboriginal and Torres Strait Islander representatives, Disability representatives, children and youth, as well as application domain experts, civil society representatives, technology developers and technology deployers. Stakeholder groups that rely heavily on volunteers would greatly benefit from being compensated for their participation. The Responsible AI Network based out of the National Artificial Intelligence Centre could be leveraged for coordinating the inclusion and participation of some of these stakeholders.

Active industry participation in bringing evaluative case studies to the fore could be encouraged by providing limited regulatory safe harbor or immunity to organisations who engage in good faith by inviting government stakeholders directly into collaborative development and deployment environments in which their feedback and ideas can be considered and integrated. This would foster an operational field-to-learn approach to more deeply and pragmatically understand AI technology and its attendant risks. To ensure maximum value from such engagements, and to avoid the risk or perception of regulatory capture by industry, it would be especially important for such activities to operate with a high-degree of transparency.

Additionally, to understand the effects of existing government and industry efforts towards encouraging responsible development and use of AI, there should be some attempt to measure the effectiveness and impact of existing efforts, including the Australian AI principles, and the Australian-ratified OECD principles. We expect that AI Ethics principles are only useful at a high-level; to make them operationally effective, they need to be transposed into standards and/or practical guidance and support that are more application and context-specific. It would be excellent to see concrete and measurable goals established for the coordination efforts under the auspices of the National AI Centre, including tracking efforts by authorities (standards bodies, regulators, industry organisations) to produce and promulgate operationally useful guidelines that drive responsible AI efforts in their respective fields.

An additional goal of this coordination activity should be to educate and engage with end users - both broadly, and with application-specific activities. This education should seek to appropriately level-set expectations, identify and provide guidance for understanding risks, potential failure modes, as well as to inform users' and data subjects of their rights and avenues for redress. Also critical will be genuine engagement that seeks out, listens to and acts on input from end users and representative bodies to identify and transparently respond to issues as they arise. Such consultation should continue beyond the initial deployment of any AI system. These education and consultative activities should be a mandatory part of larger-scale commercial and government use of AI systems that interface with the public.

Responses suitable for Australia

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?

Australia, lacking a Federal Human Rights Act or Bill of Rights, will need additional rights spelled out for data subjects and users of AI systems. One of these is a right to contest automated decisions from AI or ADM systems. Looking abroad, there is already some normative convergence around this idea. Both the EU GDPR and the recently released US White House AI Bill of Rights include a form of this idea. We encourage this to be considered in Australia's approach to AI regulation.

For high-risk AI systems, regulation should follow the precedent set by the draft EU AI Act and require ongoing risk mitigation across the operational life of the system, and mandatory notification of accidents or other incidents to regulators. Regulators should also be able to step in and require the recall of a system if harms arise that cannot be adequately prevented or remedied. Other treatments that should be considered include mandating relevant types of system redundancy, and kill switches.

Overarching such requirements should be an approach to the design of high-risk AI systems. There is a place for design standards to be applied, informed by both existing tracks of AI standards, as well as by detailed case-studies and analysis of AI system successes and failures. The disciplines of Human Computer Interaction (HCI), Human Factors, and Systems Engineering will have many ideas to offer that should be factored into any design standards.

Target Areas

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

AI adoption already spans across the public and private sectors. Risks from these systems will be realised within and across both sectors, and both low-risk and high-risk systems can be deployed in both ecosystems. There is also often complex interdependency between public and private sector organisations for AI systems. Large, multi-national technology companies increasingly mediate, influence, and profit from the lives of a very significant portion of the Australian population. This includes in public sectors such as Education, and Health. Across all sectors, Government agencies are almost always customers of private sector technology companies when deploying AI technology. Given these factors, a reasonable starting point is that regulation should apply equally to both public and private sector uses of AI.

Harmonising AI regulation across the private and public sector, as well as across state and federal governments, could help remove some barriers to responsible and effective AI adoption and promote consistent and appropriate accountability for Australian organisations and developers. There is precedent for this approach - the Commonwealth Privacy Act applies to both federal public sector and private sector organisations. Drawing together existing AI regulation and assurance frameworks - such as the NSW Artificial Intelligence Assurance Framework already operating successfully for the NSW government - at the federal level could similarly improve interoperability and consistency as well as reduce unnecessary duplication and potential inconsistency in costs of and approaches to compliance. The same might be considered for the plethora of AI ethics principles that have proliferated over recent years.

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

As has been done for managing privacy and data protection, the Government could create a binding Responsible AI Code for Government agencies - similar to the Australian Government Agencies Privacy Code for public sector agencies. This would go above-and-beyond the mandatory requirements of any AI regulation, with additional requirements to support the responsible deployment and use of AI with additional attention to designing and operating trustworthy systems.

The Government would be well advised to provide ongoing and appropriate budget to regulatory agencies charged with overseeing AI system deployment and development, to allow them to credibly exercise their powers and responsibilities, as well as to commission and conduct research, to engage with relevant experts and stakeholders, as well as to enable them to access critical resources to up-skill staff and to hire and retain staff with the required technical

and policy expertise. Funding should also be directed to fora that allow collaboration between sector-specific and other relevant regulators. We also suggest that the Government consider giving necessary funding and resources to an existing Regulator rather than establishing a new AI regulator, even if initially on an interim basis to be reviewed after 12 months. This could help more quickly bootstrap any regulatory efforts.

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.

Measures to mitigate or manage risks from AI are most sensible and practical when established and evaluated not for specific technologies - which can change rapidly - but grounded in specific contexts of use. Those contexts will best define the relevance and priorities of intended and achievable goals. Not all methods of managing or reducing harms will apply to all systems, and some approaches will, in fact, be at odds with others. In the case of incommensurable objectives for managing risk, trade-offs will be necessary. This process should recognise that an absolute attainment of some goals (e.g., eliminating all discrimination risk, or a total reduction of privacy risk) may be unreasonable in context, and often the objective should be directed more towards tolerable risk mitigation rather than complete risk cessation. Some goals and risks may need to be sensibly traded against others. A critical point, however, is that these trade-offs and corresponding consequences are openly acknowledged and explained, and are explicitly balanced between corporate, government, and individual human interests, preferably with a bias towards the last of these.

That having been said, certain themes and approaches can be applied across disparate technologies and application contexts. Given the broad application and significant power that rests in data, a trustworthy and secure data foundation is universally required. This should include (but is not limited to) high-quality data integration, pipeline management, data quality checks, system audit logs, data security and access controls, the ability to version and branch data, and collaboration features to annotate datasets and identify addressable issues over time (e.g., statistical and other forms of unwanted data bias), all done in a way that respects the individual, their individuality, agency, and supporting all that, their privacy.

Recent advancements around Large Language Models (LLMs) continue to demonstrate our broader point that the most important regulatory efforts will need to be more precision oriented, focusing on applications of the technology, rather than attempting to regulate the technology itself independent of its uses. Models - LLMs or otherwise - aren't inherently ethical or unethical; it's in how they affect the world that both harms and benefits arise. Some of the necessary guardrails are fairly universal (e.g., for Generative AI, limiting the ability to produce hate speech, or similarly illegal or objectionable content) and in those cases, accountability should lie mostly with model developers; many other

guardrails are application and context specific and require other actors (domain experts, system implementers, governing organisations) to assess and mitigate risks appropriately. Again, we expect that the most significant risks and challenges of real-world LLM and AI use will occur in the deployment to specialised, context-specific use cases where value is generated and potential harms are most present.

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

1. **where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?**

Transparency is necessary but not sufficient.

It will not be effective, for example, to address LLMs and complex ML models with transparency of their internal decision processes - it's the wrong treatment and won't really help end users. The notice-and-consent paradigm for managing privacy risks has failed in the face of complex and sophisticated uses and transfers of personal information that go beyond the ability of average system users to reasonably understand the implications of verbose privacy notices, let alone actively enforce given the often huge power asymmetry. In a similar way, providing transparency into the inner workings of AI models is unlikely to help system users provide informed consent or understand the implications or risks that stem from their use of an AI system.

Requiring transparency around data collected and used to produce those models, as well as transparency around the auditing and assurance of those models - across the full lifecycle of developing, testing, deployment and ongoing monitoring - would, however, significantly help build public trust and confidence in AI. Similarly, the publication of detailed impact assessment, audit reports, and other research findings should be required wherever possible.

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

In the cyber realm, mandatory breach notifications have become standard in recent years - notwithstanding the lessons to be drawn from multiple overlapping laws that have introduced similar but disjoint reporting obligations. In the privacy realm, publishing Privacy Impact Assessments (PIAs) for projects that involve any new or changed ways of handling personal information that are likely to have a significant impact on the privacy of individuals - regardless of the technologies involved - has been a requirement for the public sector under the Australian Government Agencies Privacy Code. While the requirement should be stronger, and the regulator more adequately resourced to ensure compliance, requiring agencies and AI providers in the private sector to divulge the data and

technology in use and assess a range of risks is worth pursuing.

Do you have suggestions for:

1. **whether any high-risk AI applications or technologies should be banned completely?**

We must acknowledge that AI technology is likely here to stay and focus on how to use the technology safely and responsibly, without turning a blind eye to the risks it poses or hoping that the technology will not be used in certain ways. Our Australian values should lead us to categorise some uses of AI technology and applications as presenting an unacceptable risk, regardless of any potential benefit. Similar to the EU AI Act, social scoring, for example, should have no place in Australian society. AI systems that use biometrics for certain forms of identifying individuals - one-to-many facial recognition, gait recognition or similar - should be restricted to public sector use, and have very stringent limitations on its use as well as genuinely independent reporting and oversight requirements that accompany any use.

1. **criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?**

There are useful starting points in the EU AI Act categorisation, namely:

1. AI systems that deploy subliminal techniques beyond a person's consciousness to materially distort their behaviour in a way that could cause them physical or psychological harm.
2. AI systems that exploit any vulnerabilities of a specific group of persons due to their age, physical or mental disability, to materially distort their behaviour in a way that is likely to cause them physical or psychological harm.
3. AI systems used for social scoring by public authorities that lead to detrimental or unfavourable treatment of individuals or groups in ways that go beyond what is allowed under EU law.

Further proposals for banning AI applications or technologies, as with other proposed legislative changes, should involve genuine consultation with the community, industry, civil society, and other stakeholders, to ensure such decisions reflect community norms and expectations.

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

A key, and unfortunately often overlooked, requirement is to equip regulators with necessary resources - including budget, technical expertise, and robust legislation - to provide genuine oversight and guidance for AI system deployment

and use. This will require both strong ex-ante regulation that mandates the appropriate reduction of risk when AI systems are developed and deployed, as well as rigorous ex-post powers to detect and sanction organisations where harms arise.

Implications and Infrastructure

How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

Banning unacceptably-high-risk activities is unlikely to have a significant negative impact on Australia's tech sector. Australia should contribute to global consistency and efficiency by joining other countries already considering banning such activities; it is likely, for example, that any activities banned in Australia would also be banned in the EU, which would be a more significant factor for Australian exports and trade.

Overall, regulation that bans specific uses of AI technology provides a helpfully clear signal for commercial organisations to respond to.

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

No response.

Risk-based Approaches

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

As already noted, we believe that regulation should relate to the context of its application and context-specific concerns (historical, social, cultural, ethical, etc.). For this reason, in defining context-specific regulation, it is important to recognise that unique contexts will have intrinsic areas of concern or risk (including privacy, civil liberties, fundamental rights, sustainability, equity, inclusion, diversity, etc.) which should be directly factored into the chosen or mandated regulation approach.

That said, all Australian laws - and the enforcement of those laws - are ultimately risk based, given finite resources for enforcement. Broadly speaking, we believe a risk-based approach is a reasonable starting point for AI regulation, with goal of ensuring that regulation and its enforcement is applied first to areas of highest risk and intensity of impact.

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

A risk-based approach offers many advantages, especially when it is designed as a process to be repeated and altered as risks and knowledge change, rather than something to be completed once and not monitored or revised over time.

A few thoughts on addressing potential limitations:

- A risk-based approach does not inherently address any aspects of compensation or civil recourse if harms arise. It would be highly beneficial for these to form part of any AI regulation.
- Coupling human rights impact assessments and/or algorithmic impact assessments, can help identify and address contestable and hard-to-quantify harms that might otherwise be missed in risk assessment.
- As suggested earlier, there is a place for design standards to be applied, informed by both existing tracks of AI standards, and by detailed case-studies and analysis of AI system successes and failures. The disciplines of Human Computer Interaction (HCI), Human Factors, and Systems Engineering will have many ideas to offer that should be factored into AI system design. Some of these approaches could include requirements for hazard analysis through Software Safety Plans, especially for AI systems deployed in safety critical contexts.

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

In theory, a consistent and uniform approach to a AI regulation across sectors could help ensure that all stakeholders, regardless of their industry, are held to the same standards and expectations. This could help to prevent confusion, uncertainty, and potential harm to individuals or groups, and could promote consistency and clarity in AI development and deployment. However, such a uniform approach is likely to have significant limitations. Regulatory requirements that are widely applicable will tend to be articulated in broad terms, often as generalised principles rather than as concrete standards for action. Ultimately, any principles will need to be transposed into specific use case requirements to address the reality that most applications of so-called general use AI technologies in fact carry unique requirements or considerations for regulation. Healthcare, for example, will have specific privacy concerns related to patient data, while finance may have specific regulatory requirements related to fraud prevention.

Considering these differences, it is likely that some sectors will require a more precautionary, non-risk-based approach to AI regulation. AI systems that are proposed to be a safety component of already regulated products - such as medical devices, cars or aircraft - pose risks of measurable physical safety harms. Such systems should be subject to the same regulations as the products in which the systems are embedded, as well as needing to comply with any additional requirements of AI regulation. These product-specific and sector-specific regulations are likely to be precautionary, and go beyond identifying and mitigating risks towards proactively imposing regulatory decisions including specific bans, requirements for licensing and accreditation, and other measures that give precedence to the protection of public health, safety, environmental or other concerns over economic interests.

Under the EU AI Act, the conformity assessment process for high-risk systems requires inspection by external, third-party “notified bodies” who validate that the provider is in compliance with the Act’s requirements. These notified bodies have significant fact-finding and inspection abilities. If the notified body finds a high-risk AI system to be in conformity with the requirements, it will issue an EU technical documentation assessment certificate, which has limited time validity and can be suspended or withdrawn. In essence, this is a licensing-like scheme for high-risk AI systems. A similar regime could be established for high risk AI systems in Australia.

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

The elements outlined in Attachment C are generally sound.

Some additional comments:

- Explanations will vary in their effectiveness and appropriateness based on the level of detail, context of use, and complexity of the system. Relying on explanations to build trust pushes the onus back to end-users to assess and manage risk, which is very likely not reasonable or feasible in many circumstances, and almost certainly risks information overload as AI systems proliferate across a larger number of application contexts. The notice-and-consent model that this element builds on arguably suffers from the same systemic weaknesses despite its prominence in existing Australian privacy regulation
- Standards and/or certifications (e.g., trust marks) are another important element that should form part of a risk assessment. Compliance with emerging and existing Australian and/or global standards for AI systems should serve as a form of risk mitigation
- External assurance mechanisms are another important element that should be considered as part of Monitoring and documentation, especially to manage the ongoing risks across the full lifecycle of the AI system. Third-party verification / auditing of system operation and performance should be mandatory in higher-risk settings before deployment. For lower-risk systems, verification could be self-certified. Given the likely resource constraints on any regulator, we suggest these approaches in preference for a formal licensing scheme that would carry higher regulatory overheads.
- While human-in-the-loop oversight can be critical in many use cases, we agree with the observation that this may not be appropriate, possible, or necessary in all circumstances. Human factors need also to be considered to ensure that any human-in-the-loop is actually effective - humans-in-the-loop in autonomous vehicles, for example, can end up making emergency situations worse when automation fails, because they cannot reasonably be mentally and physically prepared to intervene after long phases of inactivity.
- Testing should also be a significant component of risk assessment and mitigation. Testing can take the form of both validation testing ("is the product correct according to its specification") and verification testing ("has the product been built correctly"). Any testing regime also needs to extend across the full life-cycle of the AI system, especially to address ongoing monitoring and maintenance requirements for AI systems, that are prone to specific brittleness issues that we outlined earlier.

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?

One approach would be to incorporate assessment of AI risks into existing risk assessment processes wherever possible - alongside Privacy Impact Assessment, Human Rights Impact Assessment, Algorithmic Impact Assessments and similar processes.

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

Risk-based approaches to regulating general purposes AI systems such as LLMs and MFMs are more effective when linked to specific use cases, rather than to the technology itself. LLMs offer additional regulation challenges, however, in that they are specifically designed to be highly accessible, easy to use, and readily applicable across very different application domains. Such systems have swiftly been adopted into workflows - both productive and malicious (e.g., producing misinformation at previously unprecedented quality and scale). Despite this generality, the harms and risks remain highly contextual, and are best addressed for specific use cases and domains.

Consequential applications that deploy LLMs should require governance mechanisms that mitigate and manage the inherent pitfalls of these models, and should clear acknowledge that that, at best, LLMs provide a stochastic mimicry of understanding. They should also take account of the propensity of such models to induce levels of end-user trust that exceed the system's competence.

When it comes to specific approaches for regulation these models, there are a range of possible treatments by developers, deployers, and users that could be encouraged or required to address risks. These include:

- Transparency around when users are interacting with such a system - specifically to address expectations given the fluency and human-like realism of some outputs from these systems. Similarly, outputs should be watermarked, tagged, or otherwise marked with clear attribution of being produced by automation.
- Transparency from developers about data used to train models and known limitations of what is being deployed.
- Transparency from deployers / system integrators (to the deploying organisation rather than end-users) about the potential harms anticipated, specific risks identified, and how these have been mitigated or minimised.

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

Mandated through regulation. There is copious evidence that self-regulation will not be enough to combat profit motives to mis-apply and carelessly deploy technology.

As just one example, the eSafety Commissioner, Julie Inman Grant, recently noted that tech giants have a poor track record of enforcing voluntary pledges. [0] We agree that commitments that do not legally bind companies to action are unlikely to be an effective way of protecting against the potential harms of AI.

As evidence, Inman Grant further notes that "more than 30 major technology companies signed up to combatting CSAM (child sexual abuse material) with the Five Eyes governments [Australia, Canada, New Zealand, the UK and US], and none of them were living up to the pledges that they signed up for." [1]

[0] https://www.afr.com/technology/be-sceptical-when-big-tech-promises-to-self-regulate-ai-esafety-boss-20230725-p5dr08, 2023-07-25

[1] Ibid.

And should it apply to:

1. **public or private organisations or both?**

Both, as we've outlined earlier.

1. **developers or deployers or both?**

Both, but with clearly articulated spheres of responsibility covering both individual and joint obligations and accountability.

For developers, AI regulation should be careful to avoid privileging infrastructure that gives exclusive advantage to a small number of big tech companies. For this reason, consideration should be given to imposing strict liability upon for-profit model developers of general purpose LLMs and similar models. Open-source and non-commercial model developers should be subject to limited liability.