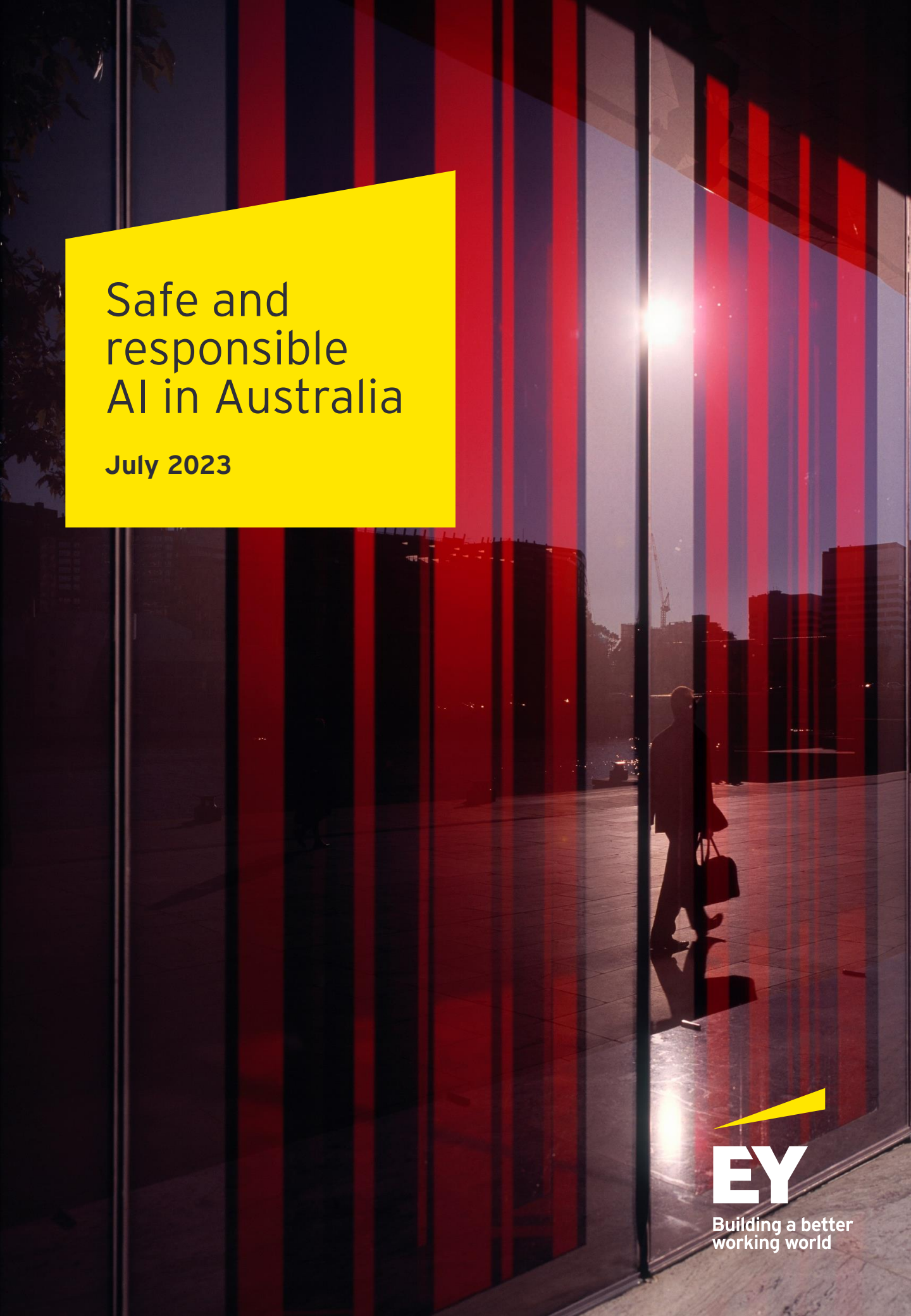


Safe and responsible AI in Australia

July 2023

EY

Building a better working world

Building a better working world

The Hon Ed Husic MP
Minister for Industry and Science
PO Box 6022
House of Representatives
Parliament House
Canberra ACT 2600

26 July 2023

Dear Minister,

Safe and responsible AI in Australia

We welcome the opportunity to respond to the consultation questions in *Safe and Responsible AI: Discussion Paper* (the Discussion Paper), launched by you on 1 June 2023.

Artificial Intelligence technology has undergone rapid development in recent years, and has the potential to revolutionise productivity and democratise a wide range of important capabilities. Businesses and government are poised to unlock this value. It is critical to position Australia at the forefront of this opportunity while maintaining a cautious stance towards potential social harms.

The time for establishing a framework for safe and responsible AI use is now. A robust governance framework based on ethical principles can rise to the challenge of rapid change, and position Australia as a leading digital nation by 2030.

Our responses represent extensive global experience partnering with government and industry in addressing AI's challenges and opportunities. They are not based on any specific scope of work undertaken by EY in Australia or elsewhere.

EY supports the adoption of a consistent nationwide risk-based regulatory framework. We also recommend the establishment of an agency tasked with championing responsible AI, oversight of governmental use of AI and the ability to issue binding market regulation.

We thank you and the Department for the opportunity to participate in public consultation on these important issues and would welcome any opportunity to contribute further to the development of safe and responsible AI in Australia.

Kind regards,

Catherine Friday
Oceania Managing Partner, Government and Health Sciences
Global Education Leader
EY

Definitions

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

EY agrees that definitions of AI should be aligned to international standards wherever possible, in order to facilitate interoperability with Australia's major trading partners. Accordingly, we support the definitions of Artificial Intelligence (AI) and Machine Learning as being based on ISO standards.

In some cases, definitions may still be evolving, or further clarity can be conveyed.

Large Language Model (LLM) and Multimodal Foundation Model (MfM)

The Discussion Paper's definition of LLM focuses on the model's outputs. It may increase understanding to further detail how these models can:

- Analyse text (at each stage of their development).
- Generate human-like language outputs.
- Interact with systems that require text input to perform tasks (including programming).

The definition of Multimodal foundation model could be similarly adapted, and both definitions could refer directly to machine learning (of which the current generative AI systems are a subset). This would result in the following definitions:

A **Large Language Model (LLM)** is a type of machine learning model that can process and generate human language to work with content and perform tasks.

A **Multimodal Foundation Model (MfM)** is a type of machine learning model that can process and generate multiple data types, including text, images and audio, as both inputs and outputs, and which is optimised for generality and versatility of outputs.

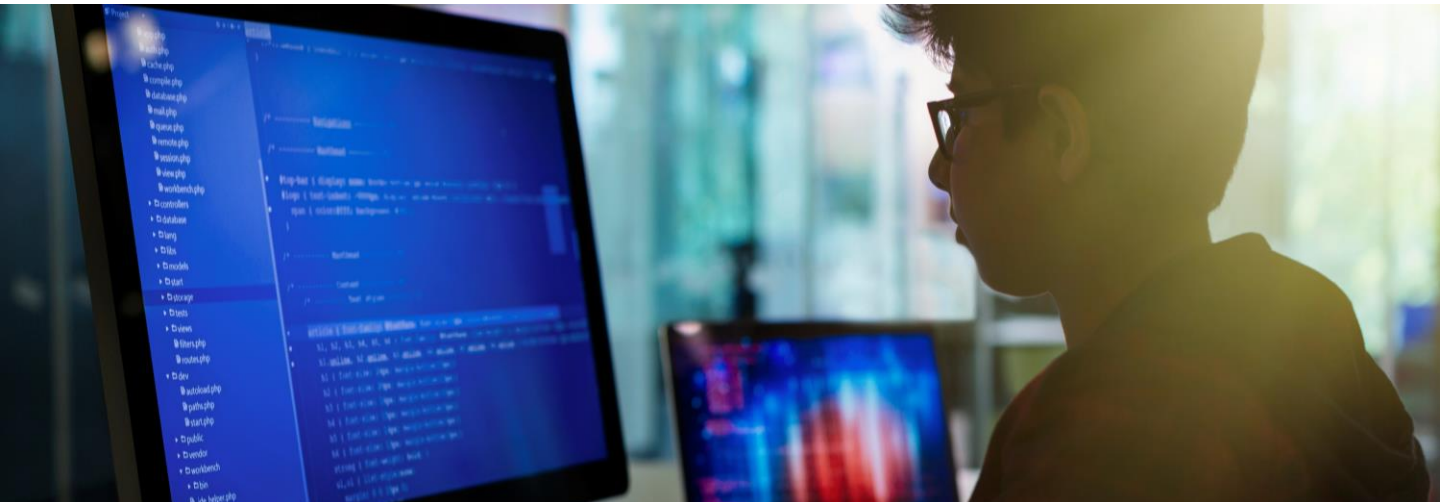
Although the terms are industry standard, the references to large in LLM and foundation in MfM are somewhat misleading. In the long term, it may be preferable for the sector to move towards more general terms such as "intelligent language model" and "intelligent multimodal model."

Automated Decision Making (ADM)

As noted in the Discussion Paper, the proposed definition of ADM covers a broad use of technology to assist in any decision-making process.

Using AI for decision-making raises different issues to, for example, a conventional credit-scoring algorithm, or systems that retrieve data from a database using simpler information retrieval techniques that might still be said to "automate aspects of the fact-finding process". Accordingly, it may be useful to use another term to refer to ADM when it is powered by AI, such as "AI-driven decision-making", which has appeared in academic papers. This approach may help to avoid the interchangeable use of ADM and AI noted by the Royal Commission into the Robodebt Scheme.[1] We would recommend against using the term ADM in legislation without further definition, given the very wide range of behaviour that it potentially captures.

[1] *Report of the Royal Commission into the Robodebt Scheme*, page 472.



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?

Many AI risks arise from its ease of access and ability to quickly generate output at scale. At a conceptual level, much of this activity is already controlled by existing laws and regulations, including privacy laws, cyber and data protection, anti-discrimination, tort, criminal law, deceptive practices, product liability, and so on. There could be substantial challenges at a practical level, however, as the regulatory system implicitly assumes some limits around the volume of activity to be controlled. Examples of the kinds of activities that could strain enforcement include:

- The automated generation of massive numbers of high-quality deepfakes, for the purposes of personally targeting individuals for embarrassment, fraud or influencing public opinion
- The widespread availability of easy-to-use tools and resources for computer hacking. The automation of these tools could also greatly increase the volume and speed of attacks (noting that the current average time for detection of a manual cybersecurity breach is in the order of 90 days).
- Greatly increased volumes of high-quality, personalised spam that is designed to evade detection (e.g., robocalls)
- The widespread availability of information and planning capabilities that are dangerous to the public and previously difficult to obtain (e.g., information on the manufacture of certain kinds of weapons or illicit drugs that may evade the bounds of existing prohibitions)

Noting the breadth of the issue, it is not feasible to make specific recommendations about what changes to legal frameworks will be required to address these concerns within the scope of this submission. We do recommend that the government continue to undertake detailed consideration and consultation, ideally before significant harms emerge. Regulatory responses could include requiring developers to include 'content guardrails', mechanisms for user feedback or guidelines for 'red team' testing of AI products and services.

AI also raises novel risks that don't necessarily have direct equivalents in existing harms. It is difficult to generalise about these as the implications of the technology are evolving and take time to be understood. But they could include, for example:

- Opaque decision-making that is based on errors or bias. These will be partially addressed by requiring further transparency in decision-making, which may involve further technical innovations.
- Novel privacy violations through the detailed profiling of individuals or groups using innovative inference methods that may not align with current understandings of private information
- The dependence on a small group of suppliers. Leading-edge AI technologies are being developed by a handful of global companies. Although barriers to model development are reducing, advancing the boundaries of the technology still currently requires substantial financial and intellectual resources. Further concentration into the hands of fewer companies, combined with widespread use of particular models, could create systemic risks to the economy and society.
- Risks associated with "natural" (i.e., unstructured) decision-making. Leading developers are envisaging a near future (three to five years) in which AIs are making decisions in a range of environments outside of a conventional software process (and at a massively increased pace) such as reaching broad diagnostic conclusions and making legal recommendations). This has implications for a wide range of regulations addressing responsibility and liability, for example the Corporations Act 2001 (Cth). Such AIs may well come into contact with each other, sharing information and responding to each other's actions, creating further challenges in attributing responsibility for outcomes.
- The risk of adverse unknown outcomes from progressively more advanced AI. AI is a novel technology that has undergone massive rapid development over recent years, and the pace appears to be accelerating. Although some existential risk scenario may sound far-fetched, many credible industry figures have raised concerns, particularly in the event of recursive self-improvement in which AI systems rapidly iterate their own future versions. There are also unpredictable interactions with other developing technologies, with, e.g., a potential for AI to experience substantial further increases in capability due to the expected commercialisation of quantum computing.
- Risks associated with input data quality. AI model development depends on ingesting vast quantities of data. If the data is of poor quality, or reflects historical biases or prejudices, there is a risk that such biases and prejudices may be perpetuated, with potentially widespread consequences.

Our recommended regulatory actions to address these issues are set out in our responses to the other questions in this paper.

Some further risks are outside the current scope of consultation, but merit noting as issues for future consideration:

- Intellectual property rules for the use of datasets to train large-scale models, noting that it may not be possible to directly attribute any particular data to an AI system's outputs.
- Potentially substantial impacts on the labour market.
- Increased exposure of the financial markets to new styles of algorithmic trading, with the potential to change market behaviour and increase volatility.
- The role of AI in national defence by Australia, allies and potential adversaries.

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.

A number of non-regulatory initiatives could be implemented to support responsible AI.

- The government can fund collaborative research, encourage the sharing of resources, and continue to support the development of internationally aligned standards and best practices across government, academia and industry, as it does already through its funding of, e.g., CSIRO's National Artificial Intelligence Centre. This could extend to supporting research into measuring AI's compliance with environmental standards and fundamental rights, societal impacts of AI on the labour market (job redesign, job transition) and more broadly.
- Education will be critical to ensuring that Australia has the skills needed to participate in AI and to build public trust. The government can support training opportunities across the range of needed technical skills, and encourage its responsible use in the university sector. Education can also play a vital role in improving access to skilled work in the sector for disadvantaged groups.
- In addition to skills training, consideration should be given to supporting programs to educate the public about AI by raising awareness and addressing myths. Information packages could also be developed to support better, more accurate media coverage of AI myths.
- Governments should actively consider their own training needs to ensure that they are informed and capable buyers and users of AI services. We note that, during this open consultation period, the Digital Transformation Agency issued Interim Guidance for agencies on government use of generative AI platforms. We support this initiative to provide clear internal procedural guidance for application across Commonwealth government agencies (and indeed as an example for the private sector). Government could create roles within the APS at a senior level to oversee AI initiatives with a focus on safety and responsibility.
- Government should continue working with international partners to promote the responsible use of AI (e.g., through the Global Partnership on Artificial Intelligence established in 2020) and aligning Australia's standards and practices (e.g., through the work of Standards Australia). Businesses should be provided with easy reference to preferred technical standards to facilitate compliance.
- Consideration could be given to establishing an independent panel of recognised experts to evaluate significant / new generative AI solutions that are to be commercialised in Australia (with the ability to recommend sandboxing / other testing procedures before widespread implementation).
- Consideration could be given to establishing a senior governmental presence in the sector to provide appropriate leadership across the Commonwealth, states and Territories, such as a committee of the National Cabinet.

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.

Cohesive governance across government will be important, both within the Commonwealth and across Australian Governments. A consistent approach will foster:

- The identification of opportunities and mitigation of potential harms
- Innovation and learning
- Growing the market for AI implementation by reducing the need to comply with conflicting rules and processes

Many of the recommendations made in our response to question 3 are relevant here. Ethical principles and frameworks can provide important guidance to agencies developing AI applications, as can the adoption of standards and providing tools and systems to validate the operation of AI systems.

Our recommendation to establish a senior governmental presence to provide leadership and direction, mentioned in our response to question 3, can also play an important role, noting that this should cascade down into coordination at the level of key departments.

Within the Commonwealth, consideration should be given to establishing a task group within one of the central agencies or a key delivery agency to advise on AI opportunity identification, prioritisation, planning, procurement, delivery and ongoing management. This could serve as a centre of excellence and share knowledge across departments, and would be established for a term sufficient to build up the internal capability of all departments with ongoing AI requirements.

Responses suitable for Australia

5 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 Discussion Paper provides a comprehensive summary of approaches in most of the jurisdictions that Australia would usually consider comparable, or that are relevant from a trade perspective. A risk-based approach to AI regulation is consistent with the approach in most of these markets.

Target areas

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

For both the public and private sector, a risk-based approach suggests that regulatory responses should be proportionate to the risks involved. In general, the public sector has a greater impact than any individual business, due to the importance of many of its services and the scale and monopoly nature of its provision. These factors alone suggest that, even if the overall approach is the same, the public sector should be held to the highest standards in the practical application of a risk-based approach.

Several further factors suggest the public sector should be adopting higher standards for itself than for the private sector:

- The direct nature of the accountability of government to its citizens, unlike the relationship between the private sector and its customers
- The heightened ethical obligations arising from government's combined role as legislator, regulator and actor within the regulatory system
- The differing roles of the public and private sectors in driving innovation (with the public sector most focused on creating the right conditions for innovation, and its funding role), at least insofar as innovation might place at risk the achievement of other priorities and obligations

This approach should be embedded in the public sector's AI regulatory framework, particularly in relation to transparency, the pace of adoption, and establishing rights to complain and obtain an effective remedy.

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

Our response to this question is similar to our responses to questions 3 and 4. We support:

- Adopting ethical principles and frameworks to guide agencies in AI implementation
- Establishing a task group within a department or agency to advise on AI matters, as mentioned in our response to question 4
- Targeted training for officials to ensure that they have the capability to be informed and capable buyers of AI services. Developing guardrails or guidelines for implementation and maintenance of AI platforms, including the management of the data assets that are accessed and created
- Creating new roles within the APS focused on ethical uses of AI

8 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.

The categorisation of AI regulation set out in Figure 2 of the Discussion Paper addresses important functional areas and vertical sectors where specific legislation may be required, due to the critical importance of the sector (health and financial services), the safety issues (health and automated vehicles) or their other unique characteristics (defence, government).

We agree with the categories adopted and the breakdown into generic and specific types. The generic regulatory frameworks will still likely need to be adapted to address AI-specific risks occasioned by the scale of infringement that AI may unfortunately make possible. It may still be expedient to explicitly ban some applications that could be addressed within a general category, such as widespread facial recognition, predictive policing and social scoring.

Consideration could be given to adding intellectual property as a category, to address issues that have arisen around the use of data for training.

We also suggest adding another category for critical infrastructure. This could take the form of generic standards applying equally to all such infrastructure, with sector-specific requirements addressed through regulations or guidance.

We recommend considering the adoption of a specific category covering the development of foundational AI technology, noting some of the concerns mentioned in our response to question 2. Whether foundation model development is occurring at sufficient scale (or sufficiently at the leading edge) in Australia to merit regulation should be evaluated to ensure that the regulatory response is proportionate to the risk.

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

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

A high degree of transparency will help mitigate risks and improve public trust in:

- The acquisition, sharing and use of data by governments and the private sector for model training and developing applications. In may be sufficient to rely on existing (and proposed) privacy frameworks to AI. However, consideration should be given to the potential increase in value and interest in some datasets (e.g., public video, public text corpuses) due to their potential value for model training. Datasets may need to be carefully reviewed to mitigate inaccuracies and historical bias and prejudices.
- Using personal data in the provision of services to users. The Attorney-General's Report on the Privacy Act 1988 (Cth) is indicative of the future direction of privacy policy in Australia. The proposed requirement to act fairly and reasonably when collecting, using and disclosing personal information (Proposal 12) would apply readily to AI in respect of both model training and the use of data in the provision of services.[2] We believe Australians will increasingly expect to have control over what personal information is shared. Consideration should be given to the intersection and compliance with other frameworks and obligations, such as the Consumer Data Right (CDR) in the management of personal data, noting possible changes to a Data Holder or Data Receiver as a result of using AI.
- The use of AI specifically in relation to ADM. The Attorney-General's Report (addressing all forms of ADM, and not specifically AI) identifies this as a distinct area of concern and recommends disclosures where ADM has a legal or similarly significant impact on individual rights, along with a right to request information on how decisions have been made (Proposal 19).[3] This will be particularly important as, without a clear line of sight into how decisions have been made, it may only be possible to identify defects in algorithms after the accumulation of enough data to enable statistical analysis.
- The implementation of high-risk technologies, as described further in our response to question 10(a). Public disclosure of such implementations can help to identify key areas of public concern and promote an informed public debate on how technologies are developing.
- The development of foundation AI models. The rapid progress in capabilities may become an area of significant public concern. Although much of the advanced work is undertaken overseas, consideration could be given under a risk-based approach to requiring the submission of applications to undertake domestic model development meeting certain thresholds, in much the same way as applies to research in other regulated industries, such as biotechnology and nuclear energy.

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

Transparency is key to increasing public trust in AI. The drive for transparency should also recognise the need for innovators to protect their intellectual property and the cost of regulatory compliance.

At a minimum, we would suggest that all developers producing AI for public use should be required to:

- Provide basic documentation describing their AI systems, including their purpose, functionality, limitations (e.g., if they are unsuitable or untested for use in safety critical applications) and the types of data they handle, through concepts such as System or Model Cards.
- Disclose their data handling practices, including data collection, storage, use, and privacy protection measures (as they are already required to do under existing legislation, with any AI-related uses specifically noted).
- Conduct and publish an AI Impact Statement that outlines the potential effects of their AI systems, including any associated risks and the steps taken to mitigate them.

[2] *Attorney-General's Department, Privacy Act Review Report 2022*, pages 8 and 9.
[3] *Attorney-General's Department, Privacy Act Review Report 2022*, page 12.

Something analogous to the Australian Bureau of Statistics' Five Safes Framework relating to data protections might assist as a framing device for best practice in applying AI.

Government agencies and systemically important businesses should be subject to additional obligations, including:

- Providing a technical summary of their AI systems, including information about the underlying algorithms, training processes, and measures taken to ensure fairness and avoid bias
- Undergoing regular third-party audits to verify compliance with AI ethics guidelines, data privacy laws, and other relevant regulations
- Implementing frequent monitoring and reporting mechanisms for their AI systems, and sharing these reports with relevant regulatory bodies
- Being required to report promptly on any significant incidents or failures of their AI systems, along with the steps taken to address such issues

Factors helping to determine whether a business is systemically important could include:

- Its market share or monopoly status
- Its involvement in critical sectors or where data is particularly sensitive, such as infrastructure, healthcare, or financial services
- Its social importance, considering the size of its user base and the nature of its services
- Interdependencies with other businesses or sectors that could result in consequential effects
- The potential for harm if its AI systems were to fail or behave unexpectedly

10 Do you have suggestions for:

a. Whether any high-risk AI applications or technologies should be banned completely?

Banning in general should be considered a last resort when other methods of regulation are inadequate. In an Australian social context, the key areas where a ban could be considered comprise:

- Social scoring algorithms
- Predictive policing
- Widespread / indiscriminate facial recognition and other forms of biometric identification (potentially with limited exceptions requiring appropriate judicial oversight)

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

The first area where bans could be considered relate to infringement on existing established rights (e.g., life, health, a fair trial and due process, property, freedom of expression, privacy, freedom from discrimination) and so on. These rights are covered by existing regulatory frameworks, but these should be assessed for their effectiveness in an AI context. Tying the rationale for banning technologies to an infringement of an established right grounds the issue in established law and principle and reduces the scope for giving the government the right to simply ban things it doesn't like.

Some other social ills not previously considered as relating to a specific right may also need to be addressed. For example, whether there should be a "right" to access and participate on the internet and social media. Some privately operated social media and sites have become analogous to a digital town square, but with the ability to algorithmically ban users with little information on how the decision was made or how it could be reviewed.

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

Governments can play a crucial role in shaping perceptions of AI and can implement measures to improve public trust.

It may be common sense, but it is worth stating that AI will only succeed in the public's eyes if it is in fact successful, i.e., that it works and offers tangible advantages over alternative approaches. Especially at this relatively early stage, governments should invest in humans in the loop, system redundancies and other modalities to ensure that AI implementations are given the time and resources to iron out issues with minimal adverse impacts on the public.

Turning first to a government context, it is unlikely anyone wants remote, unresponsive systems making opaque decisions that are justified by cost or productivity savings that are not passed on. At a minimum, the rationale for adopting AI should be clearly explained and, for large-scale implementations, subject to widespread consultation.

Government AI implementations should lean into increasing the transparency of decision-making. Public expectations that AI systems will be able to explain their decision-making are likely to grow, and this should form a standard part of every interaction with the people affected by decisions. In line with mandatory release provisions under information access laws, governments should be required to publish impact assessments and the results of external audits of AI implementations at the system level.

In some cases, it may be possible to structure incentives to encourage AI usage, e.g., a small additional refund on personal tax returns in exchange for a filing that is handled with minimal human intervention, with comprehensive information provided by the system on how decisions have been made and subject always to an entitlement to human review. People can then identify AI in government as a tool that makes their lives easier and better, with an appropriate safety net if the technology gets it wrong and with jobs preserved by freeing up resources to focus on exceptional cases.

For the private sector, in many cases the market may move in the direction of greater interaction with consumers (noting the fall in the costs of doing so) and better explanations of its decision-making for competitive reasons. This dynamic should potentially be given time to develop before considering regulating a requirement for AI transparency for the private sector.

The government should encourage AI literacy in much the same way as it has promoted broader digital literacy. The costs of producing educational material are rapidly decreasing and governments can take advantage of this to embed AI literacy into the curriculum.

The Government can play a role in sponsoring the development of tools and services to assess compliance with AI frameworks and requirements, along the lines of capAI (developed by Oxford University) and AI Verify (developed by the Singapore Government in cooperation with industry, and now open-sourced with the potential to be adapted to Australian requirements).

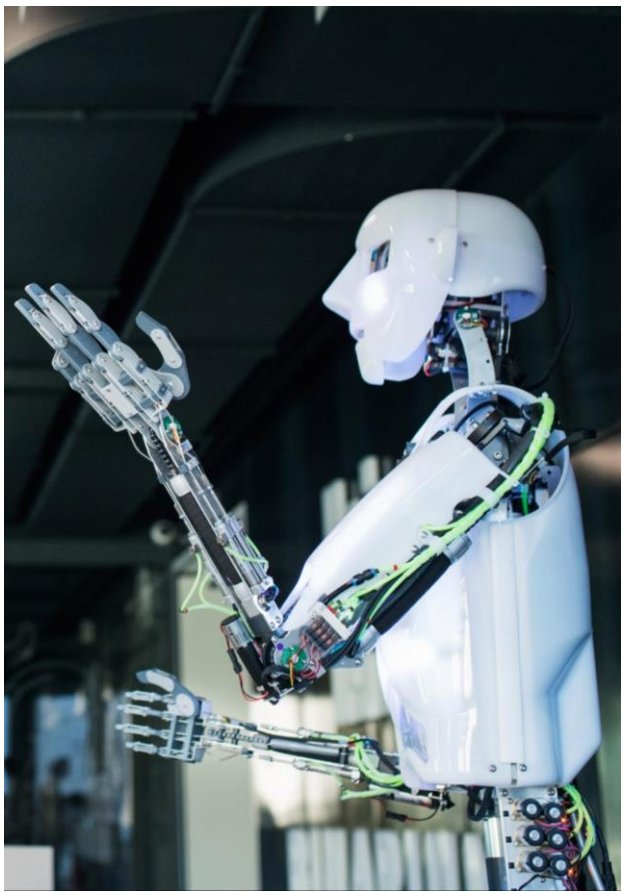
Finally, ensuring that regulatory frameworks are fit for purpose in an AI era will greatly assist public confidence. Breaches of trust should be dealt with comprehensively and with clarity so that the public can be confident that abuses will be appropriately addressed.

Implications and infrastructure

12 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?

In terms of direct impacts, the banning of undesirable technologies in Australia will necessarily reduce the potential trade benefits of supplying services to support those activities, as with the banning of other undesirable activities. This should of course be seen as a price worth paying.

More broadly, technologies are often "dual use", with the ability to be applied to desirable or undesirable purposes. In such cases, the approach taken in other industries could be adopted, involving restricting the sale of technologies for those undesirable purposes, or to markets where such purposes are foreseeable or likely. That should minimise the possible adverse trade impacts of restricting sales. If the uses are not capable of being adequately separated, then consideration may need to be given to restricting sales more broadly.



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

There are many components to a comprehensive system of assurance addressing the potential risks of AI, some of which are in place today. Many of the key components are addressed elsewhere in our submission, including:

- Ethical principles and frameworks for AI, such as the NSW Government AI Ethics Principles and AI Assurance Framework
- A mix of adaptations to existing legislation and new legislation to provide mandatory rules in areas of significant risk
- The ongoing work of Standards Australia, which can help inform the private sector and lay the basis for future regulation
- The role of education, of both the private and government sectors, to ensure that decision-makers have an in-depth understanding of AI's risks and opportunities
- Funding research and fostering collaborations into AI safety and ethics across government, academia and industry

The conformity infrastructure could be further supported by:

- Establishing a central regulatory body to oversee the sector, issue binding market guidance, monitor compliance and undertake necessary enforcement actions. It would need to have the authority, resources, and expertise to effectively manage these responsibilities
- Establishing certification bodies to independently assess and validate AI systems against both regulatory and voluntary standards

Regulatory frameworks will need to be designed with the volume of activity in mind to ensure that the requirement for regulatory resources is proportionate to the risks addressed.

Risk-based approaches

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

EY supports the careful and considered adoption of risk-based regulation for AI. We believe that making the burden of compliance proportionate to the risks is critical to balancing the need to avoid misuse while encouraging important innovation. Its relevance can be seen in its growing role in AI regulation in a number of jurisdictions comparable to Australia.

However, the concept will require substantial elaboration to develop into a fully worked regulatory regime. It will need, for example, clear guidance on how AI risks will be assessed so that developers and investors can make confident choices about where they should invest their resources. We have set out further considerations relating to implementing the risk-based model in our response to question 15.

15 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 will tend to:

- Increase the overall deployability of AI in the economy while protecting citizens from substantial risks.
- Channel innovation into areas where the benefits are likely to outweigh the risks, and away from innovation in higher-risk areas, guiding the private sector to opportunities with greater social value.
- Position regulators to identify and respond to risks over time, as opposed to a more static approach. This will be particularly important given the fast-paced nature of AI development.
- Direct resources to where they will achieve the most benefit in terms of harm reduction, achieving greater benefits at a lower resource cost.

Although the concept of a risk-based approach can result in a matching of risks to enforcement activity, it is not a "cure-all":

- It depends on a strong technical understanding of risk which government may find challenging in such a complex, fast-moving area.
- The principles of risk assessment need to be clearly articulated for both internal regulatory purposes and to inform technology developers and investors.
- The technology is likely to become more deployable, at lower cost and with a lower observable profile, over time. Concerning cases are likely to arise from smaller entities that may not be readily identifiable. If risks cannot be seen and assessed until they reach a threshold of visible concern, significant harms may have already occurred.
- Even a specific AI system is likely to evolve over time due to ongoing developments in the underlying technology, changes made to systems at the implementation level and built-in feedback loops. Governments are likely to face significant challenges in maintaining sufficient visibility to regulate such systems.
- As AI becomes more widely distributed in the economy, it may become less feasible to monitor numerous small-scale implementations, particularly as they may not meet regulatory standards of transparency that may be more suited to major industry players.
- The mobility of data and computing power means that Australian citizens are likely to encounter AI systems that are not based in Australia or subject to its laws or effective enforcement.
- The harms of biased or inaccurate AI decision-making may still be serious without being readily observable, except over time and with a sufficient sample size. Even where differences are detected, mere differences may not be indicative of bias and could require detailed and resource-intensive technical investigation.
- As the market matures it will become more likely that toolchains will fragment, and an AI system may involve components of multiple systems. Attributing responsibility and liability in such circumstances may be challenging.

16 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 a general principle, a risk-based approach is a suitable model regardless of the sector, application, organisation size, AI maturity or resources. In applying a risk-based approach, it may be relevant to consider that:

- Some sectors (e.g., healthcare) are inherently higher risk than others (e.g., book publishing).
- Some applications (e.g., an ADM conducting credit analysis) are inherently higher risk than others (e.g., a retailer's chatbot), noting that the issues may be at a sub-system level—a car's braking system is inherently more safety critical than its climate control.
- Larger organisations (or the government) are systematically more important than smaller organisations, and may be more likely to undertake significant novel or risky innovation.

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

The mechanisms set out in Attachment C of the Discussion Paper outline a balanced approach to monitoring risks and informing consumers of when AI is used. Additional approaches might include:

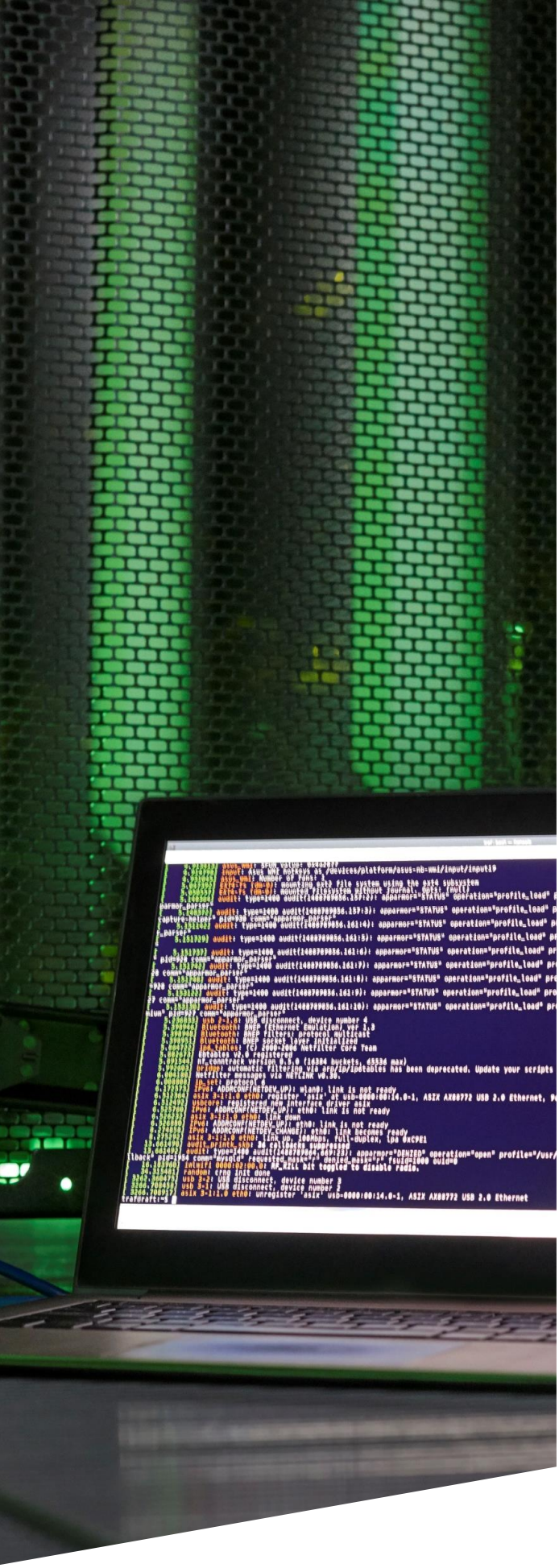
Robustness and security testing—AI systems should be robust and secure to prevent malicious attacks and to ensure they work reliably. Higher-risk AIs should be subject to regular testing for robustness against adversarial attacks and penetration testing to evaluate system security.

Third-party auditing—Independent external audits can help ensure that AI systems are complying with established standards and regulations. Auditing could cover areas such as data handling practices, algorithmic fairness, system security, and privacy protections.

Bias and fairness evaluations—Checks for potential bias and fairness in AI decision-making processes could involve a mix of auditing and the implementation of fairness metrics.

Redress mechanisms—Mechanisms should be implemented to challenge decisions made by AI systems and seek redress for harms. This could involve dispute resolution processes, ombudsman services, or other mechanisms.

- Data governance and privacy frameworks—Existing frameworks should be reviewed to ensure they are fit for purpose in an AI context (noting the Attorney General's recently completed review of the Commonwealth's Privacy Act 1988 (Cth)).
- Safety and resilience planning—Higher risk AI applications could be made subject to additional requirements for demonstrating planning and capabilities addressing how it would mitigate and then recover from a failure and any consequential effects
- Banning—The development, deployment, or usage of certain AI applications could be prohibited where the risks to basic rights or societal norms are deemed too high. A ban could be temporary or permanent depending on the severity of potential harm. The process for implementing a permanent ban should reflect the serious nature of the remedy and include a thorough risk assessment, stakeholder consultation, and be subject to ongoing review. Interim bans could be implemented to allow time for due consideration of possible permanent bans.
- Licensing—Certain high-risk AI technologies could be subject to licensing requirements. Work on these technologies could be limited to parties (organisations or individuals) meeting criteria such as competence and responsibility in managing AI risks. This could also apply to licensing of the deployment of AI in certain critical sectors, such as critical infrastructure, healthcare, or autonomous vehicles.



18 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?

A pre-cursor to integrating an AI risk-based approach into any existing frameworks is a comprehensive review of all assessment frameworks in operation. Similar to that undertaken by NSW Treasury, *Regulating for NSW's Future* (July 2020), any such review should provide the necessary insight into where duplication exists, and the extent to which existing frameworks are fit-for-purpose, and capable of responding to the rapidly changing technological landscape.

To this effect, we anticipate the Government's response to the Attorney General's report on its review of the Privacy Act 1988 (Cth) to be delivered later in 2023.

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

Foundation models (FMs) are truly "foundational" in serving as the basis for a range of AI use cases. Their potential for benefits and harms is widespread and varied. Some key factors for consideration:

- The recent rapid progress in FMs has made it challenging to assess their implications for the wider technology sector, business and society at large.
- Advances in the development of FMs and their compression and distribution (after they are developed) are making FM technology progressively more widely available with a smaller visible footprint. It is already possible to install and use an LLM while remaining offline on a consumer PC or laptop, and use it to generate information that may be usable for criminal purposes. This problem is likely to grow. As an example of possible future directions, OpenAI's Code Interpreter plugin gives ChatGPT access to programming resources and the ability to solve multi-step problems involving complex data analysis. A future offline version of such a tool could potentially make current advanced capabilities in computer hacking trivially available to any person.

- Training costs for FMs are currently high due to the large amount of computation required. Consideration should be made between the environmental impact of many FMs being individually developed and trained, versus the challenges highlighted in our response to question 2 of the concentration of knowledge in a small group of suppliers.
- Although some scenarios of the possible outcomes of continued FM model development may be overblown, there are "unknown unknowns" in the form of a possibility of future widespread adverse consequences, either directly as a result of technological developments or from the use of advanced FMs by hostile actors. The possible scale of such an event justifies more than usual caution in regulating the enabling technology.
- As FMs may be applied to a mix of high- and low-risk use cases, assessing the risk level of potential use cases does not usefully inform the regulatory response for FMs.
- FMs are not only potentially risky in their own right, but are potentially systemically important when used as the basis for AI applications.

Concerns about the rapid development of AI recently led over 1,000 industry experts, and others, to propose a temporary moratorium on the further development of FMs to allow time for risks to be assessed.[4]

In terms of a regulatory response for Australia, we would recommend that:

- Any fundamental or novel research in FMs undertaken in Australia should be subject to reporting obligations and oversight from technically qualified independent experts, with a view to implementing further restrictions as appropriate.
- Unusually, oversight should take a more precautionary approach to risks, focused on safety and the prevention of adverse outcomes, than might normally be applied in other regulatory domains.
- Because of the difficulties of assessing FMs for their specific technical applications, oversight should focus on how FMs will be built and trained. Specific capability gates could be identified and subject to further review (such as access to the internet, limits on stored context, administrator access to software systems, recursive self-improvement) although some of these gates have been passed already and others may be difficult to observe or enforce.

The oversight should be balanced, to some extent, by the awareness that other countries are likely to proceed apace in this area, and that innovation should be encouraged to the greatest extent possible that is consistent with safety.

20 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?

Yes, a risk-based approach should be mandated across public and private sectors (noting our responses to questions 6 and 19).

EY has written previously on the need to establish a central regulatory agency with authority to introduce binding market regulation for AI and ADM.[5] This approach can still allow for self-regulation where appropriate (as the regulator may choose not to issue regulations in some areas) in order to continue to foster innovation and ensure that regulation is targeted to significant public policy goals. But it will still allow for a rapid, binding regulatory response as and when a policy need arises.

[4] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[5] https://www.ey.com/en_au/government-public-sector/building-a-trusted-ai-framework-for-the-public-sector

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

EYG no. 000371-24Gbl

ey.com