Australian Government
Department of Industry, Science and Resources

Safe and responsible AI in Australia consultation

Australian Government's interim response

Contents

Overview

The potential for AI systems and applications to help improve wellbeing, quality of life and grow our economy is well known. It's been estimated that adopting AI and automation could add an additional $170 billion to $600 billion a year to Australia's GDP by 2030.[1]

While AI is forecast to grow our economy, there is low public trust that AI systems are being designed, developed, deployed and used safely and responsibly. This acts as a handbrake on business adoption, and public acceptance. Surveys have shown that only one-third of Australians agree Australia has adequate guardrails to make the design, development and deployment of AI safe.[2]

Submissions to the *Safe and responsible AI in Australia* discussion paper (the discussion paper) said that more needs to be done to ensure that the development and deployment of AI is safe and responsible. In considering the right approach, it is important to recognise there are a range of ways AI technologies are used. Many AI tools and applications – for example, to filter spam emails or to optimise business operations – are not new, and are low risk. Other AI applications – for example, to predict a person's likelihood of recidivism, suitability for a job, or in enabling a self-driving vehicle – are considered higher risk. These uses can result in negative impacts for people that are difficult, or impossible, to reverse. It was for this reason that the discussion paper proposed a risk-based approach, focused on setting up additional guardrails to reduce the likelihood of harms occurring in high-risk settings.

Submissions highlighted specific risks associated with newer, very powerful AI models. These 'frontier' models exceed the capacity of previous models and can generate new content quickly and easily. Submissions noted that 'frontier' AI models may require targeted attention. These models, developed by a small number of companies, can be embedded in a wide variety of settings. It was also highlighted that AI services are being developed and deployed at a speed and scale that could outpace the capacity of legislative frameworks, many of which have been designed to be technology-neutral. This speed and scale is also driving concern that these systems, deployed for legitimate purposes that nonetheless can result in harm, should be subject to greater testing, transparency and oversight. Responses also indicated public concern that AI systems are not being tested appropriately, and have limited detail on how they function. As a result this is reducing public and business trust and reliability in the outcomes of these systems.

A preliminary analysis of submissions found at least 10 legislative frameworks that may require amendments to respond to applications of AI. Many AI risks outlined in submissions were well known before recent advances in generative AI. These include:

- inaccuracies in model inputs and outputs
- biased or poor-quality model training data
- model slippage over time
- discriminatory or biased outputs
- a lack of transparency about how and when AI systems are being used.

The pace at which advances in generative AI are being made accessible by companies, without perceived oversight, has sharpened the focus of governments worldwide on ensuring there are sufficient guardrails for AI development and deployment. Increasingly, governments are demanding that companies developing and deploying AI in high-risk contexts take proactive steps to make their products safe to use.

1 Taylor et al., 'Australia's automation opportunity: Reigniting productivity and inclusive income growth', McKinsey & Company, 3 March 2019, accessed 12 December 2023.

2 Gillespie et al., 'Trust in Artificial Intelligence: A Global Study', The University of Queensland and KPMG Australia, 2023, accessed 12 December 2023, doi:10.14264/00d3c94.

This includes introducing obligations for:

- testing (for example, testing of products before and after release)
- transparency (for example, labelling of AI systems in use or watermarking of AI-generated content)
- accountability (for example, requiring training for developers and deployers and clearer obligations to make organisations accountable and liable for AI safety risks).

Some jurisdictions have introduced voluntary commitments from companies. The United States (US) sought voluntary commitments from seven leading AI companies. Singapore introduced standardised self-testing tools ('AI Verify') to enable businesses to check AI models against a set of principles. Other jurisdictions, including Canada and the European Union (EU), are seeking to make commitments mandatory for higher risk AI systems through new legal frameworks. Both Canada and the EU have also sought voluntary commitments from companies ahead of the enactment and enforcement of these proposed legal frameworks. In December 2023, for example, the EU announced the establishment of an AI Pact to foster early implementation of obligations that will be mandated under the EU AI Act.

The pace of advancements in AI was also a catalyst for the recent AI Safety Summit (the summit) hosted by the United Kingdom (UK) in November 2023. At the summit, Australia joined the EU and 27 countries in signing the Bletchley Declaration, committing to international collaboration on AI safety testing and the building of risk-based frameworks across countries to ensure AI safety and transparency. On the eve of the summit, the US published an extensive executive order on AI safety. It includes new AI safety and transparency standards and mandated testing and notification requirements for companies developing 'foundation' models through their domestic *Defense Production Act*. The Bletchley Declaration with the US executive order represent a seismic shift in the way that governments approach technology regulation. Many responses to the discussion paper also reflected this shift.

There was broad consensus in submissions that voluntary guardrails are insufficient. Submissions proposed that if additional mandatory guardrails are introduced they should only apply to high-risk applications of AI. This would allow low-risk innovation to flourish largely unimpeded. Views differed as to the most appropriate regulatory response from government to embed safety and safety guardrails for high-risk AI. Existing laws could be updated (for example, through healthcare sector or financial sector laws) or new laws could be introduced for AI (for example, modelled on the EU or Canadian approach). Submissions emphasised that any regulatory response should be interoperable with international approaches and adapted as necessary.

Submissions also highlighted that non-regulatory approaches will also be needed to build public trust and boost adoption. This included in areas such as education, business awareness and capability and capacity building.

The Australian Government's interim response

The government recognises that many applications of AI do not present risks that require a regulatory response. For example, AI can help monitor and measure biodiversity or help automate internal business processes.

However, we have heard from consultations that the current regulatory framework likely does not sufficiently address known risks presented by AI systems, which enable actions and decisions to be taken at a speed and scale that hasn't previously been possible. In particular, existing laws likely do not adequately prevent AI-facilitated harms before they occur, and more work is needed to ensure there is an adequate response to harms after they occur.

The government is already undertaking work to strengthen existing laws in areas that will help to address known harms with AI. This includes the implementation of privacy law reforms, a review of the *Online Safety Act 2021*, and introduction of new laws relating to misinformation and disinformation. The government will continue to work with states and territories to consider opportunities to further strengthen regulatory frameworks.

However, the submissions identified gaps where it was assessed laws do not sufficiently prevent harms from the deployment of AI systems in legitimate but high-risk contexts. When AI is used in high-risk contexts, harms can be difficult or impossible to reverse such that specific guardrails for AI design, development, deployment and use may be needed.

The government will consider mandatory safeguards for those who develop or deploy AI systems in legitimate, high-risk settings. This will help ensure AI systems are safe when harms are difficult or impossible to reverse.

The government will consider possible legislative vehicles for introducing mandatory safety guardrails for AI in high-risk settings in close consultation with industry and the community. It is committed to a collaborative and transparent approach to developing obligations, whether that is through amendments to existing laws or through a new dedicated legislative framework.

The government also recognises the need to consider specific obligations for the development, deployment and use of frontier or general-purpose models. It will continue to collaborate with international partners to establish safety mechanisms and testing of these systems during the AI product lifecycle, noting that models developed overseas can be built into applications in Australia.

In considering the right regulatory approach to implementing safety guardrails, the government's underlying aim will be to help ensure that the development and deployment of AI systems in Australia in legitimate, but high-risk settings, is safe and can be relied upon, while ensuring the use of AI in low-risk settings can continue to flourish largely unimpeded. Our immediate focus will be on considering what mandatory safety safeguards are appropriate, and how best to implement them, informed by developments in other countries. In addition, and building on the investments in Australia's AI capability in the 2023–24 Budget, the government will continue to explore steps it can take to support the development and diffusion of AI technologies across the Australian economy, including the need for an AI Investment Plan.

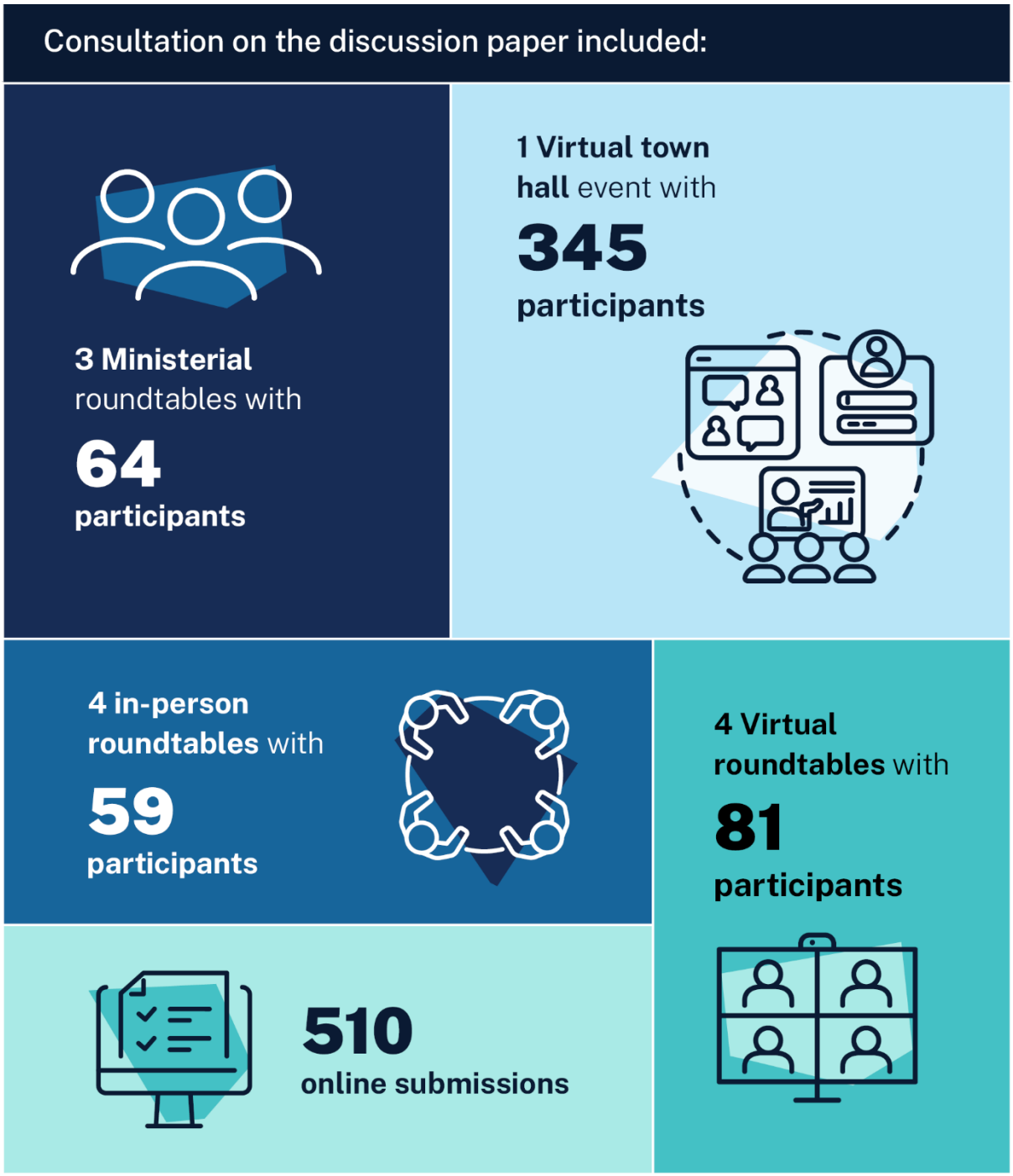
While the government considers mandatory guardrails for AI development and use and next steps, it is also taking immediate action through:

- working with industry to develop a voluntary AI Safety Standard, implementing risk-based guardrails for industry
- working with industry to develop options for voluntary labelling and watermarking of AI-generated materials
- establishing an expert advisory body to support the development of options for further AI guardrails.

What we heard

Where and how did we engage

Between 1 June 2023 and 4 August 2023, the Australian Government consulted on its *Safe and responsible AI in Australia* discussion paper (the discussion paper). The discussion paper sought views on whether Australia has the right governance arrangements in place to support the safe and responsible use and development of AI.



We heard from:

- members of the public
- community, digital rights and advocacy groups
- academia and research institutions
- industry and peak bodies
- legal firms
- health and care organisations
- government agencies.

The Department of Industry, Science and Resource has published all non-confidential submissions on its consultation hub.

The discussion paper has been effective in focusing this public debate. The largest proportion of submissions, about 20%, were from people writing in a personal capacity. Members of the public increasingly expect government to take a substantive response to support the safe and responsible use of AI.

What did we hear?

Australians are excited about the potential for AI to benefit our society and economy. However low trust, a lack of skills, inadequate IT infrastructure, financial barriers and regulatory uncertainty are barriers to AI adoption in Australia and around the world.[3] Australians worry there is not enough being done to mitigate risks around the use of AI. There were broad calls for government to do more to harness the opportunity and address these risks, though differing views on the action government should take.

Australians see the opportunity of AI

Submissions highlighted that AI is already benefiting our society and economy. AI systems are helping to analyse medical images, optimise engineering designs and better forecast and manage natural emergencies.[4] Submissions also identified that AI has the potential to bring transformative benefits to the way Australians live and work. AI can create new jobs and benefit consumers. It can change the way we learn, power new industries, boost productivity, uplift healthcare and facilitate a smooth transformation to net zero.

3 R Grünbichler, 'Implementation barriers of artificial intelligence in companies' [conference paper], FEB Zagreb International Odyssey Conference on Economics and Business, Zagreb, 22-25 May 2023, accessed 12 December 2023, doi:10.22598/odyssey/2023.5; S Alsheiabni, Y Cheung and C Messom, 'Factors Inhibiting the Adoption of Artificial Intelligence at organization-level: A Preliminary Investigation' [conference paper], Twenty-fifth Americas Conference on Information Systems, Cancun, 15-17 August 2019, accessed 12 December 2023; M Hoffman and L Nurski, 'What is holding back artificial intelligence adoption in Europe?', Policy Contribution, Bruegel, 30 November 2021, accessed 12 December 2023.

4 Bell et al., 'Rapid Response Information Report: Generative AI – language models (LLMs) and multimodal foundation models (MFMs)', Australian Council of Learned Academies, 24 March 2023, p 10; W Sun, P Bocchini and BD Davison, 'Applications of artificial intelligence for disaster management', Natural Hazards, 3 July 2020, accessed 12 December 2023, 103:2631–2689, doi:10.1007/s11069-020-04124-3.

Submissions highlighted that AI systems are already powering and improving the lives and wellbeing of Australians.

> "AI's potential as a powerful transformative force that delivers good in the world must be thoughtfully explored and supported. For example, AI can reduce the instance of human error, can take on high-risk activities such as deep sea or space exploration, and can take on mundane and monotonous tasks without experiencing fatigue, which in a high skills-based country like Australia, can enhance national productivity and enable employers to diversify and broaden roles." (anonymous)

AI has the potential to improve educational outcomes for children. It can offer personalised learning experiences which address their unique needs or help them to collaborate and develop critical thinking and problem-solving skills.

> "AI systems show promise in improving educational opportunities, from early learning to virtual mentoring to school management." (UNICEF)

> "Adaptive learning platforms have the potential to provide personalized learning experiences to address each user's unique needs. When combined with traditional teaching methods, such customization and one-on-one intelligent tutoring could be greatly beneficial to children with learning difficulties. Other types of AI-enabled educational tools can help teachers generate curricula without having to develop them from scratch" (UNICEF).

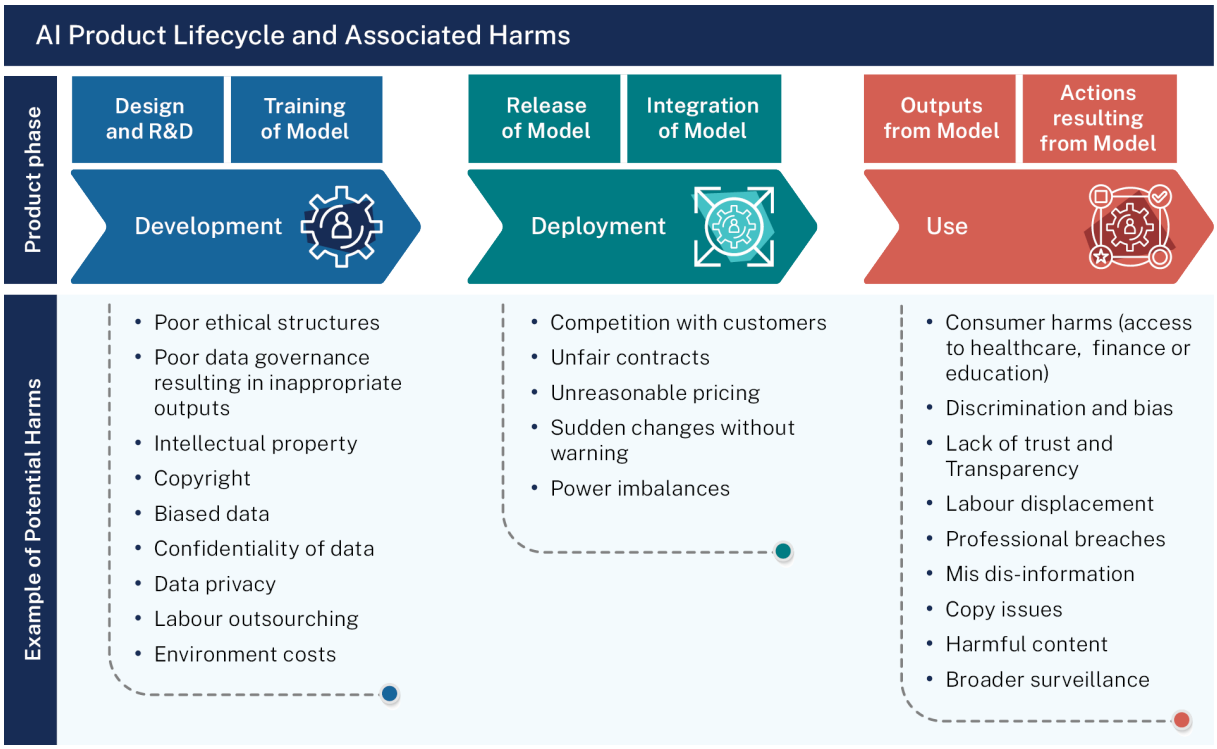
Australians have concerns about AI

However, alongside excitement for the opportunities of AI, there is community concern about the potential for harms along the AI design, development and deployment lifecycle.

Conceptualising the harms of AI

The below diagram draws on the submission from the ARC Centre of Excellence for Automated Decision-Making and Society (ARC ADM+S). It provides a visual representation of the AI development lifecycle and identifies the harms that may occur at each stage.

Diagram of impacts through AI lifecycle:



As outlined in the diagram, developing an AI system begins with the gathering, cleaning and processing of data. Harms from this initial phase stem from the data collection processes used and are widely documented. For example, data sets could use intellectual property without approval from the owner in a way that breaches relevant intellectual property laws and undermines the legitimate commercial interests of rightsholders. Any biases or errors in the data can become embedded in the model and be repeatedly expressed in its outputs. If the data quality is compromised harms can occur regardless of the sophistication of the AI model itself.

The next stage is training of the model by feeding data into a training system. This is followed by a refinement process to ensure the model achieves its intended goals. The potential harms in this stage include the environmental impacts of processing data and the design of systems that do not adequately mitigate the risks of harmful downstream application. As with the data collection phase, biases and errors embedded in training may flow through to the outputs of the model.

The model can then be made available to users through an app interface or integrated into existing services. At this stage, potential competition issues become relevant to consumers. The dominance of one or few products may create dysfunctional markets that result in unreasonable pricing, unfair contracts and unequal access.

When the model delivers outputs, potential harms multiply. Harms can be individualised and include discrimination, deception or malice. But harms may also manifest at a systemic level, with outputs potentially compromising political and social cohesion, stability of labour markets and human rights. Influenced by the processes of data collection and training, submissions most commonly identified the risk of catastrophic and unforeseen harms emerging at this stage of output compared to earlier stages.

Risks identified by submissions

Many of the risks of AI are not new. There have been examples over the years of AI systems causing harms where known risks, like bias, discrimination and scams, have not been adequately mitigated. But submissions also spoke of new and emerging risks. Existing regulation addresses some of these new and emerging risks but may not capture others. The speed and scale of the development and deployment of AI is driving concern that AI systems, deployed for legitimate purposes that nonetheless can result in harm, should be subject to greater testing, transparency and oversight. Responses also indicated public concern that AI systems are not being tested appropriately, and have limited detail on how they function.

Submissions presented diverse views on the most significant, urgent and probable risks of AI. We have broadly categorised these as technical risks, unpredictability and opacity, contextual risks, systemic risks and unforeseen risks.

Technical risks

Submissions raised well-documented concerns that the outputs of AI systems can be compromised by technical limitations, including inaccuracies in system design or biases in training data. This can result in inaccurate or unfair outcomes for people or groups. For example, if AI models in healthcare are trained on non-representative data, they may contribute to disparity in health outcomes for underrepresented groups.

Unpredictability and opacity

Submissions identified that opaque AI systems can make it difficult to identify harms, predict sources of error, establish accountability, explain model outcomes and assure quality. For example, if job applications are assessed by 'black box' AI systems (where internal workings are automated and invisible), people affected by discriminatory outcomes may have limited ability to understand or question decisions.

Domain-specific risks

New or exacerbated risks can arise where AI interacts with existing harms, systems or legislative frameworks. Examples include the generation and spread of online harms like deepfake pornography or AI generated cyber-abuse, and the undermining of social cohesion through misinformation or disinformation generated, tailored and spread by AI.

Systemic risks

Many submissions also emphasised that potentially systemic risks may arise from emerging AI developments. Submissions identified risks associated with the development of highly capable and potentially dangerous frontier models as well as the greater accessibility and useability of generative AI models. Submissions noted that these developments can lead to unpredictable harms on a scale and at a speed not previously possible.

Unforeseen risks

Submissions acknowledged that AI is evolving with a speed and complexity that will likely pose unforeseen risks. The rate of technological change makes it difficult to ensure that regulation is future-proofed and can meet unforeseen challenges without stifling innovation. Submissions called for regulatory approaches that are flexible and responsive to risks as they emerge. Governments must respond with agility when known risks change and new risks emerge.

Australians want more to be done

Almost all submissions called for the government to act on preventing, mitigating and responding to the harms of AI. However, views were mixed on how this should be done. Where harms are likely, significant or difficult to reverse, submissions broadly support the establishment of regulatory safeguards to prevent harms from occurring in the first place. Multiple pathways were proposed to establish appropriate regulatory safeguards, including improving existing laws and introducing new laws.

While the need for regulatory action to prevent and respond to harm was a key message, there was also a broad acknowledgement that low-risk AI applications (for example, the optimisation of business operations) should not be subject to onerous guardrails.

Further, there was an acknowledgement that non-regulatory initiatives, including standards development, assurance frameworks and capability building, must play an important role in supporting safe and responsible AI in Australia.

Submissions called for regulatory action

Submissions broadly agreed that voluntary commitments from companies to improve the safety of systems capable of causing harm are insufficient. Views differed, however, on the most appropriate form of regulation to make these commitments mandatory. Industry groups preferred an approach that focused on strengthening existing laws, through amendments or providing regulatory guidance. On the other hand, consumers and academic groups were more likely to call for new laws or a specific AI Act like those being pursued in the EU, Canada and Korea. They noted any new laws would need to be tailored to Australia's context. Submissions identified the importance of the Australian Government working with the states and territories to harmonise regulatory settings on AI development.

Strengthening existing laws

Submissions noted existing laws and regulation may be suitable to regulate some applications of AI or could be adapted without introducing entirely new ex-ante approaches.

Ex-ante laws

Many submissions suggested establishing ex-ante regulation, particularly for the deployment of AI systems in legitimate, but high-risk settings, and for 'frontier' AI models where harms could be difficult or impossible to reverse. Ex-ante regulation is a regulatory intervention to limit harms before they occur. These preventative interventions would apply early in the AI lifecycle, particularly in the design of systems.

Submissions proposed a range of preventative interventions broadly categorised as testing, transparency and accountability:

- testing, including internal and external testing before the release of an AI system and ongoing auditing requirements
- transparency, including requiring generative AI systems to incorporate digital labelling or 'watermarks', so that AI-generated content is identifiable
- accountability, including mandating 'human-in-the-loop' requirements where critical points of AI decision-making have human oversight to mitigate AI misalignment with human objectives, or introducing licensing schemes for high-risk AI development.
- introducing outright bans of AI uses that present unacceptable risks – suggestions in submissions included behavioural manipulation, social scoring and real-time widescale facial recognition.

A number of submissions considered that the lack of ex-ante rules in current technological development allows for the development and proliferation of high-risk applications of AI. Where there is limited ex-ante regulation to ensure AI models are safe, laws that address harms after they've already occurred can be over-relied on. Submissions identified that the speed and scale of recent advances in AI is changing the risk profile and creating challenges for existing laws that will increasingly struggle to contain harms.

Risk-based approach

The discussion paper outlined a risk-based management approach to regulate AI. In a risk-based regulatory framework, AI development and application is subject to regulatory requirements commensurate to the level of risk they pose. A risk-based approach allows low-risk AI development and application to operate freely while targeting regulatory requirements for AI development and application with a higher risk of harm.

Among submissions, there was consensus that a risk-based approach to adopt AI guardrails is appropriate. However, there were mixed views on implementation, including whether assessment should be voluntary, internal or external. Many submissions cited examples of AI applications that they believed should not be captured by additional guardrails, like screening parcels and optimising internal business functions. A risk-based approach to AI regulation would reflect this by ensuring low-risk applications could be implemented with little or no regulatory intervention.

Submissions identified benefits of a risk-based approach, including the potential ability to:

- give regulatory certainty by categorising risks and obligations
- minimise compliance costs for businesses that do not develop or use high-risk AI
- incorporate a well-defined 'menu' of risk-management options that could be imposed
- balance the costs of regulatory burden with the value of risk reduction
- be flexible and responsive as technology develops.

Submissions also identified limitations to a risk-based approach, including that:

- frameworks will not accurately and reliably predict and quantify risks
- context-specific risks will not be well captured by categorisation
- unpredictable risks will not be considered, particularly for frontier models designed for general purpose application

- assessment, if voluntary or carried out via self-assessment, will underestimate risk
- the categorisation of risk is reductive and ineffective
- there is no appropriate legislative foundation or regulator to administer a risk-based framework
- there are diverse views on what defines 'high-risk' AI.

Further work is needed to define the criteria of risk categorisation to ensure that safe AI systems would not be overregulated and that dangerous AI systems would not be underregulated. Likewise, further work is needed to identify what would be the most appropriate guardrails and regulatory interventions for each category of risk. It will also be important to further refine definitions of 'high-risk' AI, including taking into account developments in overseas jurisdictions.

Existing definitions of 'high-risk'

The discussion paper sought stakeholder feedback on a risk-based approach for addressing potential AI risks. In doing so, it proposed a concept of 'high-risk' that was focused on impacts that were 'systemic, irreversible or perpetual'. The use of AI-enabled robots for medical surgery and the use of AI in self-driving cars to make real-time decisions were presented as 2 examples.

The EU has adopted a list-based approach in its proposed *Artificial Intelligence Act*, citing specific uses of AI that it considers to be high-risk in terms of safety and rights implications of the individuals impacted by the use of that AI. This includes:

- certain critical infrastructure (water, gas, electricity)
- medical devices
- systems determining access to educational institutions or recruiting people
- systems used in law enforcement, border control and administration of justice
- biometric identification
- emotion recognition.

Canada's proposed AI legislation allows for a definition of a 'high-impact system' to be prescribed by regulation. In the interim, the legislation's companion documents set out a range of principles to determine whether an AI system is high-impact, including the potential severity of the harm to an individual's safety or rights.

Updating existing laws

Businesses and individuals who develop and use AI are already subject to various Australian laws. These include laws such as those relating to privacy, online safety, corporations, intellectual property and anti-discrimination, which apply to all sectors of the economy. There are also sector-specific laws that impact some development and deployment of AI, for example in medical devices, motor vehicles, airline safety and financial services.

Submissions identified at least 10 existing legislative frameworks that may need updating or clarification to be relevant in an AI context. These include:

- uncertainties relating to whether individuals who use AI to generate deepfakes can be liable for misleading and deceptive conduct (competition and consumer law)
- whether AI models used by health and care organisations and practitioners could lead to clinical safety risks (health and privacy laws)
- the way creative content may be used to train generative AI models, including potential remedies for any infringement (copyright law).

Some submissions went further to propose specific amendments to legislation across these and other legal domains.

Targeted but technology-neutral approach

Given the rapid pace of AI technological development, some submissions called for introducing regulatory settings that are agile and responsive to emerging risks. If regulatory actions are too rigid in their application to the current state of AI technology, there is a risk that they will not apply as intended when AI technology advances in unpredicted ways. Some called for regulatory actions to be 'technology neutral' or 'outcomes focused' so they keep pace with technological advancement. Likewise, to ensure there aren't gaps in the regulatory environment, many submissions called for a cohesive 'whole- of- government' approach to ensure actions are not fragmented.

Security

While security issues were not in the scope of the discussion paper, many submissions emphasised the importance of embedding national and cyber security measures in AI development and deployment. AI models can exacerbate cyber risks, amplify risks such as disinformation and increase the identification and exploitation of cyber vulnerabilities at scale. But AI technologies can also play an important role in bolstering Australia's cyber resilience. A number of submissions called for AI development to incorporate secure-by-design practices to ensure security is a core business goal from the start of development. Security was identified as a necessary foundation to build community and business trust in AI.

This is a rapidly evolving area and Australia will continue to work with global partners. Building on the Bletchley Declaration at the first Global AI Safety Summit in November 2023, Australia can contribute to international governance mechanisms and regulation to ensure AI models are secure by design. This will help protect against the risk of malicious corruption of models and potential malicious use of AI for cybercrime and other digital threats.

The 2023–2030 Australian Cyber Security Strategy will help the government achieve its vision to be a world-leading cyber secure and resilient nation by 2030. The strategy focuses on sovereign capability and takes a whole-of-nation approach to building cyber resilience. Security considerations will continue to inform Australia's response to the opportunities and risks of AI.

Submissions call for non-regulatory action

Submissions proposed a diverse range of non-regulatory actions government can take to implement safe and responsible AI. These include:

- establishing an expert AI advisory body
- considering regulatory sandboxes
- investing in domestic AI capability
- adopting and promoting international standards for AI development and deployment
- for the government to lead by example in its own safe and responsible use of AI.

AI advisory body

Many submissions recommended the establishment of an AI advisory body. A body could provide regulators and policymakers with access to ongoing, multi-disciplinary, technical and regulatory expertise on AI. If formed, a body could have a diverse membership and expertise from across industry, academia, civil society and the legal profession.

Regulatory sandboxes

A number of submissions call for the introduction of 'regulatory sandboxes' to support domestic AI innovation. A regulatory sandbox is a controlled environment where the government provides approval for industry participants to test innovative concepts in the market under relaxed regulatory requirements at a smaller scale, on a time-limited basis, and with appropriate oversight and safeguards in place. Submissions suggested an Australian AI regulatory sandbox could allow the government to work with industry to approve controlled trials of new AI technologies and applications under relaxed regulatory settings with safeguards in place to ensure safe practice and outcomes. A regulatory sandbox could benefit industry by providing the conditions to innovate and grow domestic capability. It could benefit regulators by strengthening collaboration with industry and providing an opportunity to evaluate regulatory settings.

Government as an exemplar

Submissions emphasised that government should be an exemplar in its safe and responsible adoption and use of AI technologies. This includes harnessing AI and machine learning to predict service needs, gain efficiencies in agency operations, support evidence-based decisions and improve user experience.

Some submissions recommended government acts to mitigate the risks of frontier models developed overseas. This includes engaging in international initiatives to coordinate AI governance, contributing to standards development and promotion, and ensuring that our domestic responses can 'scale up' to meet global risk.

Many submissions identified that voluntary mechanisms are important but insufficient by themselves to harness the opportunity of safe and responsible AI.

AI capability

Submissions called for more investment in Australia's AI capability and identified that Australia is in the early stages of adopting AI. The extent to which Australia will benefit from AI will depend on our ability to develop and adopt AI solutions, as well as our willingness to trust their application.

> "We support the development of practical guidance and educational initiatives surrounding the use of AI to educate the public and generate trust and confidence in AI. We think that this would not only increase AI adoption, but would also be conducive towards encouraging the legal and responsible development and deployment of AI." (Insurtech Australia)

"The Australian Government should play an active role in fostering collaboration and knowledge sharing among various stakeholders, including researchers, industry leaders, and civil society organisations. Establishing partnerships and platforms for dialogue will facilitate the exchange of best practices, promote responsible AI innovation, and enable continuous learning and improvement in AI governance." (Professor Rocky Scopelliti)

The government's interim response

Through the discussion paper, the government started a conversation with the Australian community on how to best harness the opportunities and mitigate the risks of AI. The government is committed to continuing this conversation.

The initial analysis of the submissions, work across government, and global conversations at forums like the AI Safety Summit, has made clear:

- AI can create new jobs, power new industries, boost productivity and benefit consumers. Highlighting the benefits presented by AI will boost community confidence
- many applications of AI do not present risks that require a regulatory response, and there is a need to ensure the use of low-risk AI is largely unimpeded
- our current regulatory framework does not sufficiently address risks presented by AI, particularly the high-risk applications of AI in legitimate settings, and frontier models
- existing laws do not adequately prevent AI-facilitated harms before they occur, and more work is needed to ensure there is an adequate response to harms after they occur
- the speed and scale that defines AI systems uniquely exacerbates harms, and in some instances makes them irreversible, such that an AI-specific response may be needed
- consideration needs to be given to introducing mandatory obligations on those who develop or use AI systems that present a high risk, to ensure their AI systems are safe
- the need for government to work closely with international partners to establish safety mechanisms and testing of these systems, noting that models developed overseas can be built into applications in Australia.

In considering the right regulatory approach, the government's underlying aim will be to ensure the development and deployment of AI systems in Australia in legitimate, but high-risk settings, is safe and can be relied upon, while ensuring the use of AI in low-risk settings can continue to flourish largely unimpeded. With this in mind, the government's immediate focus will be on considering whether mandatory safeguards are appropriate. If they are it will consider how to best implement them. This may be through existing laws or new approaches. This work will be undertaken in close consultation with industry, academia and the community.

Principles guiding the government's interim response

Industry continues to invest in the development and rollout of AI. AI technologies will also continue to evolve at pace. With this in mind, the following principles will help guide the government's approach to AI. They reflect the government's commitment to creating a regulatory environment that builds community trust and promotes innovation and adoption while balancing critical social and economic policy goals. They seek to ensure the development and deployment of AI systems in high-risk settings is safe and reliable.

Principles guiding the Australian Government's interim response to support safe and responsible AI

Risk-based approach

The Australian Government will use a risk-based framework to support the safe use of AI and prevent harms occurring from AI. This includes considering obligations on developers and deployers of AI based on the level of risk posed by the use, deployment or development of AI.

Balanced and proportionate

The Australian Government will avoid unnecessary or disproportionate burdens for businesses, the community and regulators. It will balance the need for innovation and competition with the need to protect community interests including privacy, security and public and online safety.

Collaborative and transparent

The Australian Government will be open in its engagement and work with experts from across Australia in developing its approach to the safe and responsible use of AI. It will ensure there are opportunities for public involvement and draw on technical expertise. Government actions will be clear and make it easy for those developing, implementing or using AI to know their rights and protections.

A trusted international partner

Australia will be consistent with the Bletchley Declaration and leverage its strong foundations and domestic capabilities to support global action to address AI risks. This includes substantial risks to humanity from frontier AI, addressing the high-risk applications of AI, as well as near-term risks to individuals, our institutions and our most vulnerable populations.

Community first

The Australian Government will place people and communities at the centre when developing and implementing its regulatory approaches. This means helping to ensure AI is designed, developed and deployed to consider the needs, abilities and social context of all people.

Next steps

In line with the Australian Government's overall objective to maximise the opportunities that AI presents for our economy and society, proposed next steps relate to:

- preventing harms from occurring through testing, transparency and accountability
- clarifying and strengthening laws to safeguard citizens
- working internationally to support the safe development and deployment of AI
- maximising the benefits of AI.

Preventing harms from occurring through testing, transparency and accountability

In response to calls for further guardrails to prevent harms from arising, particularly where the deployment of AI presents a high risk, the government will consult further on options for introducing new regulatory guardrails with a focus on testing, transparency and accountability.

Testing could include requirements relating to:

- internal and external testing of AI systems before and after release, including, for example, by independent experts
- sharing information on best practices for safety
- ongoing auditing and performance monitoring of AI systems
- cyber security and reporting of security-related vulnerabilities in AI systems.

Transparency could include:

- users knowing when an AI system is used and/or that content is AI generated, including labelling or watermarking (see below)
- public reporting on AI system limitations, capabilities, and areas of appropriate and inappropriate use
- public reporting on data a model is trained on and sharing information on data processing and testing.

Accountability could include:

- having designated roles with responsibility for AI safety
- requiring training for developers and deployers of AI products in certain settings.

This work includes defining 'high risk' in an Australian context. It also includes work to clearly communicate when AI is legitimately deployed with high risks to individual safety, individual rights and national security.

It will also consider links to other initiatives across the Australian Government as well as state and territory governments. These include:

- the **AI in Government Taskforce's** work to support the safe and responsible deployment of AI in the Australian Public Service, including by developing policies, standards and guidance
- related work under the **Data and Digital Minister's Meeting** to develop a nationally consistent approach to the safe and ethical use of AI by governments
- reforms to **Australia's privacy laws**, including an in-principle agreement to require non-government entities to conduct a privacy impact assessment for activities with high privacy risks to identify and manage, minimise or eliminate risks (which is already a requirement for government entities). Further, corresponding proposals agreed in the **Robodebt Royal Commission** report focus on increasing transparency and integrity of automated decision- making which uses personal information
- the registration under Australia's **online safety laws** of new mandatory industry codes and the development of 2 mandatory industry standards which will require industry to provide appropriate community safeguards to deal with certain types of illegal and harmful content (including child sexual abuse material) online, including that generated and spread by AI
- **cyber security considerations** consistent with the Cyber Security Strategy, as well as work underway in the Australian Signals Directorate through its Ethical AI Framework.

To complement longer-term consideration of regulatory options to mandate testing, transparency and accountability obligations in high-risk AI settings, the Australian Government will take immediate steps to

help businesses to operationalise safe and responsible AI. This will be through 3 related initiatives working with industry:

1. developing an AI Safety Standard, implementing risk-based guardrails for industry
2. considering watermarking or similar data provenance mechanisms
3. establishing a temporary expert advisory group

AI Safety Standard

Australian business provided feedback that they find it difficult to navigate the plethora of responsible AI principles, guidelines and frameworks. They have also indicated that, while principles are useful, they need to be translated into practical actions. The National AI Centre will work with industry to draw these frameworks together to produce a best-practice and up-to-date voluntary AI risk-based safety framework for responsible adoption of AI in Australian businesses.

This will create a single source for Australian businesses seeking to develop, adopt or adapt AI. It will complement work to develop a similar framework for government use of AI.

Watermarking or similar mechanisms

The Department of Industry, Science and Resources will work with industry to consider the merits of voluntary watermarking or similar data provenance mechanisms for AI developed and used in high-risk settings.

Temporary expert advisory group

The Department of Industry, Science and Resources is establishing an interim expert advisory group to support the government's development of options for AI guardrails. Submissions urged the creation of a permanent advisory body, which may be considered in future.

Actions

The Australian Government will consider and consult on the case for and the form of new mandatory guardrails for organisations developing and deploying AI systems in high-risk settings. In any high-risk settings where mandatory guardrails already exist, the Australian Government will look to leverage existing requirements.

The Australian Government will ask the National AI Centre to work with industry to develop an AI Safety Standard to provide industry with a practical, voluntary, best-practice toolkit that ensures that AI systems being developed or deployed are safe and secure.

The Australian Government will commence work with industry, including developers and deployers, on the merits of voluntary labelling and watermarking of AI-generated material in high-risk settings.

Clarifying and strengthening laws to safeguard citizens

Significant work is underway or planned across the government to address issues raised during consultation on regulatory and policy frameworks. The department is working with agencies leading this work to ensure that views in submissions can inform these processes. This includes:

- developing new laws that will provide the Australian Communications and Media Authority with powers to combat online **misinformation and disinformation**
- an independent statutory review of the ***Online Safety Act 2021*** to ensure that the legislative framework remains responsive to online harms
- working with the state and territory governments, industry, and the research community to develop a regulatory framework for **automated vehicles** in Australia, including interactions with work health and safety laws
- ongoing research and consultation by the Attorney-General's Department and IP Australia, including through the AI Working Group of the IP Policy Group, on the implications of AI on copyright and broader IP law
- implementing the **privacy law reforms**
- strengthening Australia's competition and consumer laws to address issues posed by **digital platforms**
- agreeing an **Australian Framework for Generative AI in Schools** by education ministers to guide the responsible and ethical use of generative AI tools in ways that benefit students, schools and society while protecting privacy, security and safety
- ensuring the **security** of AI tools, such as using principles like security by design, through the government's work on the Cyber Security Strategy.

In addition, the Australian Government, as well as state and territory governments, will continue to consider areas where existing laws could be strengthened to address risks and harms posed by AI.

Action

Building on recent (for example, privacy law) and proposed (for example, online safety and mis- and disinformation) reforms, the Australian Government will consider suggestions put forward in submissions on further opportunities to strengthen existing laws to address risks and harms from AI.

Working internationally to support the safe development and deployment of AI

Global AI Safety Summit

As highlighted at the recent AI Safety Summit, Australia has a strong interest in working with international partners to mitigate the risks of AI through global governance mechanisms. A focus is the risks from frontier AI that is not being developed in Australia.

In November 2023, Australia, alongside the EU and 27 countries including the US, UK and China, signed the Bletchley Declaration at the first Global AI Safety Summit (the summit). The declaration affirmed that AI should be designed, developed, deployed and used in a way that is safe, trustworthy, responsible and takes a community-first approach. The declaration highlights that proactive, risk-based international collaboration is required to help ensure the safety of frontier AI. It signals Australia's commitment to work with international partners to ensure AI is developed within the right guardrails.

One important challenge identified before and during the summit was the fragmented and incomplete understanding of frontier AI. Participating countries agreed that, to fully realise the opportunities presented by AI, governments need to take the lead in building public trust. This requires clarity about the technology itself. Countries, including Australia, committed to working together on a *State of the Science* report on frontier AI. This will facilitate a shared science-based understanding on the risks and capabilities associated with frontier AI. Australia, through CSIRO Chief Scientist Bronwyn Fox, will join the Expert Advisory Panel overseeing the annual AI *State of the Science* report.

Australia, with other governments and companies, agreed to test and evaluate the next generation of AI models against a range of critical national security, safety and societal risks. To support this, both the UK and the US are setting up AI safety institutes. Likewise, Singapore's Infocomm Media Development Authority has signed an agreement with the UK to build evaluation tools.

International responses

Australia will continue to monitor and work with partners to understand action being undertaken in key jurisdictions and in key international forums that develop technical standards for AI. Australia has an interest in ensuring that any domestic responses to support safe and responsible AI are interoperable with global processes.

Since the discussion paper (which included a high-level summary of domestic responses by key jurisdictions at the time) was released, there have been several significant developments. This highlights the significant pace of reform underway globally.

European Union

In December 2023, the European Parliament and European Council reached agreement on the EU's *Artificial Intelligence Act* (EU AI Act) paving the way for formal approval before the regulation becomes law. The EU AI Act includes a risk-based approach under which high-risk systems will be required to undergo rigorous quality assurance before and after they are deployed.

There was also an agreement to apply dedicated rules to general-purpose AI models to ensure transparency along the value chain, with a new European AI Office to supervise the implementation and enforcement of these new rules.

Industry is also being encouraged to sign onto a new AI Pact to begin to implement guardrails proposed in the EU AI Act ahead of its formal commencement.

United States

In October 2023, the US issued a landmark executive order to harness the opportunities of AI while managing the risks. It outlines action across the themes of:

- safety and security
- protecting privacy
- advancing equity
- supporting consumers
- promoting innovation and competition
- working with international partners
- supporting government as an exemplar of AI.

Specific actions include:

- establishing standards for red-team testing to ensure AI safety before public release
- guidance for watermarking
- standards requiring developers of powerful AI systems to share safety test results and other critical information with the US Government.

A number of AI-specific bills have also been introduced into the US Congress. This includes an Artificial Intelligence Research, Innovation, and Accountability Bill introduced in November. The Bill, which has bipartisan support, proposes new transparency and certification requirements for AI systems considered to be 'high-impact' or 'critical-impact'.

Canada

The government of Canada is introducing a new voluntary AI code of conduct for Canadian companies' use of advanced generative AI systems. It includes commitments to increase transparency and avoid bias.

Actions

The Australian Government will take forward the commitments it made in the Bletchley Declaration, including supporting the development of a State of the Science report.

The Australian Government will continue to engage internationally to help shape global AI governance. It will also identify and consider opportunities to support the safe and responsible deployment of AI technologies in our region.

The Australian Government will consider ways to bolster the engagement of Australian experts in key international forums that develop technical standards for AI.

The Australian Government will continue to engage with international partners to understand their own domestic responses to the risks posed by AI. This will help strengthen and ensure interoperability with Australia's own domestic responses.

Maximising the benefits of AI

The 2023–24 Budget contains $75.7 million of funding for AI initiatives to help realise these opportunities. This includes:

- $17 million for the AI Adopt program. This program will create new centres giving SMEs support and training to make more informed decisions about using AI to improve their business.
- $21.6 million to expand the remit of the National AI Centre, doing important research and providing leadership for AI industry around Australia. This builds on the existing funding of $8 million from the 2021–22 Budget ($2.6 million in 2023–24).
- $34.5 million of continued funding for the Next Generation Artificial Intelligence and Emerging Technologies Graduates programs to attract and train the next generation of job-ready AI specialists.

This complements the significant private investments that businesses large and small are making in Australia's technology sector. In 2022, private investment in AI in Australia totalled $1.9 billion, up from $1.8 billion in 2021. This brings total private investment in AI to $4.4 billion since 2013.

Building on these important investments, the Australian Government will continue to consider opportunities to support the adoption and development of AI and other automation technologies in Australia, including the need for an AI Investment Plan. This complements efforts to ensure that Australia has in place the necessary guardrails to build trust and confidence in the use of AI.

Action

Building on the Australian Government's existing investment to grow our national capability to develop and adopt automation technologies like AI and robotics, it will consider opportunities to ensure that Australia can maximise the benefits of such technologies.