Australian Government
Department of Industry,
Science and Resources

Safe and responsible AI in Australia

Proposals paper for introducing mandatory guardrails for AI in high-risk settings

September 2024

Our purpose is to help the government build a better future for all Australians through enabling a productive, resilient and sustainable economy, enriched by science and technology.

| **consult.**industry.gov.au/

Contents

Executive summary

Content warning: this discussion paper contains content that some readers may find distressing. It refers to non-consensual deepfake pornography and child sexual abuse material.

The potential for Artificial Intelligence (AI) to improve social and economic well-being is immense. AI development and deployment is accelerating and is already permeating institutions, infrastructure, products and services. This often occurs undetected by those engaging with it.

The Australian Government's consultations on safe and responsible AI have shown that our current regulatory system is not fit for purpose to respond to the distinct risks that AI poses. Internationally, governments are reforming existing regulations and introducing new regulations to address the risks of AI, with a focus on creating preventative, risk-based guardrails that apply across the AI supply chain and throughout the AI lifecycle. To unlock innovative uses of AI we need a modern and effective regulatory system.

In the Australian Government's interim response to the *Safe and Responsible AI in Australia* discussion paper[1], a commitment was made to a develop a regulatory environment that builds community trust and promotes AI adoption. The guardrails outlined in this paper complement those in the Voluntary AI Safety Standard, and set clear expectations from the Australian Government on how to use AI safely and responsibly. They aim to address risks and harms from AI, build public trust and provide businesses with greater regulatory certainty. Implementing the Voluntary AI Safety Standard now will help businesses start to develop practices required in a future regulatory environment.

This paper outlines options the Australian Government is considering to mandate guardrails on those developing and deploying AI in Australia in high-risk settings. These options include adapting existing regulatory frameworks to introduce additional guardrails on AI, or creating new frameworks such as through framework legislation or by introducing an Australian AI Act. A risk-based approach has been developed, with emphasis on measures including testing, transparency and accountability, consistent with developments in other jurisdictions.

In this paper, feedback is sought on:

- **Defining high-risk AI** – the proposed principles for determining high-risk AI settings and their potential application to general-purpose AI (GPAI) models.
- **Mandatory guardrails** applying across the AI supply chain and throughout the AI lifecycle – 10 guardrails proposed for AI systems in high-risk settings to reduce the likelihood of harms occurring from the development and deployment of AI systems. These preventative measures require developers and deployers of AI in high-risk settings to take steps to ensure their products are safe, including in relation to:
    - *testing* during development and in deployment to ensure systems perform as intended and meet appropriate performance metrics
    - *transparency* about how AI products are developed and used with end-users, other actors in the AI supply chain and relevant authorities
    - *accountability* for governing and managing the risks associated with AI systems.
- **Regulatory options to mandate guardrails** – this section considers 3 options for implementing the proposed mandatory guardrails:
    - Option 1 – Domain specific approach – adapting existing regulatory frameworks to include the proposed mandatory guardrails
    - Option 2 – Framework approach – introducing framework legislation, with associated amendments to existing legislation, or
    - Option 3 – Whole of economy approach – introducing a new cross-economy AI Act.

[1] Department of Industry, Science and Resources, *Safe and Responsible AI in Australia*, Department of Industry, Science and Resources, 2023.

These options build on the Australia Government's current work to strengthen and clarify existing laws impacted by AI.

Active engagement and feedback is sought on the issues above from a broad cross-section of industry, civil society, academia, workers and the community.

Introduction

Artificial Intelligence (AI) continues to be developed by organisations at a rapid pace. AI can improve wellbeing and quality of life, grow our economy and support a future made in Australia. However, as AI becomes more pervasive, there is increasing evidence that in some settings, these technologies present risks to people, community groups and society, and result in harm. Public trust in AI remains low, which is slowing adoption by businesses and organisations.

The community expects governments to safeguard the safety and security of AI systems.[2] Governments have an essential role in building public confidence in AI. They are also crucial actors in creating an environment that supports safe and responsible AI use, while reducing the risks posed by these technologies.

Internationally, there has been a shift in how governments regulate AI technology. Actions so far this year include:

- the newly elected UK Starmer Government has announced the intention to introduce an AI bill[3]
- the European Union's approval of the landmark EU AI Act on 13 March 2024[4]
- the unanimous passage by the United Nations General Assembly of the first global resolution on AI on 21 March 2024[5]
- the Council of Europe adopting the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the Convention), the first international legally binding treaty on AI, human rights and the rule of law on 17 May 2024[6]
- 6 states in the US (including California) adopting AI-related resolutions or enacting legislation, for example concerning government use of AI or automated systems, consumer protections for AI and the use of AI or AI-generated content in education settings and political contexts.[7]
- the Canadian Parliament's continued consideration of the proposed Artificial Intelligence and Data Act (AIDA).[8]

[2] A Saeri, M Noetel, and J Graham, *Survey assessing risks from artificial intelligence*, University of Queensland, 8 March 2024.

[3] UK Government, The King's Speech 2024, 17 July 2024.

[4] European Parliament, Artificial Intelligence Act: MEPs adopt landmark law, European Government, 13 March 2024.

[5] United Nations, General Assembly adopts landmark resolution on artificial intelligence, UN News, 21 March 2024.

[6] Council of Europe, *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law,* Council of Europe: Committee of Ministers, 17 May 2024 *Negotiating and observing countries will be invited to sign the Convention from 5 September 2024.

[7] US National Conference of State Legislatures, Summary: Artificial Intelligence 2024 Legislation, US Government, 3 June 2024.

[8] Parliament of Canada, 'Amendments to Bill C-27 (44-1)', Minister of Innovation, Science and Industry, 28 November 2023.

The Australian Government recognises that implementing appropriate guardrails and regulation will support innovation and increase adoption, leading to benefits for Australia. In Australia, the development and use of AI is shaped by a range of legal frameworks, including areas such as consumer, privacy, anti-discrimination, competition, data protection, and copyright law. The Australian Government announced an investment of $4.2 million in the 2024–25 Budget to clarify and strengthen existing laws impacted by AI, with reviews in priority areas of health care, consumer and copyright law.

There are few rules specifically targeting the upstream development and deployment of AI technologies in Australia, as exist in relation to other types of technologies: rules shaping the development and use of gene technologies, or nuclear technologies, for example. This has enabled the rapid development and proliferation of AI technologies, but has contributed to increased community concern that safety concerns specific to AI are not being taken seriously by organisations seeking to develop and use them in high-risk settings. Increasingly, governments are focused on upstream guardrails to prevent harms occurring.

The Australian Government released its Interim Response to the *Safe and Responsible AI in Australia* discussion paper in January 2024, which concluded a broad consensus that voluntary guardrails and commitments from companies are insufficient to prevent harms occurring. It also advocated a risk-based approach to AI regulation, acknowledging that a vast number of uses of AI – for example, in optimising parcel delivery, or monitoring weather patterns, or managing traffic flow – are considered low risk and should be enabled to flourish unimpeded. This paper defines high-risk settings, proposes a set of mandatory guardrails, and outlines 3 possible approaches to mandating those guardrails via legislation.

About this paper

The Australian Government has acted on the 4 immediate actions in the interim response. This includes establishing a temporary AI Expert Group to provide advice to the Department of Industry, Science and Resources (DISR) on options for mandatory guardrails for the development and deployment of AI systems in high-risk settings.

The expert group has a broad skill set including law, ethics and technology as well as Aboriginal and Torres Strait Islander (First Nations) values. The diversity of skills ensures the expert group can explore a breadth of issues. It also reinforces opportunities for Australia to be a leading international voice on regulatory approaches for AI that address sustainability, Indigenous Cultural and Intellectual Property (ICIP), Caring for Country and First Nations values. As well as the advice of the expert group, this paper captures information from:

- previous consultations
- ongoing engagement with industry, academia, civil society and the community
- engagement across government.

This paper supports the Australian Government's commitment to consider options for mandatory guardrails for the use of AI in high-risk settings. The paper has 4 parts:

- **The case for regulating guardrails (section 4):** Why we need guardrails that focus on the development and deployment of AI to mitigate risks for the use of AI in high-risk settings.
- **Defining high-risk AI (section 5):** A principles-based approach to defining high-risk AI with known or foreseeable uses, and a definition to capture general-purpose AI (GPAI) models.
- **Guardrails ensuring testing, transparency and accountability for AI (section 6):** Proposed mandatory guardrails, their aims, and how they could apply across the AI supply chain and throughout the AI lifecycle by different actors.
- **Regulatory mechanisms to mandate guardrails (section 7):** Options to mandate guardrails from the adaptation of Australia's existing legal frameworks through to enacting new AI-specific legislation or framework legislation.

The proposed risk-based response places obligations on those across the AI supply chain and throughout the AI lifecycle who can most effectively prevent harms *before* people interact with, or are subject to, an AI system (an *ex ante* or pre-market approach). These obligations should apply to AI developers and deployers. If adopted, Australia's approach would come closer into line with jurisdictions including the European Union, and proposed approaches of Canada, and the United Kingdom who join Australia as signatories to the multilateral Bletchley Declaration.[9] Adopting an approach that focuses on the most advanced AI systems would also build on commitments developers have made through the Hiroshima AI Process Code of Conduct and the Frontier AI Safety Commitments.

In considering guardrails, the government is motivated to harmonise with emerging legal standards where it can, while preserving Australia's local needs and context. This includes Caring for Country, ICIP and Indigenous Data Sovereignty ensuring that AI systems are culturally appropriate. Interoperability with other countries' approaches to AI regulation will help to minimise regulatory burdens on Australian businesses exporting to those countries. It will also help to attract more foreign business investment in Australia, and avoid creating a regulatory 'race to the bottom' where lax regulatory frameworks do not ensure safety of these systems both domestically and overseas.

The Australian Government's integrated approach to AI regulation

AI is considered a field of critical technology that is in the national interest.[10] It has the potential to increase productivity, address skill shortages, facilitate innovation and enable more tailored and accessible services, including in healthcare. For example, generative AI alone could contribute $45 billion to $115 billion to the Australian economy.[11] The Australian Government is taking an integrated approach to mitigating the risks of AI while supporting these innovations in the development and deployment of AI. Some examples of support for innovation that also applies to AI include:

- incentivising R&D development in AI industries under the Research and Development Tax Incentive (RDTI). In 2022–23 the RDTI program supported AI, machine learning and computer vision projects worth a total of $478 million
- the allocation of $1 billion in the National Reconstruction Fund (NRF) to support growth ready critical technology companies
- investing $392 million in the Industry Growth Program to support innovative Small to Medium Enterprises undertaking commercialisation or growth projects in NRF priority areas
- investing $17 million to establish four new centres under the AI Adopt Program giving Small to Medium Enterprises (SMEs) support and training to make more informed decisions about using AI to improve their business
- providing a prize pool of $500,000 worth of research and development support for the 3 winners of the AI Sprint, a program for startups and entrepreneurs to develop AI solutions to solve challenges
- implementing the Next Generation Graduates Program to attract and train the next generation of job ready AI and emerging technology specialists to drive growth of the Australian technology sector.

Effective regulation, uplift of governance skills and capabilities, promotion of best practice and education on how to use AI responsibly are essential to securing the benefits of AI for the Australian community.

[9] On 1 November 2023, Australia alongside the EU and 27 countries signed the Bletchley Declaration affirming that AI should be developed, deployed and used in a manner that is safe, human-centric, trustworthy and responsible.

[10] Department of Industry, Science and Resources, *List of Critical Technologies in the National Interest*, Australian Government, 2023.

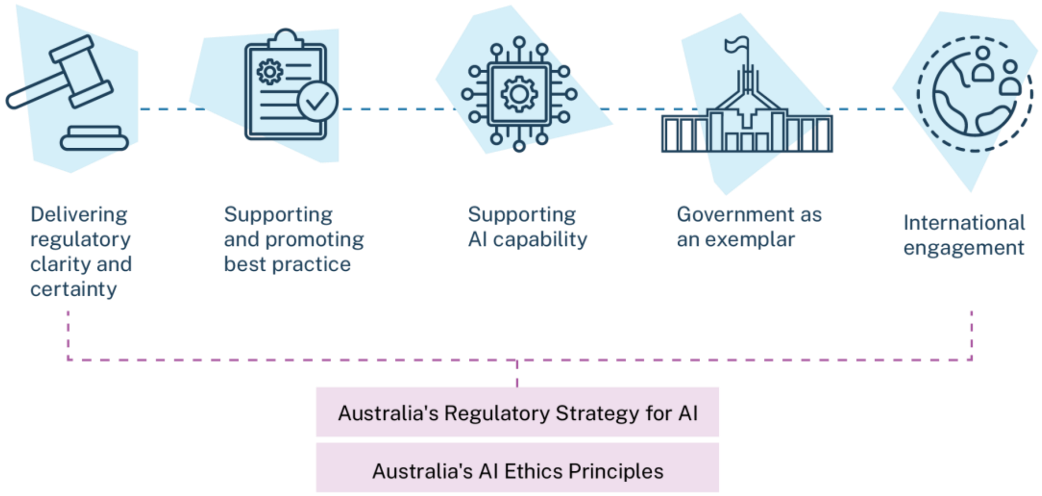
[11] Microsoft and the Tech Council of Australia, *Australia's Generative AI opportunity*, Tech Council of Australia, 2023.

The Australian Government is operationalising Australia's AI Ethics Principles across all its AI policy initiatives.[12] These principles provide important values-based guidance for the intent of regulatory design and align with internationally recognised principles on ethical and responsible AI.[13]

AI systems are becoming increasingly powerful and pervasive across our economy and society. The overwhelming view in submissions to the discussion paper was that voluntary compliance with Australia's AI Ethics Principles is no longer enough in high-risk settings. Adopting ethics principles can help improve safe and responsible practices within organisations developing and using AI. But effective regulation and enforcement is also needed, to create the right settings for AI innovation and adoption in Australia across all sectors.

This paper focuses on proposed mandatory guardrails for the use of AI in high-risk settings. It sits alongside a broader suite of governmental actions (Figure 1) enabling safe and responsible AI under 5 pillars. The details of these actions, including reform processes underway relating to the use of automated decision-making (ADM) by the Australian Government and the Privacy Act Review, have been provided in Attachment B. Collectively, these aim to strengthen capabilities and create the settings needed to support safe and responsible innovation through AI across the economy.

Figure 1: Actions the government is taking to support safe and responsible AI in Australia.


Delivering regulatory clarity and certainty
Supporting and promoting best practice
Supporting AI capability
Government as an exemplar
International engagement
Australia's Regulatory Strategy for AI
Australia's AI Ethics Principles

[12] Department of Industry, Science and Resources, *Australia's AI Ethics Principles*, Department of Industry, Science and Resources, 2019.

[13] Organisation for Economic Co-operation and Development, OECD AI Principles Overview, OECD, 2019

Safe and responsible AI work plan

The Australian Government is taking a coordinated approach to regulating AI. The proposed mandatory guardrails outlined in this paper sit alongside other initiatives, including the development of a Voluntary AI Safety Standard, work to strengthen and clarify existing regulation, international engagement to ensure consistency and coherence, supporting government as an exemplar of AI and maximising the benefits of AI. Our overall objective is to maximise the benefits of AI to the Australian community while minimising harms.

The 5 pillars of the Australian Government's approach to supporting the adoption of safe and responsible AI are outlined below (with further detail at Attachments B and C). It also includes work on a regulatory strategy that reinforces the connections between these pieces of work. It will be developed alongside the government's consideration of options for mandatory guardrails.

1. Delivering regulatory clarity and certainty

Regulatory clarity and certainty for those developing and deploying AI is essential. The Australian Government has taken initial steps to support this goal by:

- working to strengthen and clarify existing regulation applying to the use of AI, including committing to reviews of priority areas in health care, consumer and copyright law
- proposing mandatory guardrails for high-risk AI as outlined in this paper
- developing a concise regulatory strategy to support updating statements of expectation for regulators.

We are also progressing reforms following the review of Australia's Privacy Act, new proposed powers to combat misinformation and disinformation and have brought forward the independent statutory review of the Online Safety Act. Other linked reforms across government will also assist with broader regulatory clarity, including the development of a consistent legal framework for automated decision-making by government.

2. Supporting and promoting best-practice governance

It is important that industry uplifts its governance skills and capabilities to innovate responsibly with AI. To help support this, the government is developing a Voluntary AI Safety Standard for organisations using AI to provide organisation-level and system-level best practice for developers and deployers.

3. Supporting AI capability

The Australian Government will continue to consider how best to support AI capability, including through existing programs like the AI Adopt Program and investments through the National Reconstruction Fund.

4. Government as an exemplar

To support our commitment to be an exemplar in the safe and responsible adoption of AI, we established the AI in Government Taskforce last year. The Taskforce concluded on 30 June 2024. While the Taskforce has concluded, the Digital Transformation Agency (DTA) will continue to develop and implement policies to position government as an exemplar in the use of AI. On 15 August 2024, the DTA released the policy for the responsible use of AI in government to position government as an exemplar in the use of AI. The DTA will also soon pilot a draft Commonwealth AI Assurance Framework to support a more consistent approach by agencies to assessing and mitigating the risks of AI use.

5. International engagement

We engage internationally to share our regulatory approaches and best-practice policies, to secure efficient and effective international AI governance structures, and to learn from other countries' experience. We seek to ensure that our domestic regulations are interoperable where appropriate and that our interests are advanced and protected in international agreements and principles. We work with and through the international AI Safety research network to share research and evaluate risk in AI models and systems.

The case for regulating guardrails

The term ‘artificial intelligence’ was first coined in the 1950s. Precursors to modern AI systems first appeared in the early 2000s, though development, deployment and use were limited and niche in these early days. AI systems, with their ability to be trained using pre-existing data rather than written instructions, were a response to the limitations of simpler software. Unlike traditional software models, AI is developing the flexibility and capacity to try to capture the uncertainty and complexity of reality.

AI technologies have continued to evolve, driven by improvements in algorithms, increased computational power, and widespread creation and availability of training data. Today AI technologies are increasingly part of our day-to-day lives, and many of our interactions with these models go largely unnoticed. Most of the AI models we interact with are ‘narrow AI’: models developed to perform a particular task, such as speech recognition (see the table below on definitions). From the moment we wake up each morning and reach for our phones, or glance at our wearable devices, we’re exposed to, and influenced by, AI.

As outlined in the introduction, AI technologies are evolving at a rapid pace. General-purpose AI (GPAI) models (that is, AI systems that could be used for a wide range of purposes) are the next evolution of AI. GPAI models, such as GPT-n, DALL-E and Sora can now generate ‘human-like’ text, images and videos based on simple user prompts. They can perform levels of human-like general cognition previously only seen in humans.

Key definitions used in this paper (see also Glossary)

AI lifecycle: all events and processes that relate to an AI system's lifespan. This spans from inception to decommissioning, including its design, research, model development, training, deployment, integration, operation, maintenance, sale, use, and governance.

Artificial intelligence (AI) system: a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Agentic AI: AI that is capable of accomplishing multi-step tasks in pursuit of a high-level goal with little or no human oversight.

AI supply chain: the complex network of actors and organisations that enable the use and supply of AI throughout the AI lifecycle from model design, testing and fine tuning to deployment and integration into the local IT system.

Artificial intelligence (AI) system and model

- *AI model:* the raw, mathematical essence that is often the ‘engine’ of AI applications.
- *AI system:* for the purposes of differentiating an AI system from an AI model, the AI system is the ensemble of several components, including one or more AI models, that is designed to be particularly useful to humans in some way.

Applying the above, for example, the ChatGPT app is an AI system. Its core engine, GPT-4, is an AI model.

Developer: organisations or individuals who design, build, train, adapt, or combine AI models and applications.

Deployer: any individual or organisation that supplies or uses an AI system to provide a product or service. Deployment can be for internal purposes or used externally impacting others, such as customers or individuals.

End user: any intended or actual individual or organisation that consumes an AI-based product or service, interacts with it or is impacted by it after it is deployed.

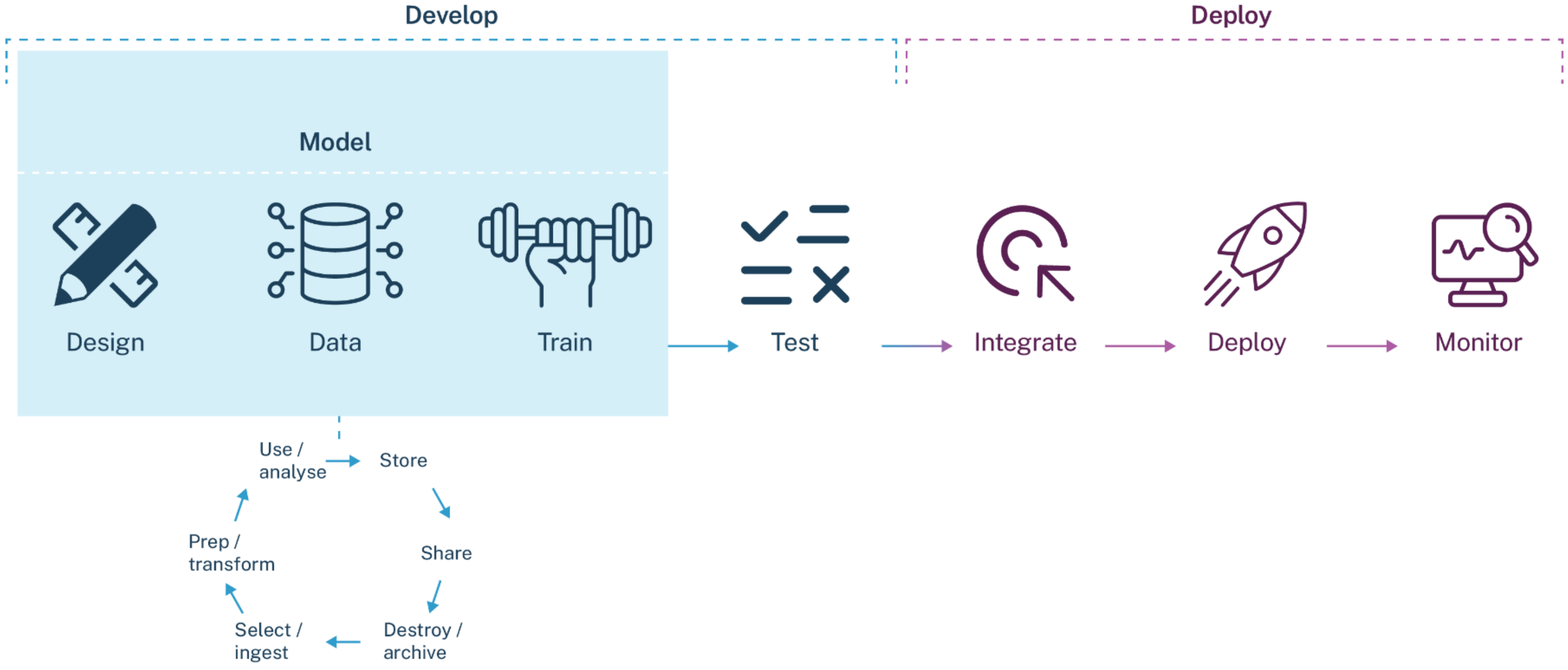
General-purpose AI (GPAI) model: an AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems.

Generative AI model: an AI model with the capability of learning to generate content such as images, text, and other media with similar properties to its training data.

Narrow AI system: a type of AI system or model that is focused on defined tasks and uses to address a specific problem. Unlike GPAI models, these types of AI systems cannot be used for a broader range of problems without being re-designed.

The figure on the following page maps the role of developer and deployer onto the stages of an AI lifecycle from design to model development to deployment and integration into the local IT environment.

Figure 2: AI lifecycle.



Design	Data	Train	Test	Integrate	Deploy	Monitor
Define the problem, design specifications and success measures.	Prepare data for training.	**Pre-train** Select an algorithm for training. **Fine tune** Select an existing model for fine-tuning.	Test and evaluate model against success measures.	Integrate model into IT systems to end-users.	Deploy systems to end-users.	Perform ongoing testing and monitoring.

Why is AI different?

AI shares many similarities with other technologies. Like other technologies, its goal is to solve problems, find efficiencies and accelerate innovation. Past waves of technology development – for example advances in gene technologies, nuclear technologies, and transport technologies – have in some cases, necessitated technology specific infrastructure and rules to ensure safety and community expectations are also respected. Submissions to the discussion paper identified characteristics in advanced AI models that necessitate focus. The temporary AI Expert Group also agreed that AI has characteristics, as distinct from other types of software programs, that warrant a specific regulatory response. This was reinforced in the recent *International Scientific Report on the Safety of Advanced AI: Interim Report (the International Scientific Interim Report)*, a key deliverable from the Bletchley AI Safety Summit.

These differences are summarised below:

- **Autonomy**: Services and products embedded in AI technology or stand-alone AI applications are becomingly increasingly autonomous. AI systems can make decisions autonomously and pervasively, without human intervention at any stage of the decision-making process, if designed by organisations to function that way.
- **General cognitive capabilities**: General purpose systems like large language models can exhibit behaviour that, in humans, would require general cognitive capabilities, as opposed to sophisticated but specific capabilities to solve a task. This includes the capacity to transfer learning across domains and apply it to unseen or new tasks. For example, reading comprehension, writing, drawing, composing music and writing code would require general reasoning capabilities in any human.
- **Adaptability and learning**: AI systems can improve their performance over time and adapt by learning from data. As noted above, this differs from simpler software programs, which often follow pre-defined rules and need explicit programming. For example, GPAI are often trained by scraping, analysing and processing publicly available data from the internet. As AI has become capable of generating data – and even programming code – it has also become a creator of information and technology.
- **Speed and scale**: AI has an unparalleled capacity to analyse massive amounts of data in a highly efficient and scalable way. It also allows for real-time decision-making and distribution of outputs at a scale that surpasses the capabilities of, and diversity of those tasks as previously undertaken by humans.
- **Opacity or lack of explainability:** This is often referred to as the 'black box' problem. The most advanced AI models are trained on data that is often too vast and too complex for humans to efficiently process, and which may not have been curated or documented prior to ingestion for training. Techniques used to reason from data are multi-layered and under-studied, contributing to a limited understanding of their outputs. Decisions that AI systems make are not always traceable.

  This opacity presents several challenges:

  - for those engaging with or being impacted by an AI system who may not be aware of it
  - for those governing or overseeing AI system use in organisations
  - as a barrier to effective enforcement efforts
  - for those engaging with the administrative law system where a decision needs to be clearly documented.
- **High realism**: AI's advancement – and particularly, generative AI – has reached a point where AI can emulate human-like behaviours. This includes creating realistic outputs that make it challenging for end-users to identify when they are interacting with AI or a human (the measure used in the Turing Test), or distinguish between artefacts that are AI-generated rather than human-generated.

- **Versatility**: AI models are a multipurpose technology that can perform tasks beyond those intended by their developers. This is particularly evident with GPAI, even if it is deployed for a particular purpose.
- **Ubiquity:** AI, particularly generative AI, has become an increasing part of our everyday lives and continues to be developed and adopted at a significant rate. It is readily accessible.

AI amplifies and creates new risks

Many of the characteristics of AI outlined above underpin why AI is proving to be such a transformational technology. It is already impacting and often improving various facets of our lives. However, as AI has become more pervasive, there is increasing evidence that these same characteristics can present risks and are contributing to harms.

In some instances, AI amplifies existing risks such as bias. AI systems have the potential to embed human biases and create and amplify algorithmic biases in new ways. This can lead to systemic impacts on groups because of attributes protected under Australian law, such as race or gender. This bias may arise in different ways, including:

- bias due to inaccurate, insufficient, unrepresentative or outdated data
- bias due to the design, configuration of or deployment of the AI system itself.[14]

Harms from AI misuse and failure can occur on several levels:

- harms to people (such as physical or psychological injury, a breach of privacy and exclusion from access to opportunities and services)
- harms to groups of people (such as bias and discrimination based on protected attributes)
- harms to organisations (such as introduction or identification of unknown vulnerabilities in common enterprise software or a cyber-attack exposing training data causing reputational and commercial harm)
- collective harms to society more broadly (such as growing economic inequality, mis- and disinformation and spread of extreme or misogynist content, or eroding social cohesion).

AI systems can also better equip actors threatening Australia's national security. For example:

- AI can accelerate the speed and scale of information creation and dissemination, enabling tailored foreign information manipulation and interference activities on the Australian population[15]
- AI-enabled disinformation campaigns can reach wide audiences with tailored content to erode public trust in our democratic institutions and damage social cohesion
- AI lowers the barriers for non-sophisticated actors to engage in malicious cyber activity, increasing the threat to the public, whether through targeted cybercrime and scams or threats to our critical infrastructure.

In some instances, the risks presented by AI have been realised and harm has occurred. For example:

- Some AI resume screening applications have discriminated unfairly and perpetuated stereotypes and social biases against candidates belonging to a certain ethnic group or gender. This can be because of biased training data, any data design parameters the AI system received, or the outcome the AI-system was trained to provide. AI systems may make decisions about resumes

[14] Australian Human Rights Commission, 'The Need for Human Rights centred Artificial Intelligence - Submission to the Department of Industry, Science and Resources,' p 30, July 2023.

[15] Department of Home Affairs, 'Department of Home Affairs submission to the Inquiry into Adopting Artificial Intelligence', 10 May 2024.

autonomously with limited transparency for candidates, which can make it difficult to identify the bias in training data sets and decisions.[16]

- AI facial recognition software has been used to perpetuate bias and discrimination. One case involving such software falsely identified an African American man as the perpetrator of a crime, leading to his arrest. This was because the AI system was trained mostly on photographs of Caucasian people, due to the photographs' availability and quality of data.[17] In a 2019 study, an AI system falsely matched African American faces up to 100 times more often than it did Caucasian faces.[18] Research shows women with darker skin experience the highest error rates and risks of misidentification.[19]
- Content moderation algorithms designed to detect prohibited content such as sexualised images have censored or suppressed legitimate images of women's bodies and items of clothing. This can affect online businesses or health and other content that targets women.[20]
- First Nations cultural material has been used without consent and misappropriated by AI. For example, generative AI tools are being trained on First Nations artworks without the artists' permission. They are then used to create inauthentic First Nations works of art, reportedly available for sale on popular online marketplaces.[21] This poses further risks to ICIP by not giving attribution to those who hold cultural knowledge and by disregarding cultural protocols that affect how stories are told and by whom.
- AI technologies used in an educational context can contain and perpetuate bias, with serious consequences for a student's future. Research suggests that software used to detect whether an essay is AI-generated can discriminate against non-native English speakers and lead to false accusations of cheating.[22] Likewise, educational grading algorithms have been found to favour students in higher performing schools. [23]

People can experience discrimination or exclusion from engaging with an AI system. Adopting AI in the workplace can also affect workers, who may feel excluded from discussions around how AI is integrated into business contexts.[24] When a poorly designed AI system is adopted at scale, it can cause systemic

[16] For example, Amazon terminated an AI hiring program that favoured male over female applicants, reinforcing existing structural biases in male-dominated roles like software engineering. J Dastin, 'Insight – Amazon scraps secret AI recruiting tool that showed bias against women,' Reuters, 11 October 2018.

[17] J Purtill, 'AI facial recognition scanned millions of driver licenses. Then an innocent man got locked up', ABC, 1 November 2023.

[18] P Grother, M Ngan and K Hanaoka, *Face Recognition Vendor Test (FRVT),* National Institute of Standards and Technology, U.S Department of Commerce, Report 8280 2019, DOI 10.6028/NIST.IR.8200.

[19] J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,' *Proceedings of Machine Learning Research*, 81:1-15, 2018 and Najibi A, 'Racial Discrimination in Face Recognition Technology,' Science in the News blog, Harvard University 2020.

[20] Mauro G and Schellmann H (2023) 'There is no standard' – investigation finds AI algorithms objectify women's bodies, The Guardian, 8 February 2024.

[21] C Wilson, 'AI is producing "fake" Indigenous art trained on real artists' work without permission,' Crikey, 19 January 2024.

[22] Liang W, Yuksekgonul M, Mao Y, Wu E and Zou J, 'GPT detectors are biased against non-native English speakers', *Patterns,* 2023, https://doi.org/10.1016/j.patter.2023.100779; White D, 'Rushi was accused of using AI to cheat. It took him weeks to clear his name', *Sydney Morning Herald*, 13 May 2024, accessed 23 May.

[23] Porter J, 'UK ditches exam results generated by biased algorithm after student protests', The Verge, 18 Aug 2020, accessed 23 May.

[24] UTS Human Technology Institute, 'Invisible Bystanders: How Australian workers experience the uptake of AI and automation', 2024.

social inequality and marginalisation of groups including women, people of colour and people with disabilities.

Recent advances in GPAI also present a new range of serious AI-related risks. The malicious use of these versatile, powerful and general-purpose technologies has already shown their capacity to generate harms. These include:

- harms to a person's dignity, mental health, sense of safety and reputation through the generation of deepfake pornographic content
- harms to society through increasing the efficacy of mis- and disinformation campaigns to fuel political conflicts and undermine democratic systems.

Women and girls are overwhelmingly the targets of non-consensual pornographic deepfake images and videos, including high-profile women and celebrities.[25] Almost 150,000 new deepfake pornographic videos were uploaded to the 40 most used deepfake pornography sites in the first three quarters of 2023 – more than in all the previous years combined.[26] Specific incidents include singer Taylor Swift and Italy's Prime Minister Giorgia Meloni being targeted by deepfake pornographic videos.[27]

In politics, deepfakes' disinformation campaigns are becoming more sophisticated. For example, 2 days before Slovakia's last general election, a deepfake audio recording of the Progressive Party leader, Michal Šimečka, circulated online. The video appeared to show him discussing how he had compromised the integrity of the election.[28]

Even seemingly benign AI systems, such as chatbots, can be harmful. In 2023, a Belgian man reportedly ended his life after a chatbot encouraged him to 'sacrifice' himself to stop climate change.[29]

As GPAI models become more powerful, and the 'frontiers' of AI continue to advance, it becomes more difficult to predict all the foreseeable applications and risks of AI. This is particularly the case where GPAI models have high levels of autonomy (including agentic AI) and adaptability, and can function in multiple uses.

The International Scientific Interim Report highlighted a range of risks presented by GPAI covering malicious use risks, risks from malfunction, systemic risks and cross-cutting factors. It said that:

> General-purpose AI outputs can be biased with respect to protected characteristics like race, gender, culture, age, and disability. This can create risks, including in high-stakes domains such as healthcare, job recruitment, and financial lending. In addition, many widely-used general-purpose AI models are primarily trained on data that disproportionately represents Western cultures, which can increase the potential for harm to individuals not represented well by this data.[30]

The report also outlined:

- well-documented cases of AI systems showing discriminatory behaviour based on race, gender, age and disability status causing substantial harm

[25] Ajder H, Patrini G, Cavalli F, and Cullen L (2019) *The State of Deepfakes: Landscape, Threats, and Impact*, Deeptrace Labs, September 2019. Also for example, Hao K (2021) 'Deepfake porn is ruining women's lives. Now the law may finally ban it,' MIT Technology Review.

[26] N Badshah, 'Nearly 4,000 celebrities found to be victims of deepfake pornography', The Guardian, 22 March 2024.

[27] J Zitser, 'Most victims of deepfake porn never get justice, but Italy's prime minister is out for vengeance', Business Insider, 21 March 2024.

[28] C Devine, D O'Sullivan and S Lyngaas, 'A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning', CNN, 1 February 2024.

[29] I El Atillah, 'Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change,' EuroNews, 31 March 2023.

[30] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 12.

- GPAI makes it possible to generate and disseminate disinformation at unprecedented scale and it is difficult to detect
- GPAI can be used to increase the scale and sophistication of scams and frauds and generate fake compromising content featuring people without their consent.[31]

There has been increased interest in developing agentic AI systems that can autonomously interact with the world with little to no human oversight or intervention. This means users can access faster and cheaper applications of general-purpose AI. Some examples of uses are sending emails, sending instructions to physical equipment, or acting as virtual web-browsing assistants. Current capabilities of agentic AI are limited or in the early stages. However, a subset of researchers believe that there is potential for agentic AI to grow to be advanced enough to take on autonomous powers and complete tasks on its own.

With these developments comes an increased concern over potential 'loss of control' that may arise when these automated processes deviate from the constraints set by humans. The severity and likelihood of potential loss of control situations remains unclear as there is limited research assessing it. Additionally, with its associated reduction in human oversight and intervention, agentic AI may cause more accidents as it is increasingly applied.

Risks may also arise through emergent effects, where an AI model discovers patterns in the training data that developers could not anticipate. This occurs when training the system with larger models, more computing power and more data. Such risks increase with the speed at which AI technology is developing, and the speed and scale of uptake of the technology. This may lead to systemic risks, making it difficult to predict which unintended capabilities may emerge at scale and when.

The modern era of AI calls for a risk-based approach to regulation

Businesses and people who develop and use AI are already subject to Australian laws. These include economy-wide laws like those on privacy, administrative law, online safety, corporations, intellectual property, competition and consumer protection, and anti-discrimination. These also include sector-specific laws, for example applying to medical devices, motor vehicles, airline safety and financial services.

These laws can, and in some cases are, responding to some of the risks of AI. This includes through reforms to Australia's privacy regime and automated decision-making in government or through the actions and guidance from existing regulators. The Australian Government continues to review updates to existing legislation to address AI risks.

However, some AI characteristics are limiting the ability of existing laws to effectively prevent or mitigate risks. Submissions to the discussion paper pointed to these gaps and uncertainties. Examples include:

- clarifying accountability and ensuring legal responsibility is distributed appropriately to developers and deployers best placed to manage the causes of potential harms from AI decisions and applications, particularly as many existing laws were originally drafted on the presumption that humans are taking actions and making decisions
- the need for transparency in the development and use of AI to support affected persons and regulators to establish that harm has occurred when relying on redress available under existing laws, and identify and anticipate potential harms that may arise in future
- regulatory gaps at the development phase of training AI models, such as understanding how developers have trained their models to enable governments, regulators or independent experts to monitor for risk or consider necessary guardrails

[31] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 41, 42, 49, 2024.

- enforcement gaps due to lack of remedies or indirect ones, and practical challenges in questions of proof to exercise rights available under existing laws.

The *International Scientific Interim Report* reinforced this point, outlining:

> …a recurring theme in the discourse on GPAI risk is the mismatch between the pace of technological innovation and the development of governance structures. While existing legal and governance frameworks apply to some uses of general-purpose AI systems and several jurisdictions (like the European Union, China, the USA or Canada) have initiated or completed efforts to regulate AI and general-purpose AI specifically, there often remain regulatory gaps. In a market that is as fast-moving as the general-purpose AI market currently is, it is very difficult to fill such gaps ex-post, because by the time a regulatory fix is implemented it might already be outdated.[32]

A risk-based approach, with a focus on *ex ante* (preventative) measures

Consistent with submissions and findings of the Productivity Commission's *5-year Productivity Inquiry: Australia's Data and Digital Dividend* (2023),[33] the Australian Government will adopt a risk-based approach for any new mandatory guardrails for the use of AI in high-risk settings. In designing a risk-based regulatory regime for AI, government will consider:

- the *levels* of risk and *key characteristics* of known risks
- the balance of *ex ante* (preventative) and *ex post* (remedial) regulatory measures to effectively target and mitigate known risks.

Other scientific policy domains where regulation has sought to mitigate the risk of catastrophic harm under significant uncertainty have used the precautionary principle. These domains include climate change, food safety, pharmaceuticals and chemicals regulation. The precautionary principle is usually applied where a technology has significant uncertainties about its environmental, ethical, health and safety risks.[34]

Adopting a risk-based regulatory framework for AI with a focus on preventative measures aims to avoid catastrophic harm before it occurs. Preventative measures could include risk assessments and management frameworks, or prohibition of certain activities with unacceptable risks. This approach is consistent with a precautionary approach, rather than waiting for post-market liability measures and lengthy, often costly, litigation to shift industry practice. It recognises that in some circumstances, for example where there are system-wide effects, there are limitations in making people accountable for negotiating their own safety.

Several features of AI make it well-suited to a risk-based and preventative approach to regulation:

- potential for significant harms to spread across the economy and community at speed
- potential for catastrophic harm, such as via weaponisation
- potential harms arising not only to people, but to groups of people and society at large
- highly context-specific harms. For example, an AI system deployed within one sector for a particular purpose may present very low risk of harm. However, once applied in a different sector it

[32] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 66, 2024.

[33] Productivity Commission, *5-year Productivity Inquiry: Australia's data and digital dividend*, Vol. 4, Inquiry Report no. 100, 2023.

[34] Organisation for Economic Co-operation and Development, 'Understanding and Applying the Precautionary Principle in the Energy Transition', OECD, 2023.

may present a high-risk of harm due to significant differences in domain-specific risks or impacted parties

- uncertainty about how and what types of AI harms might arise as technology evolves, such as greater autonomy. This uncertainty will require regulatory measures and enforcement tools which can successfully adapt to new forms of AI.

For these reasons several countries, including Australia, are considering regulatory responses focusing on preventing harms from AI technologies before they arise.

The response will also allow consumers, workers and citizens to meaningfully exercise existing legal rights.[35] This response places clear obligations on developers and deployers across the AI supply chain and throughout the AI lifecycle where they can control risks arising during development and deployment, and reduce the likelihood of harm occurring.

Preventative regulatory interventions that are specific to AI are a relatively new consideration in Australia.[36] However, other Australian laws already impose safety obligations before a product or good enters the market. For example, the therapeutic goods regime. [37] Recognition of the potential value of such preventative interventions is increasing globally. This particularly relates to the context of competition regulation of digital markets, with major competition reforms for digital platforms now underway in a number of jurisdictions.[38] The Australian Government has recently provided in-principle support for mandatory service-specific codes to address anti-competitive behaviours of certain digital platforms.[39]

Preventative intervention also aligns to the Australian Government's proactive approach to cyber security. It does so by encouraging 'secure-by-design' and 'secure-by-default' practices in digital technologies and software development by mandating standards and collaboration with industry and international partners.[40]

Risk-based guardrails which clearly allocate responsibility for AI harms across the AI supply chain and throughout the AI lifecycle are essential. These provide developers and deployers of AI systems with the guidance they need to innovate responsibly with AI and invest with confidence. These guardrails would work in complement to Australia's existing regulatory regimes. Obligations set out under these guardrails would not replace, or exempt AI firms from, any existing obligations under existing legislation. This existing legislation includes the *Privacy Act 1988*, *Copyright Act 1968*, *Criminal Code Act 1995*, *Corporations Act 2001, Fair Work Act 2009* and the *Competition and Consumer Act 2010* and administrative law.

Effective and meaningful guardrails will be critical in building the trust and confidence in AI technology that are needed to support its broader adoption across our economy and society. Australians expect to be afforded at least the same protections available to citizens in comparable jurisdictions.

[35] Kaminski, M, 'Regulating the risks of AI', *Boston University Law Review*, Vol 103:1347, 2023.

[36] As outlined in the final section of this paper, to date explicit preventative obligations have been proposed by the eSafety Commissioner under existing provisions of the *Online Safety Act 2021*.

[37] Department of Health and Aged Care, Therapeutic Goods Administration, Therapeutic Goods Administration Regulatory Framework, 1 September 2020.

[38] Organisation for Economic Co-operation and Development, 'Ex Ante Regulation and Competition in Digital Markets', OECD, 2021.

[39] Treasury, 'Government Response to ACCC Digital Platform Services Inquiry', Treasury, 2023.

[40] Australian Government, 2023–2030 Cyber Security Strategy.

Defining high-risk AI

Questions for consultation

1. Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove?

   Please identify any:

   - low-risk use cases that are unintentionally captured
   - categories of uses that should be treated separately, such as uses for defence or national security purposes.

2. Do you have any suggestions for how the principles could better capture harms to First Nations people, communities and Country?
3. Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more defined approach, with a list of illustrative uses, needed?

   - If you prefer a list-based approach (similar to the EU and Canada), what use cases should we include? How can this list capture emerging uses of AI?
   - If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?

4. Are there high-risk use cases that government should consider banning in its regulatory response (for example, where there is an unacceptable level of risk)? If so, how should we define these?
5. Are the proposed principles flexible enough to capture new and emerging forms of high-risk AI, such as general-purpose AI (GPAI)?
6. Should mandatory guardrails apply to all GPAI models?
7. What are suitable indicators for defining GPAI models as high-risk? For example, is it enough to define GPAI as high-risk against the principles, or should it be based on technical capability such as FLOPS (e.g. 10^25 or 10^26 threshold), advice from a scientific panel, government or other indicators?

This section sets out the Australian Government's proposed approach to defining high-risk AI where the mandatory guardrails outlined in section 6 of this paper would apply. Two broad categories are proposed.

The first category relates to instances where the proposed uses of the AI system or GPAI model are known or foreseeable. This is to do with regulating the *use or application* of AI technology. In these cases, risk has been determined with reference to the context in which that AI system will be used or the foreseeable applications for the AI system or GPAI model. Principles have been proposed to guide organisations deciding whether a particular use represents a 'high-risk'. This section includes examples of AI system or GPAI model uses that may be considered high-risk.

The second proposed category of high-risk AI relates only to advanced, highly-capable GPAI models, where all possible applications and risks cannot be foreseen. The risk lies in the potential for these models to be used – or misused – for a wide range of purposes with emergent risks. Australia is considering whether to specially define and capture GPAI models because of their capacity to cause harms to people, community groups and society at a wide-scale and speed. By the time a risk or harm may be foreseeable, it may be too late to apply preventative measures. Feedback is sought on the definition of GPAI models to be subject to mandatory guardrails and what these guardrails should be.

The EU AI Act also explicitly prohibits certain uses of AI. These include exploitative AI, subliminal or deceptive techniques impairing a person's ability to make an informed decision, and biometric categorisation systems inferring sensitive personal information.[41] Feedback is sought on types of AI use that could present an unacceptable level of risk in Australia and should be banned.

High-risk AI based on intended and foreseeable uses

Proposed principles

The following principles are proposed to guide consistent assessment by organisations of whether the use of an AI system is high-risk, and so subject to mandatory guardrails shaping its development and use, based on intended and foreseeable uses. The principles are similar to those adopted or proposed in comparable jurisdictions, such as the EU and Canada. This supports interoperability.

The principles should be considered as a whole when assessing whether the use of the AI system is high-risk. They are intended to complement and be broadly consistent with the various regulatory regimes that exist to protect Australians from risks across a range of domains.

The proposed approach provides flexibility to prevent inadvertently catching low-risk applications in a list of high-risk settings, and to accommodate step-up changes in AI technology as it advances. This is discussed further in 'Principles or a list-based definition' section where it is recognised that some jurisdictions, like the EU and Canada, have explicitly provided a list of high-risk use cases.

Proposed principles:

In designating an AI system as high-risk due to its use, regard must be given to:

a. *The risk of adverse impacts to an individual's rights recognised in Australian human rights law without justification, in addition to Australia's international human rights law obligations*

b. *The risk of adverse impacts to an individual's physical or mental health or safety*

c. *The risk of adverse legal effects, defamation or similarly significant effects on an individual*

d. *The risk of adverse impacts to groups of individuals or collective rights of cultural groups*

e. *The risk of adverse impacts to the broader Australian economy, society, environment and rule of law*

f. *The severity and extent of those adverse impacts outlined in principles (a) to (e) above.*

Australia is committed to the responsible research, development, deployment, acquisition and use of AI across the whole economy, including in the defence and national security domains. Australia proposes to align with other jurisdictions, which have treated national security and defence applications separately from civilian applications, including in the US[42] and in the EU.

[41] EU AI Act, Article 5.

[42] The US Executive Order 141110 and the memorandum for the Heads of Executive Departments and Agencies on *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence* do not cover AI when it is used as a component of a national security system. The Executive Order directs the development of a National Security Memorandum to govern the use of AI as a component of a National Security System. Agencies have also adopted responsible AI principles such as the Department of Defence's *Responsible Artificial Intelligence Strategy and Implementation Pathway* and the Office of the Director of National Intelligence's *Principles of Artificial Intelligence Ethics for the Intelligence Community*.

Human rights

a. The risk of adverse impacts to an individual's rights recognised in Australian human rights law without justification, in addition to Australia's international human rights law obligations

Australia's federal human rights legislation focuses largely on questions of discrimination, particularly on the basis of age, disability, race or sex. Consistent concerns have been heard from the community and experts about the potential for AI systems to discriminate based on these protected attributes. There are multiple examples of this already occurring, including:

- AI software that was designed to review resumes and systematically discriminated against women applying for software engineer positions[43]
- AI powered systems used in the criminal justice system to predict criminal behaviour which have undermined the presumption of innocence and discriminated against minorities[44]
- AI biometric identity verification which has included risks of racial discrimination errors, where women with dark skin experience the highest rates of misidentification[45]

This principle also recognises the notion of proportionality where in some circumstances interference with traditional rights and freedoms, including human rights may be justified. Examples where this may be the case include where the interference is in a pursuit of a legitimate objective, is suitable and necessary to meet that objective and on balance the public interest outweighs the harm done to the individual right.[46] This can apply where the interference is needed to protect public order, national security, public health and safety.

As well as ensuring the principles appropriately capture a key area of potential harm from AI systems, a human rights-based approach also supports interoperability. This is due to other jurisdictions also adopting this approach. These include the EU, Canada and the US (see box below). A notable difference between Australia and these jurisdictions is, however, that Australia does not have a singular, centralised domestic human rights charter or legislation. Recognising this, the proposed principle links not only to Australia's domestic human rights laws but also to rights recognised in Australia's international human rights law obligations. This includes the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social and Cultural Rights (ICESCR).

[43] Australian Human Rights Commission, '*The Need for Human Rights centred Artificial Intelligence - Submission to the Department of Industry, Science and Resources*,' p 30, July 2023.

[44] Women with Disabilities Australia (WWDA), 'Response to 'Safe and responsible AI in Australia' Discussion Paper,' p 5, July 2023.

[45] J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,' *Proceedings of Machine Learning Research*, 81:1-15, 2018.

[46] Australian Law Reform Commission, '*Traditional Rights and Freedoms – Encroachments by Commonwealth Laws (ALRC Report 129)*', 12 Jan 2016.

Human-rights based approaches to defining high-risk AI in other jurisdictions

The EU, USA and Canada all have overarching legislative or constitutional instruments enshrining human rights in law. Australia does not have a similarly comprehensive instrument at a federal level. However, many of the rights enshrined in law in other jurisdictions are implicitly woven through Australia's legal system. This includes the disallowance of acts incompatible with the 7 international human rights instruments to which Australia is a party.

In the **EU AI Act,**[47] the extent to which an AI system has an adverse impact on the rights protected by the EU Charter of Fundamental Rights is relevant in classifying a system as high-risk. The charter provides a broad suite of digital rights, which underpin the comprehensive regulatory mechanisms in the AI Act. For instance, Article 7 provides a right to respect for privacy and communications. Article 8 provides a right to the protection of personal data and the fair processing of that data, with compliance to be assessed by an independent authority.

The AI Act also includes a general provision where an AI system is always considered high-risk if it profiles individuals. This includes profiling on metrics such as work performance, economic situation, health, preferences, interests, reliability, behaviour, location or movement. This is a catch-all provision to encompass systems acting as social crediting or scoring systems.

The draft **United States** Artificial Intelligence Research, Innovation and Accountability (AIRIA) Bill[48] refers to 'critical impact' and high-impact AI use cases. These are defined as instances where AI is used to make a decision with legal effect which poses a risk to constitutional rights or safety.

The **Canadian Government** used the Canadian Human Rights Act (CHRA) to develop their list of 'high-impact' use cases in the AIDA,[49] with a focus on issues of discrimination and biased outputs. The amended Bill also requires that the Governor in Council take into account the risk of adverse impacts on individuals' rights recognized in international human rights treaties to which Canada is a party when varying, adding or removing classes of high-impact systems.

Health and safety

b. The risk of adverse impacts to an individual's physical or mental health or safety

There is significant community concern about the impact AI systems have on individual health and safety. These risks will continue to evolve as AI integrates further into everyday products and industries. This principle means to ensure that risks to individual physical and mental health and safety are considered when determining whether an AI system meets Australia's definition of high-risk.

Other jurisdictions include considerations of people's health and safety in defining high-risk. In the EU, any AI product that is part of a safety system is always considered high-risk. In the United States' AIRIA, one of several AI Bills before Congress, systems are considered 'high-impact' wherever there is a general risk to safety, in combination with risks to legal and constitutional rights.

While AI offers opportunities for improved health outcomes, there is a risk that it can be developed or implemented in a way that poses a risk to people's health and safety. AI products used in health screening to determine a patient's need for treatment or further investigation may be influenced by the representativeness of data and its applicability to the Australian population. For example, cultural differences in patients' diets may produce a data bias in a product used for bowel screening. In another

[47] EU AI Act.

[48] Congress of the United States, '*Artificial Intelligence Research, Innovation, and Accountability Bill 2023*, U.S. Government, 15 November 2023.

[49] Parliament of Canada, '*Bill C-27 (44-1)*', Government of Canada, 22 November 2021 (AIDA). See also: Parliament of Canada, '*The Artificial Intelligence and Data Act (AIDA) - Companion Document*', Government of Canada, 13 March 2023, and Parliament of Canada, '*Amendments to Bill C-27 (44-1)*', Minister of Innovation, Science and Industry, 28 November 2023.

example, an AI driven pulse oximeter was found to overestimate blood oxygen levels in patients with darker skin, resulting in the undertreatment of their hypoxia in this patient group.[50]

Legal effects

c. The risk of adverse legal effects, defamation or similarly significant effects on an individual

A legal effect is something that affects the legal rights of an individual. This is particularly relevant where it is not reasonably possible for people affected by AI systems to opt out of that system. It's also relevant where the system affects their access to essential services, such as but not limited to law and enforcement, and access to housing and finance (whether provided by government or the private sector).

This principle reflects the approach other jurisdictions take in ensuring AI does not unduly affect people's legal rights. For example, the EU's General Data Protection Regulation (GDPR) restricts entities from making solely automated decisions that have a legal or similarly significant effect on individuals. Article 22(1) states:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

The EU's Charter of Human Rights includes an explicit protection for the rights of its citizens. The EU AI Act considers the extent to which an AI system might infringe on these rights housed in this charter in its definition of high-risk AI. Australia's human rights framework is structured differently. Our rights are protected through different laws and processes rather than through a charter. Our framework also does not include an explicit protection for data rights in the same way. Given these differences, it may be appropriate to include a broad principle concerning the extent to which an AI system has an adverse legal effect on people. This would then account for rights protected elsewhere under Australian law, such as in employment legislation or administrative law guaranteeing rights to legal and procedural review.

While significant effects are more difficult to define, some examples may include impacting access to health services, and e-recruiting practices without human intervention.[51] This principle is expected to capture use of AI by government service providers that could produce adverse impacts on people.

Impacts on groups

d. The risk of adverse impacts *to groups of individuals or collective rights of cultural groups*

Previous consultations revealed that AI systems have the potential to create highly unequal or damaging outcomes for specific groups and perpetuate existing inequalities.[52] This principle acknowledges the need to protect marginalised groups such as First Nations people, as well as the need to prevent adverse impacts based on gender and other intersecting factors.

[50] A Gregory and A Hearn, '*AI poses existential threat and risk to health of millions, experts warn*', The Guardian, 10 May 2023.

[51] WP29, '*Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*', European Parliament, 17/EN WP251 rev.01, 6 February 2018.

[52] Australian Human Rights Commission, '*The Need for Human Rights centred Artificial Intelligence – Submission to the Department of Industry, Science and Resources*', July 2023.

Systemic impacts

e. The risk of adverse impacts to the broader Australian economy, society, environment and rule of law

This principle addresses the potential systemic risks to the Australian community due to the scale of AI use and adverse impacts on society at large. This principle also addresses instances where an AI system may have a direct adverse impact on the environment or care for Country (for example, through poorly designed automated environmental approvals).

AI can enable the creation and dissemination of harmful synthetic content, leading to mis- and disinformation and manipulation of mass public opinion. This can also turbocharge scam activity. These activities can have broad effects on Australian society. In the political sphere, bad actors can use AI-generated content to unduly influence public perception of candidates or spread disinformation about electoral processes and democratic systems. They can also inundate political actors with AI-generated content with the aim of inaccurately representing constituent sentiment.[53] Dissemination at speed and scale can undermine trust in our institutions and public information, and have far-reaching impacts on social cohesion at an aggregate level. Some examples include:

- algorithms on popular social media and media websites have been shown to amplify racist and misogynistic content and push it at young people, normalising sexist and other negative ideologies[54]
- chatbots can provide inappropriate and harmful responses to user prompts
- hyper realistic generative AI deepfakes are spreading, as is the creation of synthetic child sexual abuse material[55]
- the amplification of content that promotes discrimination, such as sexism, misogyny, homophobia or racism, can have adverse effects by normalising prejudice or hate. This may also contribute to radicalisation towards terrorism and violent extremism[56]
- scammers can use AI-generated voice clones in scam calls to convincingly imitate government agencies. This can result in large scale fraud, digital identity theft where a person's biometric details are used to bypass existing voice authentication protocols and damage public trust in institutions.

As the Australian Electoral Commission outline in their submission[57] to the Senate Select Committee on Adopting Artificial Intelligence, in 2024 alone there have been many documented instances globally of AI-produced content being used to influence the integrity of electoral events. Specific examples include:

- in the US, a robocall using AI voice cloning technology replicating President Joe Biden's voice was used to encourage voters to skip the primary election[58]

[53] Department of Home Affairs, '*Department of Home Affairs submission to the Inquiry into Adopting Artificial Intelligence*', 10 May 2024.

[54] Reset Australia (2022) *Algorithms as a weapon against women*, discussion paper April 2022. See also Regehr K, Shaughnessy C, Zhou M & Shaughnessy N (2024) *Safer Scrolling: How algorithms popularise and gamify online hate and misogyny for young people*, University College London and University of Kent.

[55] Generative AI – position statement | eSafety Commissioner.

[56] Recommender systems and algorithms – position statement | eSafety Commissioner.

[57] Australian Electoral Commission, '*Select Committee on Adoption Artificial Intelligence (AI) Australian Electoral Commission (AEC) Submission'*, 10 May 2024.

[58] J Purtill, 'AI is changing how elections are fought, from deepfake endorsements to chatbot campaigners', ABC News, 21 February 2024. E Steck, and A Kaczynski, 'Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's democratic primary', CNN, 22 January 2024.

- in India, before elections in May, an AI-generated deepfake of deceased former Tamil Nadu Chief Minister and Bollywood star M Karunanidhi praised the leadership of his son and current Tamil Nadu Chief Minister[59]
- in Pakistan, imprisoned former Prime Minister Imran Khan claimed election victory for his party in an AI-generated video.[60]

An academic roundtable convened by the Strengthening Democracy Taskforce in Home Affairs in October 2023 further detailed the impact of AI on democratic processes. They identified 7 dimensions.

7 dimensions of AI’s impact on democracy[61]

1. **Connectedness – Polarisation**

A more connected society has greater democratic resilience than a polarised one. Current AI technologies have the potential to push democracies toward greater polarisation.

2. **Transparency – Opacity**

Transparency is a key pillar of democracy. A large proportion of current AI technologies are characterised by opaqueness including how and what data is collected, how decisions are made, who is responsible for those decisions and where the technology is leading democracies.

3. **Decentralisation – Consolidated control**

Democracy, by definition, decentralises power to the public by way of elections, the rule of law, a free press and other principles and processes. The current trajectory of AI technology is leading to greater control by small number of players.

4. **Democratising voices – Narrowing voices**

Liberal democracy defends and benefits from a plurality of voices, whereas AI technology has the potential to be exclusionary or biased against certain voices in its development and application.

5. **Truth and deliberation – Deception**

Democracy is founded on the ability to have an equal stake in the future, which requires access to factual information in order to make informed decisions. The current trajectory of AI is increasing the prevalence of deceptive material in the information environment.

6. **Public good – Private gain**

Liberal democracy as a system has at its foundation the aim of providing good to the broadest element of the public without undermining the rights of the minority. AI technologies are often seen to further private gain rather than the public good.

7. **Information engagement – Information transmission**

Democracy is based on deliberation of ideas, a discussion of what the public values and what it wants. Current AI technologies are predominantly about the supply and transmission of information rather than the deliberation of it.

The EU AI Act also recognises these risks:

> Aside from the many beneficial uses of AI, that technology can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices. Such practices are particularly harmful and abusive and should be prohibited because they contradict Union values

[59] N Christopher, ‘How AI is resurrecting dead Indian politicians as election looms’, Al Jazeera, 12 February 2024.

[60] G Peshimam, ‘Pakistan's jailed ex-PM Imran Khan claims election victory’, Reuters, 10 February 2024.

[61] Department of Home Affairs, ‘Department of Home Affairs submission to the Inquiry into Adopting Artificial Intelligence’, 10 May 2024.

> of respect for human dignity, freedom, equality, democracy and the rule of law and fundamental rights.[62]

The EU AI Act subsequently includes that AI systems used to influence the outcome of democratic processes should be defined as high-risk:

> AI systems intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda should be classified as high-risk AI systems with the exception of AI systems whose output natural persons are not directly exposed to, such as tools used to organise, optimise and structure political campaigns from an administrative and logistical point of view.[63]

Severity and extent of impacts

f. The severity and extent of those adverse impacts outlined in principles (a) to (e) above

In its interim response, the Australian Government outlined that its focus in considering mandatory guardrails for high-risk AI, was to ensure that lower risk AI can develop largely unimpeded. To ensure that guardrails are appropriately targeted at AI systems that represent a *high* risk, the types of risks and their likely severity and extent of impact will need to be assessed. This could include considering:

- who may experience impacts, including if communities or groups, such as women or culturally and linguistically diverse communities are affected as they can experience disproportionate impacts compared to other groups
- the scale and intensity of those harms in potential numbers of people affected
- the likely consequences and disproportionate impacts for the people, groups or society from the potential harms
- the extent to which adverse impacts are likely to occur (that is, their likelihood)
- the effectiveness of any measures taken to address or mitigate the relevant risk of harm.

The next 2 sections discuss how other countries have identified high-risk use cases and provides some examples of how organisations can apply this principle.

Principles or a list-based definition

The principles set a foundation for defining high-risk AI where the proposed uses of an AI system or GPAI model are known or foreseeable. Feedback received on the merits of principles vs a list-based approach, and on the regulatory options, will guide how these are implemented in high-risk settings.

For example, the Australian Government could set out use cases as part of a defined list of high-risk settings in a standalone AI-specific Act or framework legislation. This would be like the approach the EU and Canada have adopted.

Alternatively, use cases could be drawn upon to inform any centralised guidance material that the government gives under existing regulatory frameworks. As outlined in section 7, some of the areas identified below (for example medical devices) are already subject to substantive pre-market safety regimes.

The table below (Table 1) gives a high-level summary of the areas that the EU and Canada have identified as high-risk use cases. Some jurisdictions are also considering, or have, exemptions to support law enforcement or have separate policies for the application of AI in national security contexts. Feedback is sought on whether these use cases are equally applicable to Australia.

[62] EU AI Act, recital 28.

[63] EU AI Act, recital 62.

AI systems are being deployed in different ways across the economy. The downside of a defined list of high-risk use cases is that it may unintentionally capture some low-risk uses. Within many high-risk domains, there may be use cases which can be assessed as posing limited or minimal risk. The EU AI Act includes mechanisms to ensure that low-risk AI uses are not inadvertently captured.[64] Depending on the preferred regulatory mechanism to mandate guardrails (section 7), a similar mechanism may also be appropriate in the Australian context. This will ensure that any mandatory guardrails remain appropriately targeted.

Feedback or examples are sought on the types of 'low-risk' applications that may get caught, but which should not be subject to the guardrails. The Australian Government will consider this feedback as part of any policy deliberations to introduce exemptions or carve-out mechanisms like the approach taken by the EU.

Table 1: High-risk use cases identified in other countries.

Domain areas	General description
Biometrics	AI systems used to identify or categorise individuals, assess behaviour or mental state, or monitor and influence emotions.
Critical infrastructure	AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.
Education/Training	AI systems used in determining admission to education programs, evaluating learning outcomes or monitoring student behaviour.
Employment	AI systems in employment matters including recruitment, referral, hiring, remuneration, promotion, training, apprenticeship, transfer or termination.
Access to essential public services and products	AI systems used to determine access and type of services to be provided to individuals, including healthcare, social security benefits and emerging services.
Access to essential private services	AI systems used to make decisions that affect access to essential private services, including credit, insurance in a manner that poses significant risk.
Products and services affecting individual and public health and safety	AI that is intended to be used as a safety component of a product, or is itself a safety product or something that impacts on individual and public health and safety. This includes AI-enabled medical devices, food products and other goods and services.
Law enforcement	AI systems used in aspects of law enforcement, including profiling of individuals, assessing offender recidivism risk, polygraph-style technologies or evaluating evidence.
Administration of justice and democratic processes	AI systems used for making a determination about an individual in a court or administrative tribunal, such as systems used for evaluating facts, evidence and submission to proceedings.

[64] This includes where (i) the AI system is intended to perform a narrow procedural task; (ii) the AI system is intended to improve the result of a previously completed human activity; (iii) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or (iv) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases [designated as high-risk in the Act].

Domain areas	General description
	With regards to democratic processes, may include any system which can influence the voting behaviour of individuals or the outcome of an election or democratic process.

A more detailed table outlining the specific applications prescribed in the EU and Canada is in Attachment D.

Examples of how to apply the principles to use cases

Some high-level examples have been provided to illustrate areas organisations need to consider in assessing whether their use of AI is considered a high-risk setting. These examples are not exhaustive. They are intended to illustrate borderline settings for the purposes of seeking feedback on this proposals paper.

AI in employment and workplace settings

Using AI systems in employment settings can have substantial impacts on a person's opportunities. These include in recruitment and hiring, promotions, transfers, pay and termination. To decide if a particular AI system meets the principles, an organisation would need to consider:

- the type of impact it would have on people
- any potential discriminatory impacts on people from a particular cohort
- any society-wide impacts based on the scale of the deployment
- the severity and extent to which the risks are likely to occur.

For example, an automated CV scanning service that makes a determination of an individual's suitability for a job would be considered a high-risk system. This is because it has the potential to impact a person's access to work as well as discriminate against certain groups. Similarly, using an automated rostering system could be considered high-risk as it has discriminatory potential and could also impact a person's ability to participate in work where this risk was likely to occur. For example, if the AI rostering system did not take into account an employee's caring duties. An AI system automating the evaluation of worker performance for making consequential determinations on their employment – for example via monitoring activity at a computer, or on factory floors – may also be classified as high-risk.

By contrast, there are other uses of AI systems in workplace settings which are unlikely to be high-risk. For example, an AI system used to automatically pre-fill payroll information based on existing time and attendance data.

Facial recognition technology

Many high-profile national retailers have recently been trialling and deploying AI driven facial recognition technology, for uses in combatting theft and anti-social behaviour. The *Privacy Act 1988* regulates the collection, use or disclosure of biometric information that is used for automatic biometric verification and biometric identification. The Australian Government is considering the need for new, specific obligations for biometric information, such as in the context of facial recognition technology, as part of its broader Privacy Act reform process. However, using AI-powered facial recognition technology may also be relevant to several principles discussed above. For instance, these systems may pose risks to individual's human rights, such as their privacy rights and risks of discrimination due to subsequent use of the collected biometric details for criminal or other profiling.

For the purposes of applying the mandatory guardrails, not all uses of facial recognition are necessarily high-risk uses of AI, such as unlocking a personal phone using facial recognition. Assessments on whether a use of facial recognition is high-risk will depend on:

- how any outputs may be used, and the risk that these outputs pose to a person's rights (principle a) and health and safety (principle b)
- the impact of personal information collection on the extent of legal effects on affected people (principle c)
- if it is used to create decisions with adverse legal effects on individuals such as determinations to access government or other essential services (principle c).

High-risk AI: General-purpose AI

The previous section outlines potential risks of AI systems that are foreseeable. Use of GPAI models have the same potential for foreseeable risks, but also pose unforeseeable risks because they can be applied in contexts they were not originally designed for.

This section proposes a definition for GPAI, outlines international examples of the application of guardrails to GPAI models, and seeks feedback on whether the proposed mandatory guardrails should apply to all GPAI models.

Proposed approach to defining high-risk GPAI models

Drawing on Canada's AIDA, this paper proposes the following definition of GPAI:

> An AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems.[65]

This definition focuses on what the technology can do rather than its intended use, and is sufficiently broad to capture the potential risks.

International examples of the application of mandatory guardrails to GPAI models

In the US, an Executive Order issued by President Biden mandates reporting requirements on GPAI models above certain thresholds of capability.[66] These requirements relate to reporting and information sharing.

[65] Based on but not identical to the Parliament of Canada, '*Amendments to Bill C-27 (44-1)*', Minister of Innovation, Science and Industry, 28 November 2023.

[66] United States, Executive Office of the President Joseph Biden, '*Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*', U.S. Government, 30 October 2023.

Canada and the EU have gone further. Canada has proposed guardrails that apply to all GPAI. The EU has a minimum set of guardrails for all GPAI models, and further requirements for GPAI models that are deemed to pose 'systemic risk' (defined below).

Draft legislation in Canada[67] sets out that any model captured by the proposed definition of GPAI must comply with mandatory guardrails. This assumes that all GPAI are inherently high-risk because of the nature and ability of such models.

As an example, in Canada, these guardrails include:

- data measures
- assessment of reasonably foreseeable adverse impacts
- a plain language description of the system
- tests on the effectiveness of mitigation measures
- human oversight.

The EU AI Act applies some guardrails to all GPAI models, with some further guardrails only applying to a more specific category of GPAI that pose 'systemic risk'. The EU AI Act defines 'systemic risk' as:

> "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the supply chain".[68]

High-impact capabilities are quantified by various technical measurements. These measurements include the complexity of the model, how well it performs on tests, and its number of users.

Options for defining GPAI that would be captured as high-risk AI

Given GPAI models pose unforeseeable risks, the Australian Government proposes to apply mandatory guardrails to all GPAI models. Since most highly capable GPAI models are not currently developed domestically, Australia's alignment with other international jurisdictions is important to reduce the compliance burden for both industry and government and enables pro-innovation regulatory settings.

Feedback is welcome on how mandatory guardrails should apply to GPAI models in Australia, and whether any further approaches should be considered, such as where a sub-set of guardrails or different guardrails might be needed.

[67] AIDA.

[68] EU AI Act, Article 51.

Guardrails ensuring testing, transparency and accountability of AI

Questions for consultation

8. Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings? Are there any guardrails that we should add or remove?
9. How can the guardrails incorporate First Nations knowledge and cultural protocols to ensure AI systems are culturally appropriate and preserve ICIP?
10. Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately? For example, are the requirements assigned to developers and deployers appropriate?
11. Are the proposed mandatory guardrails sufficient to address the risks of GPAI? How could we adapt the guardrails for different GPAI models, for example low-risk and high-risk GPAI models?
12. Do you have suggestions for reducing the regulatory burden on small-to-medium sized businesses applying guardrails?

This section sets out proposed mandatory guardrails for AI systems used in high-risk settings, as well as high-risk GPAI systems. These guardrails aim to reduce the chance of harms occurring from the development and deployment of AI systems. In doing so, they seek to build trust and confidence in the use of such systems.

The Australian Government acknowledges that these preventative guardrails will need to operate alongside laws and regulatory instruments to hold organisations accountable when harm does occur and help people exercise their existing rights. The precise mechanisms to enable enforcement and liability of these guardrails will be shaped by our regulatory approach, discussed in section 7.

The proposed mandatory guardrails are preventative measures that would require developers and deployers of high-risk AI to take specific steps across the AI lifecycle. These measures focus on:

- *testing* to ensure systems perform as intended and meet appropriate performance metrics, both prior to and during deployment
- *transparency* regarding product development and use with end-users, other actors in the AI supply chain and relevant authorities
- *accountability* for governing and managing the risks of AI systems.

This aligns with international standards and approaches to governance of AI systems other key jurisdictions are taking.

Mandatory guardrails must be flexible and adaptable. They are not static and are intended to evolve to keep up with technological advancements. This is because:

- guardrails need to be tailored to the particular risk profile of an AI system (for example, a GPAI model compared to a narrow AI model)
- obligations need to reflect the role that different entities play in the AI supply chain
- as AI continues to evolve, so too will ‘best practice’ in the management of AI systems.

It is also important to note that the guardrails will not always require organisations to develop entirely new processes for managing risks. Where possible, organisations should integrate guardrails into existing processes.

Importance of interoperability and standards

Developing a regulatory framework that aligns nationally and with likeminded jurisdictions will reduce the compliance burden for both industry and government. It also enables regulatory settings that don't hinder investment. This is key to unlocking AI's full potential for Australian businesses and society while also ensuring the community is protected.

The mandatory guardrails are intended to be interoperable with those that other comparable jurisdictions have developed and adopted. In particular, the mandatory guardrails have drawn on the work and development of the EU's AI Act and Canada's proposed *Artificial Intelligence Data Act (AIDA)*.

The mandatory guardrails are also intended to align with national and international standards.[69] Recent advances in AI technology have led to the development of new AI-related standards. For example, ISO/IEC 42001:2023 *Artificial Intelligence Management System* has been recognised as a standard which will support AI governance in Australia and internationally (see the case study below).

Standards can play a formal role in any regulatory system – with laws requiring adherence to standards as a means of showing compliance with a principles-based law. The EU AI Act, for example, makes it clear that the use of standards will play a key role in providing technical solutions to enable compliance with its requirements.[70] This could also be considered in Australia depending on the preferred regulatory mechanism ultimately adopted.

ISO/IEC 42001:2023 Artificial Intelligence Management System

ISO/IEC 42001 is an international standard that specifies requirements and provides guidance on establishing, implementing, maintaining and continually improving an Artificial Intelligence Management System (AIMS) in an organisation.

ISO/IEC 42001 provides a baseline that helps increase the reliability, transparency and trust in AI systems while mitigating risks and concerns around bias, fairness, privacy and accountability. The protocols within the standard align with our approach to developing guardrails. The standard sets out objectives related to accountability, testing and transparency. It gives high-level guidance on process management for the safe and responsible development and deployment of AI.

The standard helps organisations to develop and use AI systems responsibly and effectively. For example, some of the topics it discusses include:

- where responsibility falls across organisations, their partners, suppliers, customers and third parties
- provision of necessary documentation, especially to third parties that organisations are supplying AI systems to
- assessing impacts of AI systems on groups, based on the intended purpose and use of these systems.

ISO/IEC 42001 applies to any organisation regardless of its size, type or nature that develops, deploys or uses AI-based products and services.

Applying the guardrails across the AI supply chain and throughout the AI lifecycle

AI systems often involve a complex network of actors responsible for different aspects of the system's development and deployment. The supply chain and lifecycle of an AI system can vary significantly

[69] Standards are documents that set out specifications, procedures and guidelines that aim to ensure products, services and systems are safe, consistent and reliable.

[70] EU AI Act, Article 40.

depending on context, such as whether systems are developed in-house or by an external developer. It is likely that multiple organisations will be responsible for carrying out different aspects of the guardrails for a single AI system. As regulatory measures are further developed, it will be important to clarify the roles and specific obligations of organisations.

The Australian Government proposes to allocate responsibility for implementing the guardrails to developers and deployers of AI systems. It is recognised that the same organisation can perform both roles depending on the context and that there are challenges in defining developers and deployers. These definitions are especially relevant when trying to clearly delineate the extent of changes to an AI model that would result in a deployer being considered a developer.

Below are draft definitions for developers and deployers. Feedback is sought on these definitions, including the extent to which a deployer should be allowed to adapt an AI model without being considered a developer. It is noted the EU AI Act states that where a deployer or other third-party has substantially modified a high-risk AI system or made it available under their name or trademark, they will be deemed a developer.[71]

Table 2: Definitions for developers and deployers of AI systems.

Developer	Organisations or individuals who design, build, train, adapt or combine AI models and applications.
Deployer	Any individual or organisation that supplies or uses an AI system to provide a product or service. Deployment can be used for internal purposes or used externally impacting others, such as customers or individuals, who are not deployers of the system.

While the proposed mandatory guardrails do not specifically apply to end-users[72], users will still need to meet any legal obligations under existing laws. In addition, developers and deployers will need to consider who their end-users are. This includes how they may use or misuse the AI system in assessing and mitigating risks (see Guardrail 2, for example).

[71] EU AI Act, Article 3(4).

[72] As set out in the glossary, 'End users' refers to any intended or actual individual or organisation that consumes an AI-based product or service, interacts with it or is impacted by it after it is deployed.

Various contextual factors are important when considering how accountability will be apportioned across the AI supply chain and throughout the AI lifecycle, and how specific measures under the proposed mandatory guardrails will be fulfilled. Attachment E gives a high-level description of how the proposed mandatory guardrails could apply to developers and deployers, which feedback is sought on.

The guardrails should be distributed according to which actors are best equipped to address risks associated with a particular stage of development. This takes into consideration an actor's access to critical information such as training data, and their ability to effectively intervene and change an AI system.

For example, deployers may not have the insight or resources to assess or mitigate important risks associated with training and testing an AI model. Therefore, developers must take responsibility for mitigating certain risks upstream, especially where the deployer has limited control over the model itself.

The method by which AI models become available to deployers also affects the level of control that different actors have. For example, if a model is distributed through an application programming interface (API), the model is run remotely and controlled by the developers. On the other hand, if a model is released publicly or 'open source', a deployer can adapt the model for their own purposes.[73] Figure 3 provides an example of the complex network of actors across the AI supply chain and the respective roles of developers and deployers from a model's inception to its deployment to the end-user. It is not intended to be a comprehensive representation of the entire AI supply chain in all contexts.

The availability of AI models released as open source may give rise to additional considerations in determining the best way to apply guardrails to these models' developers or deployers. Further feedback is sought on this issue including:

- ways to appropriately define what models are open-source to avoid issues such as 'open-washing' to evade regulations[74]
- any obligations that should attach to developers and deployers of these models.

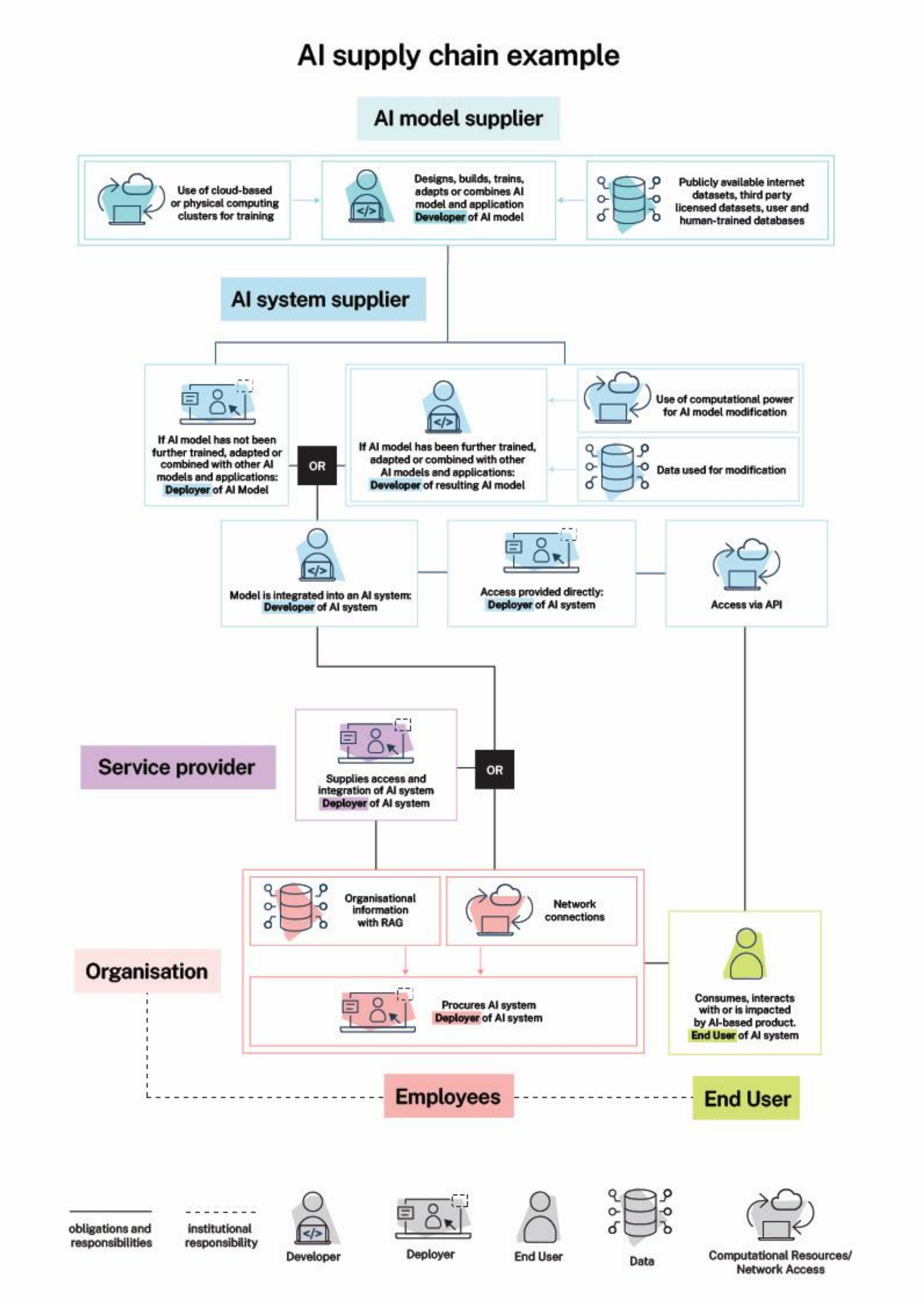
Australia will also need to monitor how the EU is developing its regulatory approach on issues of accountability and liability.[75] This approach will likely shape how developers and deployers operate in other jurisdictions and how they are held accountable. Regardless of any mandatory guardrails the Australian Government may decide to implement, existing laws will need to be strengthened and clarified to help resolve certain issues of liability. These issues are likely to arise in areas such as corporations law, product safety law, tort law and contract law.

[73] S Kuspert, N Moes, and C Dunlop, '*The value chain of general-purpose AI*', Ada Lovelace Institute, 10 February 2023.

[74] Rethinking open source generative AI: open-washing and the EU AI Act https://facctconference.org/static/papers24/facct24-120.pdf.

[75] On 28 September 2022 the European Commission made a Proposal for an Artificial Intelligence Liability Directive which would set out rules for non-contractual civil liability for damage caused by the involvement of an AI system.

Figure 3: AI supply chain example


AI supply chain example
AI model supplier
Use of cloud-based or physical computing clusters for training
Designs, builds, trains, adapts or combines AI model and application Developer of AI model
Publicly available internet datasets, third party licensed datasets, user and human-trained databases
AI system supplier
If AI model has not been further trained, adapted or combined with other AI models and applications: Deployer of AI Model
OR
If AI model has been further trained, adapted or combined with other AI models and applications: Developer of resulting AI model
Use of computational power for AI model modification
Data used for modification
Model is integrated into an AI system: Developer of AI system
Access provided directly: Deployer of AI system
Access via API
Service provider
Supplies access and integration of AI system Deployer of AI system
OR
Organisation
Organisational information with RAG
Network connections
Procures AI system Deployer of AI system
Consumes, interacts with or is impacted by AI-based product. End User of AI system
Employees
End User
obligations and responsibilities
institutional responsibility
Developer
Deployer
End User
Data
Computational Resources/ Network Access

Proposed mandatory guardrails

The proposed guardrails have been developed based on a review of other comparable jurisdictions' approaches and feedback from experts and stakeholders to date. It is proposed that the guardrails would apply to the use of AI in high-risk settings, and GPAI models. These guardrails will need to be adaptable to remain fit for purpose as AI technology continues to evolve and become more autonomous.

For the proposed mandatory guardrails to be effective, it is critical that AI developers and deployers engage openly with stakeholders across the AI supply chain and the AI lifecycle. This engagement should begin at the earliest possible stage and should seek to identify and document which key stakeholder groups may be impacted by use of the system. Open engagement with affected stakeholders will assist in identifying potential harms of the system, such as bias, discrimination or accessibility issues. Guardrail 2 below provides further guidance on the development of a risk management framework to identify these potential harms, while Guardrail 8 offers guidance on transparency requirements for organisations involved in the development and operation of high-risk AI systems.

To support businesses to develop and deploy AI safely and responsibly while the Australian Government considers options for mandatory guardrails, a Voluntary AI Safety Standard has been developed. It gives practical guidance to all Australian organisations on how to safely and responsibly use and innovate with AI, consistent with the proposed mandatory guardrails. The Standard also emphasises the importance of open and effective engagement across the AI supply chain and AI lifecycle.

Proposed mandatory guardrails for high-risk AI at a glance

Organisations developing or deploying high-risk AI systems are required to:

1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance
2. Establish and implement a risk management process to identify and mitigate risks
3. Protect AI systems, and implement data governance measures to manage data quality and provenance
4. Test AI models and systems to evaluate model performance and monitor the system once deployed
5. Enable human control or intervention in an AI system to achieve meaningful human oversight
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content
7. Establish processes for people impacted by AI systems to challenge use or outcomes
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks
9. Keep and maintain records to allow third parties to assess compliance with guardrails
10. Undertake conformity assessments to demonstrate and certify compliance with the guardrails

Guardrail 1: Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.

Organisations developing or deploying a high-risk AI system must create an accountability process outlining governance policies and clear roles to ensure compliance with the guardrails. This guardrail aligns with the Canadian Artificial Intelligence and Data Act (AIDA) requirement for organisations to maintain accountability frameworks[76] and the EU AI Act's requirement for providers to establish a quality

[76] Amendments to AIDA, s. 12.

management system.[77] This obligation should extend to publishing where this is in the public interest. For example, where the AI system or GPAI model has been or is being deployed on a wide scale.

Organisations must also make their accountability processes publicly available and accessible to improve public confidence in AI products and services.

Consistent with Canada's requirements, accountability processes will cover:

- a documented approach to regulatory compliance
- policies for data and risk management
- clear roles, responsibilities and reporting structures for staff (including contractors and third-party providers)
- details of the training organisations make available to staff, especially those responsible for overseeing the high-risk AI system once deployed.

While people might have specific roles and responsibilities, accountability for adhering to guardrails will sit at the organisational level. Internal accountability processes should also consider the size of the business, the specific risks of harm and the nuances of different lifecycle dynamics. Lifecycles for GPAI models will continue to evolve and have implications for how to distribute accountability.

Guardrail 2: Establish and implement a risk management process to identify and mitigate risks.

Organisations must establish, implement and maintain risk management processes to address risks arising from a high-risk AI system. This guardrail's aim is to eliminate or reduce the likelihood of any known or foreseeable risks. The risk management process is not just a technical risk assessment. It is crucial that developers and deployers consider any potential impacts on people, community groups and society before the high-risk AI system is in use. Where elimination is not possible, organisations must implement strategies to contain or mitigate any residual risks.

An effective risk management process will include the following:

- a process for identifying risks using the high-risk principles and assessing the impacts of these risks
- identification and application of mitigation measures[78]
- mechanisms to identify new risks and monitor the effectiveness of risk mitigation measures.

Implementation of this guardrail could be guided by AI risk management standards such as ISO/ISE 42001, which outlines a methodology for conducting AI impact assessments. Also relevant is AS ISO/IEC 23894:2023 *(Information Technology - Artificial Intelligence - Guidance on risk management)*, which helps organisations integrate AI risk management with other internal risk management processes. As well as considering process or management standards, other standards related to metrics or product/service criteria may include explicit methodologies especially useful for small-to-medium sized businesses.

Organisations must choose risk mitigation strategies appropriate for the AI system. These mitigations must be proportionate to the severity and likelihood of the risk. The type of risk mitigation will differ based on the organisation's role in the AI supply chain or AI lifecycle and the circumstances. Risk mitigation could include a duty to consider and monitor for unintended consequences as they arise. It could also involve identifying such severe risks that they ought to reconsider whether to deploy the AI system or model in the desired use case. For example, GPAI developers should take responsibility for addressing

[77] EU AI Act, Article 17.

[78] See for example Generative AI – position statement | eSafety Commissioner which sets out a range of potential mitigation measures in relation to online safety risks and harms.

risks against all foreseeable use cases by their clients. Their clients who are considered deployers should assess the risk of harm for their contained use case.

The *International Scientific Interim Report* also highlighted it was important to consider who is involved in the risk management process to identify and assess the high-priority risks. This means risk assessment requires experts from multiple domains, as well as representatives from impacted communities.[79] Consideration should also be given to including workers when assessing the impacts of adopting AI in the workplace.

Organisations should plan and carry out risk management processes across the lifecycle of a high-risk AI system. This includes downstream deployers, where specific deployment contexts might unintentionally introduce new risks. Deployers should give feedback to developers and contribute to the risk management discussion and design (see Guardrail 8). Ongoing testing, human oversight and transparency guardrails also support this guardrail. Documentation supporting risk management will need to be available or provided to relevant authorities or organisations across the lifecycle of an AI system or model (see Guardrail 8 and 9).

Guardrail 3: Protect AI systems, and implement data governance measures to manage data quality and provenance.

Organisations must ensure they have appropriate data governance, privacy and cybersecurity measures in place. Data quality directly impacts the performance and reliability of an AI model, and datasets can contain biases that may lead to discriminatory outputs. Therefore, any data used to train, fine-tune or test a model must be fit for purpose and representative. To minimise security vulnerabilities, organisations need to securely store and manage data and protect it from unauthorised access and exploitation.

Data must also be legally obtained. Datasets used to train AI systems or GPAI models must not contain illegal and harmful material such as child sexual abuse material or non-consensual intimate imagery. Data sources must be disclosed.

The EU AI Act requires GPAI model providers to implement policies to comply with EU copyright law and make publicly available a summary of the content used for training the model.[80] The Attorney-General's Department is leading Australia's approach on copyright and AI in consultation with the Copyright and AI Reference Group. This work includes considering the intersection of this proposed mandatory guardrail and copyright laws.

Consistent with ISO/IEC 42001 and the EU AI Act[81], this guardrail will cover:

- the origin and legality of the dataset and collection processes
- documentation of data provenance
- consideration of the principles of Indigenous Data Sovereignty and ICIP
- assessments of data quality and whether data is fit for purpose
- approaches to data preparation processes, including labelling and categorisation methods
- identification of biases in the dataset and methods to mitigate such biases
- cybersecurity measures put in place.

This guardrail is intended to complement requirements under other legislation, such as:

- the *Privacy Act 1988*, which places obligations on organisations handling personal information

[79] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 68.

[80] EU AI Act, Article 53.

[81] EU AI Act, Article 10.

- the *Copyright Act 1968*, which gives owners of certain material exclusive economic rights that include the right to copy and the right to communicate the material to the public
- the *Security of Critical Infrastructure Act 2018*, which imposes security obligations on data storage and processing assets.

Guardrail 4: Test AI models and systems to evaluate model performance and monitor the system once deployed.

Organisations must test and evaluate the performance of an AI model before placing a high-risk AI system on the market and continuously monitor the system to ensure it operates as expected.

The purpose of testing is to ensure an AI model meets specific, objective and measurable performance metrics, and manage risks associated with the model. The exact methods and metrics used to test an AI model will vary. They will depend on the intended or foreseeable use of the high-risk AI system, and any risks associated with the system. For example, developers of a facial recognition system would be required to test the accuracy of an AI model for different social groups who may interact with an AI system to gauge the potential for discriminatory impacts. Developers of GPAI models must conduct adversarial testing for any emergent or potentially dangerous capabilities.

Ongoing monitoring and evaluation of a high-risk AI system will ensure the system remains fit for purpose. This will detect any unintended consequences or changes in model performance, such as model drift. Organisations must systematically gather and analyse data to assess the behaviour of high-risk AI systems over time. This is consistent with requirements for 'post-market monitoring' in the EU AI Act.[82]

Implementation of this guardrail will be supported by specific measurement methodologies outlined in standards, such as those in *ISO/IEC TR 29119-11:2020 (Software and systems engineering ‐ Software testing ‐ Part 11: Guidelines on the testing of AI-based systems)* and *SA TR ISO/IEC 24027:2022 (Information technology ‐ Artificial Intelligence ‐ Bias in AI systems and AI aided decision making)*. The National Institute of Standards and Technology (NIST) is also developing guidelines and benchmarks for evaluating and auditing AI capabilities, including for red teaming, under the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Guardrail 5: Enable human control or intervention in an AI system to achieve meaningful human oversight.

Organisations must ensure that humans can effectively understand a high-risk AI system, oversee its operation and intervene where necessary across the AI supply chain and throughout the AI lifecycle. Human oversight helps to minimise risks during the deployment phase, particularly those that other guardrails cannot effectively mitigate, and respond to harm.

The purpose of AI is partly to automate certain activities and augment the ability of humans to process information. This means that real-time human involvement in an AI system may not always be practical, and may even make a system less reliable. In such instances, developers should design the system so that a human can review its operations and outputs and reverse a decision if necessary. Requiring human oversight will help address the risks associated with AI's speed, scale and increasing autonomy, especially where it is applied to assist or replace human decision-making.

To achieve meaningful human oversight, the person responsible for overseeing the AI system while it is in use must be able to access and interpret information regarding the system's outputs. Those responsible for oversight must be sufficiently qualified to interpret output and understand the core capabilities and limitations of an AI model to accurately assess the accuracy of algorithmic advice.

[82] EU AI Act, Article 72.

Guardrail 6: Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.

Organisations will be required to inform end-users on how AI is being used and where it affects them. Organisations should communicate this in a clear, accessible and relevant manner. As well as fostering the accountability of organisations deploying high-risk AI systems, mandatory transparency measures will build public trust in AI systems and their outputs. It is also important that people are made aware that they are interacting with AI so that they can seek redress if necessary (see Guardrail 7).

This guardrail will entail 3 distinct requirements:

1. Organisations must inform people when AI is used to make or inform decisions relevant to them.

The first requirement is consistent with the EU AI Act[83]. This emphasises the important role deployers have in informing people when they are subject to decisions made using a high-risk AI system. This requirement applies to AI systems which are used to make decisions impacting people's health, safety and human rights. This requirement intersects with proposed reforms to the *Privacy Act 1988* to enhance transparency about the use of personal information in automated decisions which have a legal or similarly significant effect on their rights.

2. Organisations must inform people when they are directly interacting with an AI system.
3. Organisations must apply best efforts to ensure AI-generated outputs, including synthetic text, image, audio or video content, can be detected as artificially generated or manipulated.

These requirements have been introduced by the EU AI Act[84] and Canada's AIDA[85] in response to the specific risks of impersonation and manipulation that systems intended to interact with people or generate content can pose. These requirements are primarily relevant to GPAI models. However, in both jurisdictions they apply to all AI systems, including low-risk uses of AI which do not meet the definition of high-risk AI systems.

Adequate methods for ensuring that the public can identify AI-generated outputs are still in development. These methods include content labelling and watermarking techniques, which involve embedding a unique marker into the output of an AI model that can be algorithmically detected. While these methods are still in development and even state-of-the art techniques carry technical drawbacks,[86] Australia is proposing a 'best efforts' approach. This will require organisations to apply the best methods available according to their own assessments and with reference to relevant standards. This threshold is similar to the approach Canada has proposed.

Implementation of this guardrail will evolve as AI detection and labelling techniques become more advanced. It will be informed by the development of international standards and codes of practice (see below).

The development of International Digital Content Provenance Standards

The Coalition for Content Provenance (C2PA)[87] has developed a standard to guide the development and use of provenance technologies. These technologies allow consumers to identify the origin and use of AI in the creation of digital content.

As well as providing a standard for the use of these technologies, C2PA has developed their own provenance technology called Content Credentials. The code for this is freely available online for anyone to use. Content Credentials allows creators to embed certain information into a digital file. This can include their name, how the content was created, and whether and how AI was used in the process. An

[83] EU AI Act, Article 50.

[84] EU AI Act, Article 50.

[85] AIDA, s. 7(1)(g).

[86] European Parliament, 'Generative AI and watermarking', European Government, 2023.

[87] Coalition for Content Provenance and Authenticity, '*Overview*', C2PA, 2023.

icon appears in the image or beside the file which consumers can click to learn about the content's origin. The technology also allows creators to embed their media with an imperceptible watermark which links back to the original Content Credentials if they were later removed. Adobe has also created a free public website called Verify that allows people to view the full history of any digital content that uses Content Credentials.

Previous attempts to combat AI-generated mis- and disinformation, such as deepfakes and fake news, have tended to focus on automated detection. However, this allows the lie to spread before it is found out. Content Credentials aims to offer a more proactive approach, giving consumers the ability to assess the trustworthiness of digital content as they interact with it.

The Content Authenticity Initiative (CAI) has been established in order to promote the C2PA standard and the use of Content Credentials.

Guardrail 7: Establish processes for people impacted by AI systems to challenge use or outcomes.

Organisations will be required to establish processes for people who are negatively impacted by high-risk AI systems to contest AI-enabled decisions or make complaints about their experience or treatment. Having such processes in place will be critical to establishing public trust. This is particularly important given the potential of AI systems to amplify bias on the basis of attributes protected under Australian law, such as race or gender.

These processes will include:

- establishing internal complaint handling functions
- assignment of responsibility for human oversight and intervention where complaints are made
- provision of sufficient information about the use or outcomes of the AI system to impacted parties to make contestability meaningful and effective.

Importantly, this guardrail requires organisations to proactively establish internal avenues for dealing with complaints.

Guardrail 7 closely links to guardrails 5 and 6 in supporting people who are impacted by high-risk AI systems. It supports them to challenge these systems' use and outcomes, and gives them opportunities to exercise their legal rights. Obligations under this Guardrail will need to work alongside existing avenues for complaints handling. These include existing rights and obligations under Australian Consumer Law and administrative law.

Guardrail 8: Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks.

Organisations will be required to share information about a high-risk AI system with other actors involved in the development and operation of the system. Where multiple organisations are responsible for implementing guardrails, some actors may lack critical information about a high-risk AI system and how it works. This is a problem that is exacerbated by the opacity and limited explainability of the most advanced AI models. Therefore, transparency across the AI supply chain will help organisations meet their legal obligations and enable them to effectively identify and mitigate risks. For example, there may be obligations for infrastructure providers to either share data, communicate the risks and limitations of what is on their platform or share responsibility for controls to mitigate risks where they are better placed to control risks.

Developers must give deployers information about how to use the high-risk AI system and ensure they can understand and interpret outputs. This will allow deployers to respond to risks as they emerge. The information provided must include a description of the characteristics, data sources a model is trained on, key design decisions regarding data training or the AI system or model's development affecting its performance, and the capabilities, limitations and risks of the system. It must also take into account the need to protect commercially sensitive information or trade secrets.

Deployers must report adverse incidents and significant model failures to developers, who can then issue improvements to the model. For example, the developer of an AI image generator would benefit from understanding how an AI model has evolved, to address risks posed by deepfakes at scale. This recognises the important role deployers have in holding developers accountable and identifying risks of AI systems as they learn and adapt.

This guardrail aligns with the EU AI Act's requirement for 'transparency and provision of information to deployers'[88]. It also aligns with Canada's AIDA requirement for developers to prepare and share model cards for machine learning models.[89] The EU AI Act also requires deployers to inform providers if they identify risks while a high-risk AI system is in use.[90]

Guardrail 9: Keep and maintain records to allow third parties to assess compliance with guardrails.

Organisations must keep and maintain a range of records, including technical documentation, about a high-risk AI system over its lifecycle. They must be ready to give these records to relevant authorities on request and for the purpose of undertaking a conformity assessment (see Guardrail 10). For example, this could include the power for regulators to access documents on enforcement and compliance in their regulatory remit. Record keeping ensures organisations are accountable for showing their compliance with the guardrails. It also enables external scrutiny of high-risk AI systems. Where appropriate these records can form the basis of transparency with other actors in the AI supply chain (see Guardrail 8).

At a high-level and in keeping with the EU AI Act[91], these records will include:

- a general description of the AI system
- design specifications from the development phase, including testing methodology and results
- a description of datasets used and their provenance

[88] EU AI Act, Article 13.

[89] Amendments to AIDA, s. 9(1)(c). Model Cards aim to provide information about the trained machine learning model. This can include intended use, evaluation data, training data, performance metrics, ethical considerations as well as any relevant model specific details such as type, version or licensing info.

[90] EU AI Act, Article 26.

[91] EU AI Act, Article 12.

- assessment of human oversight measures
- a detailed description of the capabilities and limitations of the AI system
- the risk management processes and mitigation measures implemented.

Any organisation training large state-of-the-art GPAI models with potentially dangerous emergent capabilities must disclose these 'training runs' to the Australian Government. This requirement recognises that the underlying capability and compute power to train powerful AI models can be a significant predictor of risk. This is consistent with the US Executive Order. Under the EU AI Act, providers of GPAI models with systemic risk must also report relevant information about serious incidents to the AI Office.[92]

Record keeping can be a major compliance burden on organisations, especially small and medium-sized enterprises (SMEs). Feedback is sought on ways to reduce burden on this group. One example could be for them to complete less detailed documentation under this guardrail, as per the EU approach.

Guardrail 10: Undertake conformity assessments to demonstrate and certify compliance with the guardrails.

Organisations will be required to show that they have adhered to the guardrails for high-risk AI systems by carrying out a conformity assessment. This is an accountability and quality assurance mechanism to verify whether organisations have met their legal obligations prior to deploying a high-risk AI system.

Conformity assessments could be carried out by the developers themselves, by a third-party or by government entities or regulators. Depending on the availability of organisations with the appropriate expertise to conduct assessments, the Australian Government may need to consider different options in the short term vs the long term.

Organisations will need to do a conformity assessment before placing a high-risk AI system on the market. They'll then need to periodically repeat the assessment to ensure continued compliance. However, consistent with the EU AI Act, if a deployer retrains the system or the system undergoes any changes that significantly affect compliance with the guardrails, the organisation will be required to do a new conformity assessment.

The assessment will rely on records captured under Guardrail 9 and an organisation's documented accountability process (Guardrail 1). Once completed, the organisation will attain certification of compliance which they can communicate to the public to show they are adhering to the above guardrails.

[92] EU AI Act, Article 53.

Regulatory options to mandate guardrails

Questions for consultation

13. Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach?

14. Are there any additional limitations of options outlined in this section which the Australian Government should consider?

15. Which regulatory option/s will best ensure that guardrails for high-risk AI can adapt and respond to step-changes in technology?

16. Where do you see the greatest risks of gaps or inconsistencies with Australia's existing laws for the development and deployment of AI? Which regulatory option best addresses this, and why?

This section considers regulatory options available to the Australian Government to mandate the proposed mandatory guardrails. It considers to what extent existing regulatory frameworks address the risks of AI outlined in this paper, and to what extent there is a need for a new whole-of-economy framework. Options include:

1. A domain specific approach – Adopting the guardrails within existing regulatory frameworks as needed

2. A framework approach – Introducing new framework legislation to adapt existing regulatory frameworks across the economy

3. A whole of economy approach – Introducing a new cross-economy AI-specific Act (for example, an Australian AI Act).

Option 1 represents reform of existing regulatory frameworks to implement the guardrails on a sector-by-sector basis (or domain specific approach), while options 2 and 3 represent a whole-of-economy approach specifically targeted to organisations developing and using AI technologies. Option 3 provides the definitions, thresholds and guardrails in one piece of legislation, as well as targeted mechanisms for enforcement and monitoring. Option 2 sits between these, as a whole-of-economy approach to reform. It will provide definitions, thresholds and guardrails in one legislative instrument, but relies on amendments to existing regulatory frameworks to accommodate enforcement via existing regulators.

Separate to these options and as discussed under section 2 in this paper, the Australian Government will continue to strengthen and clarify existing laws so it is clearer how they apply to AI systems and models. This work will occur alongside new regulatory approaches introducing mandatory guardrails for AI developers and deployers.

Current regulatory arrangements

Feedback received on the discussion paper highlighted areas where Australia's current regulatory system is not fit for purpose to support the safe and responsible use of AI in high-risk settings. Examples of gaps and uncertainties in existing laws are discussed in section 4.

Australia has many laws that are impacted by and shape the development of AI. This includes economy-wide laws, like copyright and privacy laws, and sector-specific laws, for example for medical devices, at both Commonwealth and state and territory levels. Different sectors are at varying stages of maturity in considering how to apply existing regulatory frameworks to respond to the unique challenges of AI. For example, heavily regulated industries like the financial sector already have a range of obligations that support the safe and responsible use of AI.

In some instances, regulators have already moved to define how existing regulatory frameworks could apply to AI. This includes recent guidance from the Therapeutic Goods Administration (TGA) about software-based medical devices that may include large language models.[93] Another example is the eSafety Commissioner's registration of the Search Engine Services Code and the recently registered Designated Internet Services industry standard. The standard imposes obligations for certain generative AI services to address illegal and restricted online material such as pro-terror and child sexual exploitation material. However, in many cases, regulators are still considering how their laws apply to AI so they can give guidance to regulated entities.

Case study – Codes and standards under the *Online Safety Act 2021*

The *Online Safety Act 2021*, which commenced on January 2022, provides for industry bodies to develop codes for 8 key sections of the online industry. The Act intends to address, minimise and prevent harms relating to illegal and restricted online material.[94] A wide range of services are covered, including social media services, messaging services, search engines and other websites/apps (known as designated internet services). If an industry code meets the statutory requirements, the eSafety Commissioner (eSafety) can register it. If a code does not meet the requirements, then eSafety can develop an industry standard instead.

Search Engine Services Code

eSafety registered an industry code for internet search engine services (SES Code) in September 2023, after requesting changes to reflect and capture recent advancements in generative AI functionality. This mandatory and enforceable code came into force in March 2024.

The SES Code applies to internet search engine services for end-users in Australia. It applies to any features integrated into the search functionality and the user interface of an internet search engine service, whether enabled by AI or otherwise.

The SES Code sets out specific obligations around generative AI for the most seriously harmful 'class 1' synthetic material (for example, terrorism or child sexual exploitation, including deepfake material). This includes requiring services to:

- take appropriate steps to improve systems, processes and/or technologies concerning certain class 1 synthetic materials generated by AI that may be accessible via an internet search engine service
- take appropriate steps to ensure that AI-enabled search engines do not return search results that contain child sexual abuse material, and
- where relevant, implement measures to make clear when a user is interacting with any features using AI.

[93] The *Therapeutic Goods Administration Act 1989*, administered by the TGA, includes a risk-based approach to the regulation of medical devices in part aimed at ensuring such devices are safe and work as intended. The September 2023 guidance issued by the TGA emphasised the importance of those that develop software that constitutes a 'medical device' to 'understand and demonstrate the sources and quality of text inputs used to train and test the model, and in clinical studies, in addition to showing how the data is relevant and appropriate for use on Australian populations.'

[94] Specifically, the Act regulates Class 1 and Class 2 illegal and restricted material. Class 1 (1A and 1B) material is material that is or would likely be refused classification under the National Classification Scheme. It includes child sexual exploitation material, pro-terror content, and extreme crime and violence. Class 2 material includes content that is likely to be classified as X18+ or R18+. This includes non-violent sexual activity, or anything that is "unsuitable for a minor to see."

Designated Internet Services Industry Standard

Following a period of public consultation, eSafety registered the Designated Internet Services (DIS) standard in June 2024. The standard is intended to come into effect in December 2024. DIS is a catch-all category to cover internet services not captured in other defined categories like search or social media.

The DIS standard takes a technology neutral approach to implementation of the compliance measures. It places specific obligations on the platforms which distribute open-source machine learning models and on the highest risk consumer-facing generative AI services.

This standard focuses on online content. Like the SES Code, it puts in place compliance measures to prevent, minimise or address harms from the most seriously harmful forms of class 1 material. The suite of measures includes:

- risk assessments
- having and enforcing policies
- facilitating user reporting
- efforts to filter out illegal material from prompts and outputs and to prevent it from being generated by users
- transparency reporting requirements.

These measures have a distinct application to a small subset of high-risk generative AI services. They are vital to preventing the generation and distribution of certain synthetic class 1 material, such as terrorism or child sexual exploitation material.

eSafety continues to work with the Department of Industry, Science and Resources to ensure alignment across workstreams, which have different timescales and focus areas.

Current regulatory arrangements for AI will continue to change even if new mandatory guardrails for AI were not introduced. For example, different regulators and agencies would continue to implement changes for their sectors. However, it is likely that without coordinated and holistic government action, simply allowing the status quo to continue could encounter the following challenges:

1. Potential gaps in regulatory coverage

Different issues and potential harms can occur at different points across the AI supply chain and throughout the AI lifecycle, from development and deployment through to use. However, many of Australia's existing laws put obligations on those who use AI models in their specific business applications (that is, AI deployers) rather than those developing the technology.

Existing regulatory regimes are also limited to their defined scope and the entities they regulate. This means that obligations on the use of AI in high-risk settings that apply under one regulator or legislative framework won't necessarily apply under another (i.e. enforcement is different across different regulatory regimes).

Australia's current regulatory frameworks also don't consistently provide the regulatory levers to respond to the speed and scale of risks, focusing on addressing harms after they occur. Preventative (or pre-market) guardrails would shape the development of AI technologies and how they can be deployed. Greater flexibility and more preventative/pre-market guardrails will better enable responses to economy-wide AI harms which can scale rapidly across sectors and settings. Relying on harm identification and regulatory intervention sector-by-sector is already contributing to fragmentation, gaps and inconsistency of protections for consumers and citizens.

The government's interim response on Safe and Responsible AI observed that existing laws do not have the level of specificity needed for effective AI risk management systems to address a broad range of high-risk settings. To achieve more proactive approaches to harm mitigation in some high-risk settings, existing legislation would likely need to be amended. While some sectors have more detailed risk management obligations than others, this is inconsistent across the economy.

2. Consistency within but not across sectors

Existing legislative frameworks vary in preventative or pre-market measures that could apply to AI. This differs across the statute book and depends on whether a particular use case of AI falls into a particular sector's definition of regulated activities.

Thresholds for action and penalties vary across different frameworks. This creates inconsistencies where developers and deployers are working across multiple sectors and laws. It can also cause regulatory uncertainty where it is not clear which regulator is responsible for an action.

Options for introducing mandatory guardrails

There are several ways that the Australian Government could implement guardrails ensuring greater testing, transparency and accountability of AI in high-risk settings to address the limitations of the current system. This includes implementing guardrails by:

- Option 1 – a domain specific approach – adapting existing regulatory frameworks to include the guardrails
- Option 2 – a framework approach – introducing framework legislation that will require other existing laws to be amended for the framework legislation to have effect
- Option 3 – a whole of economy approach – introducing a new cross-economy AI Act.

This section outlines the benefits and limitations of implementing the guardrails under 3 options.

Option 1 (a domain specific approach): Adapting existing regulatory frameworks

Under this option, a review of each relevant piece of legislation would be required to address gaps and to embed relevant guardrails in existing regulatory frameworks to address the risks of AI outlined in this paper. Government agencies who were conducting reviews of their laws would consider and update their laws to embed guardrails where they do not currently exist for their domain-specific regulatory regimes. Regulators could also give guidance to support implementation of the guardrails into their regulatory frameworks. This could include advice on how they will apply to their specific laws and to their regulated entities.

The Australian Government could also explore other non-legislative options to create consistency across different laws. This could include using the Office of Parliamentary Counsel's drafting directions or instruments like the Attorney-General Department's 'Guide to Framing Commonwealth Offences, Infringement Notices and Enforcement Powers.'

This option is similar to the approach to regulating AI announced by the then Sunak Government in the UK in February 2024. Following the election of the Starmer Government in July 2024, His Majesty King Charles III announced on 17 July 2024 that his government will seek to establish appropriate legislation to place requirements on those working to develop the most powerful AI models.[95]

[95] UK Government, The King's Speech 2024, 17 July 2024.

The Sunak Government's 'pro-innovation approach to AI regulation'

In March 2023, the UK released a white paper 'AI regulation: a pro-innovation approach – policy proposal' for consultation.[96] The whitepaper proposed 5 cross-sectional principles for existing regulators to interpret and apply in their remit to create safe, responsible AI innovation. These were:

- safety, security and robustness
- appropriate transparency and accountability
- fairness
- accountability and governance
- contestability and redress.

The focus of the approach was on supporting regulators through guidance and additional resourcing. However, in its recent response to its whitepaper consultations the UK also indicated that the 'government is considering introducing targeted binding requirements on developers of highly capable GPAI systems which may involve creating or allocating new regulatory powers'.[97]

Analysis of option 1

A benefit of this option is that it would leverage Australia's existing suite of economy-wide and sector-specific laws. Businesses, people and regulators are already familiar with these laws. This limits the disruption to those regulated and the broader regulatory ecosystem. It recognises that Australia has a mature legislative and regulatory framework that may be leveraged responding to new risks posed by AI.

Other benefits of this approach, compared to introducing new whole-of-economy legislation, are that:

- it limits the risk of regulatory duplication and an associated compliance burden where regulated entities operate in a single sector
- it enables an incremental approach to new regulation, learning as regulation is developed
- it lets existing regulatory frameworks address harms in their specific contexts.

A limitation of this approach is that it is likely to exacerbate gaps and inconsistencies within the current regulatory system. Different agencies and regulators could adopt the proposed guardrails requiring greater testing, transparency and accountability in high-risk settings in different ways. This could lead to definitional inconsistencies in how AI systems are regulated across the economy, particularly in contexts where AI applications have multiple use cases in different contexts. Some regulators may also decide not to prioritise reforms to address AI issues. This may be because of competing regulatory priorities, lack of resources or lack of technical capability.

Implementation of option 1 would be bound by existing law and the enforcement powers of existing regulators. This means that pre-market guardrails could only be introduced for those sectors or domains where existing regulatory frameworks already allow for pre-market obligations (e.g. therapeutic goods).

This option may also be a slower way of achieving regulatory coverage across the economy compared to a whole-of-economy reform. For example, individual policy agencies would need to identify specific gaps where the proposed mandatory guardrails do not already exist before pursuing amendments to legislation or regulators giving further guidance. Where separate legislative processes are needed to amend individual legislative frameworks, this may lead to inefficiencies because of lengthy legislative processes. There is also the potential for cumulative regulatory burden where consultation occurs on multiple legislative frameworks at the same time.

[96] Department for Science, Innovation and Technology, 'AI regulation: a pro-innovation approach', Department for Science, Innovation and Technology, 2023.

[97] UK Government, '*A pro-innovation approach to AI regulation: government response*', Government of UK, 6 February 2024.

The effect of the limitations described above is that option 1 is likely to lead to regulatory siloes, with different requirements placed on different sectors over time. This is expected to increase the compliance burden for organisations that work across multiple sectors and regulated activities. It also preserves the risk of harms occurring in the gaps between regulated sectors and activities. To mitigate these issues, option 1 would require a strong coordination function from government, supported by a clear statement of expectations, to ensure the guardrails were applied in a similar way across the whole economy. To comprehensively introduce pre-market guardrails, further legislative reform may also be required in addition to option 1 – for example, through also implementing elements of options 2 or option 3.

Option 2 (a framework approach): Adapting existing regulatory frameworks through framework legislation

Framework legislation is broadly defined and can take many different forms. It can seek to create an enabling environment for decision-making or actions to be taken by the government, in contrast to legislation that prescribes specific requirements or solutions.[98]

The *Regulatory Powers (Standard Provisions) Act 2014* is an example of framework legislation that provides:

- standard regulatory terminology
- for the exercise of standard regulatory powers (such as monitoring, investigation and enforcement) under other legislation
- standard obligations on regulators in the exercise of regulatory powers.

Under option 2, government would create new framework legislation that defines the guardrails to apply and the threshold for when they would apply. This framework legislation would provide a consistent set of definitions and measures that would then be implemented through amendments to existing regulatory frameworks.

To implement the guardrails, the framework legislation could set out the proposed principles for defining high-risk settings or a list of high-risk use cases and the proposed mandatory guardrails to apply. Such provisions would only be enforceable once they are 'activated', such as when other laws are amended to refer to these provisions. Similar to option 1, framework legislation would be confined to the scope and powers of existing laws that are amended to refer to it. This cross-reference could be made in other principal or subordinate laws or through guidance from regulators issued under those laws (this latter approach would be non-binding). The provisions included in framework legislation would represent best practice which policy agencies and regulators are expected to build into their own laws. It is expected they would do so unless there are compelling policy reasons for them to take a different approach or tailor it for their regulated activities.

Responsibility for enforcing these new provisions under the framework legislation once they are referred to in other laws would fall to the existing regulator for each regime, relying on existing penalties in each regime.

Analysis of option 2

Similar to option 1, a benefit of option 2 is that it takes advantage of the familiarity that businesses, people and regulators already have with Australia's existing regulatory regimes. This approach also recognises that Australia's existing laws may be capable of giving effect, or can be amended to give effect to one or more of the proposed mandatory guardrails in individual regimes. However, the framework legislation would still be limited by the scope and powers of existing laws that are amended to refer to it.

[98] RS Magnusson, 'Framework legislation for non-communicable diseases: and for the Sustainable Development Goals?', *BMJ Global Health* 2017 DOI:*10.1136/bmjgh-2017-000385*; E Garrett, 'The Purposes of Framework Legislation', *USC Public Policy Research Paper Series*, 3 February 2004, No 04-3, *doi.org/10.2139/ssrn.504783*.

Taking a framework legislation approach would help to support a more consistent approach to reform across the economy. This includes:

- providing a consolidated source of concepts on AI, which can be centrally maintained and amended as necessary in the framework legislation
- supporting a consistent approach across regulatory frameworks – notably, regarding definitions of high-risk settings, the types of guardrails that could be applied. This could reduce regulatory burdens on organisations being regulated particularly for those developing or deploying GPAI and could find themselves regulated under multiple regulatory frameworks and by multiple regulators
- facilitate interoperability with approaches of other countries (e.g. by providing consistent definitions and requirements that align with international approaches) to better enable Australian companies desiring to be part of AI supply chains
- providing coverage where there may be gaps across regulatory frameworks to avoid situations where organisations can capitalise on loopholes or gaps across laws (commonly referred to as regulatory arbitrage).

Framework legislation would set a stronger signalling of the Australian Government's regulatory expectations. Option 2 allows existing regulatory frameworks to be amended to better address concerns about AI pursuant to a consistent set of high-risk definitions and guardrails.

A limitation of option 2 is that while it provides consistency in definitions, it retains gaps across regimes, and is limited to the scope and powers of current regulatory arrangements on AI. This could result in differences or gaps on enforcement and obligations not being able to be applied upstream to developers due to the limited scope and powers of existing regulatory frameworks. A framework legislation approach would take time to provide coverage as it would require cross-agency coordination and agreement to progress other agencies' law reform agendas at the same or a similar time. For example, should government agree to introduce framework legislation, it would be appropriate to identify existing regulatory regimes that could be amended at the same time as the development of the framework legislation.

Option 3 (a whole of economy approach): Introducing a new AI-specific Act

The third option is to introduce a new AI-specific Act to implement the proposed mandatory guardrails for AI in high-risk settings. The Act would define these high-risk applications of AI and outline the new mandatory guardrails. It would establish a monitoring and enforcement regime overseen by an independent AI regulator. The new regime would work alongside existing regulators to oversee the guardrails where there are gaps in existing approaches, informed by a mapping exercise to minimise duplication and ensure there is clarity across different regulators. This approach would be modelled on the Canadian AIDA which makes carve outs for sectors where existing laws already have the relevant guardrails.

Unlike framework legislation outlined above, a new AI-specific Act would have enforceable provisions. It is a more centralised approach to creating legislative change, as it imposes stand-alone obligations. In contrast, framework legislation would need other laws to link to the framework. As a result, an independent AI regulator may be required to oversee and enforce the legislation where there are gaps. The independent AI regulator could be either a new regulator or an existing regulator with expanded powers. Both the EU and Canada have taken the approach of developing new AI-specific legislation. Option 3 is more closely modelled on the Canadian approach, which takes into account Australia's existing regulatory frameworks.

International examples of regulating AI through AI-specific legislation

European Union

The European Parliament adopted the Artificial Intelligence Act in March 2024, and the European Council followed with its approval in May 2024. The final Act was signed into law on 13 June 2024. At the time of writing this paper, all procedures had been completed, and the Act was awaiting final publication in the Official Journal.99

The Act will unify how AI is regulated across the single market of the 27 EU Member States. It relies on a double-layered enforcement mechanism with the EU AI Office in the EU Commission at the centre of enforcement, and Member States establishing or appointing market surveillance and notifying authorities to implement the rules under the AI Act. The Act classifies AI according to risk. It introduces a range of measures related to testing, transparency and accountability for the regulation of high-risk AI systems and GPAI systems including those with systemic risks.

The Act will be largely applicable 24 months after entry into force, but some parts will be applicable sooner. These will include:

- the ban of AI systems posing unacceptable risks 6 months after the entry into force
- codes of practice will apply 9 months after entry into force
- rules on general-purpose AI systems that need to comply with transparency requirements will apply 12 months after the entry into force

High-risk systems will have more time to comply with the requirements as the obligations concerning them will become applicable 36 months after the entry into force.

Canada

The draft Artificial Intelligence and Data Act (AIDA)[100] was introduced as part of the Digital Charter Implementation Act 2022. It would set the foundation for the responsible development and deployment of AI systems in Canada. The Act classifies AI according to risk and introduces several obligations for high-impact and general-purpose systems including testing, transparency and accountability measures. The new regulator proposed under the AIDA would have a role not only in administering the operative provisions in the AIDA, but also in working with and supporting other regulators in their engagement with AI.

The Canadian model acknowledges the existing legal and regulatory frameworks which apply to the use of AI. It also acknowledges that the AIDA is designed to build on existing principles to address regulatory gaps arising from developments in AI. This includes giving consideration to the degree in which classes of AI are adequately regulated under other laws in defining high-impact systems.

Developing an AI-specific Act, like the development of framework legislation in Option 2 which envisages identifying laws to be updated, requires consideration of interaction with existing laws. It is likely that any new AI-specific Act will involve significant interaction with existing laws. A new Act would need to identify and empower a new or existing regulator with the necessary resources, capabilities, and capacity to carry out this role.

To manage these interactions, any new AI Act would be designed to establish a baseline for guardrails in high-risk settings across the economy. Where these may overlap with obligations under existing regulatory frameworks, the Australian Government will consider approaches to enable the domain-specific law's obligations to take supremacy over the AI Act where the laws meets (or exceeds) the standards required by an AI-specific Act. In practice, Commonwealth agencies and regulators could agree on any carve outs and ensure responsibilities of regulators are clear. Obligations could also be

[99] European Parliament, Legislative Observatory, 'Artificial Intelligence Act Procedure File', European Parliament, accessed 5 July 2024.

[100] AIDA.

tailored for specific contexts to meet national security or law enforcement needs should the Australian Government agree to these exceptions.

On the responsibility of regulators, existing regulators would continue carrying out their functions and role. Any new obligations or responsibilities created under the new AI-specific Act, would be managed either by a new AI regulator or expanding the responsibilities of existing regulators.

By designing the AI Act so that domain-specific laws take precedence where they already achieve the same function, this creates incentives for regulators and industry to work together to clarify the application of the domain-specific laws for their sectors.

Canada's AIDA Bill is an example of how a standalone Act interacts with existing regulatory frameworks. It specifically excludes AI in medical devices given existing legislation already regulates these. In their proposed concept of 'high-risk', it also excludes AI systems that are already subject to similar regulatory oversight. If a new AI-specific Act is developed in Australia, it could be designed recognise areas where regulatory frameworks already provide similar levels of protection. For example, like Canada, an AI Act could exclude regulated entities already subject to the use of software-based medical devices under legislation administered by the TGA.

Analysis of option 3

This option provides greater consistency across the economy than option 1 or option 2. A new stand-alone AI Act would provide consistent definitions of high-risk settings and guardrails, full coverage to avoid gaps across regulatory regimes and extend regulatory obligations to upstream developers via a new preventative or pre-market regime. A key difference is that option 3 would also provide consistency in enforcement.

This option has the following benefits:

- provides clear and consistent expectations on those developing and deploying AI across the economy without introducing duplicative or conflicting obligations. Developers would be covered because the new AI-specific Act would be set up with a new scope and powers to impose pre-market obligations on developers
- has regulatory efficiency, by taking a whole-of-economy legislative approach instead of amending the suite of Australia's existing laws
- better enabling interoperability with international approaches taken by the EU and Canada
- enables an independent AI regulator with responsibility for preventative guardrails to support the development of regulatory expertise on AI. This expertise could then be shared with and leveraged by other regulators.

Limitations of this option include:

- the potential of added complexity and duplicate obligations with existing legislative frameworks, unless mitigated through good legislative design that includes carve outs or reflects agreement between regulators
- further consequential regulatory coordination challenges across regulators, unless mitigated through good legislative design that minimises duplication, as above
- needing to identify or establish the appropriate regulator to enforce guardrails where they do not fit in the remit of a current regulator, which would take resources and time.

How to get involved

Your contributions are welcomed as the Australian Government considers regulatory responses to mitigate the potential risks of AI, and to increase public trust and confidence in its development and use.

Consultation questions are included below to guide contributions. However, feedback is welcome on any aspect of the paper. Please submit your answers at the consult.industry.gov.au/ai-mandatory-guardrails.

Attachment A: Glossary

Term	Proposed Definition
Accountability	The state of being answerable for actions, decisions and performance within a well-defined scope of responsibility and potentially sanctionable.
Agentic AI (also, ‘AI agent,’ or ‘AI autonomous agent’)	AI that is capable of accomplishing multi-step tasks in pursuit of a high-level goal with little or no human oversight.[101]
AI lifecycle	All events and processes that relate to an AI system's lifespan. This spans from inception to decommissioning, including its design, research, model development, training, deployment, integration, operation, maintenance, sale, use, and governance.
AI supply chain	The complex network of actors and organisations that enable the use and supply of AI throughout the AI lifecycle from model design, testing and fine tuning to deployment and integration into the local IT system.
Artificial intelligence (AI) system	A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.[102]
Artificial intelligence (AI) system and model	AI model: the raw, mathematical essence that is often the ‘engine’ of AI applications. AI system: for the purposes of differentiating an AI system from an AI model, the AI system is the ensemble of several components, including one or more AI models, that is designed to be particularly useful to humans in some way. Applying the above, for example, the ChatGPT app is an AI system. Its core engine, GPT-4, is an AI model.[103]

[101] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 87.

[102] Organisation for Economic Co-operation and Development (OECD), *What is AI? Can you make a clear distinction between AI and non-AI systems?*, OECD.AI Policy Observatory website, 2024, accessed 8 May 2024.

[103] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024, page 16.

Term	Proposed Definition
Deployer	Any individual or organisation that supplies or uses an AI system to provide a product or service. Deployment can be for internal purposes, or used externally impacting others, such as customers or individuals.
Developer	Organisations or individuals who design, build, train, adapt, or combine AI models and applications.
End User	Any intended or actual individual or organisation that consumes an AI-based product or service, interacts with it or is impacted by it after it is deployed.
FLOPS	Floating point operations per second (FLOPS) is a measure of computer performance. In the context of AI it provides insight into the computational capabilities and performance thresholds of AI systems.
General-purpose AI (GPAI) model	An AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems.[104]
Generative AI model	An AI model with the capability of learning to generate content such as images, text, and other media with similar properties to its training data.
Labelling	*Labelling:* a procedure that enables organisations to put their information classification scheme into practice by attaching classification labels to relevant information assets.[105] *AI Content Labelling:* applying visible content warnings to alert stakeholders to the presence of AI-generated content and its provenance.
Large language model (LLM)	A large language model (LLM) is a type of generative AI that specialises in the generation of human-like text.
Machine learning model	A mathematical construct that generates an inference or prediction based on input data or information.

[104] Based on but not identical to the definition proposed by the Parliament of Canada, Amendments to Bill C-27 (44-1), Minister of Innovation, Science and Industry, 28 November 2023.

[105] Adapted from International Organization for Standardisation, 'Information security, cybersecurity and privacy protection – Information security controls,' ISO-IEC Standard no. 27002:2022.

Term	Proposed Definition
Narrow AI system	A type of AI system or model that is focused on defined tasks and uses to address a specific problem. Unlike GPAI models, these types of AI systems cannot be used for a broader range of problems without being re-designed.[106]
Testing	The process of executing an AI model or system to verify and validate that it exhibits expected behaviours across a set of appropriately selected test cases.[107]
Transparency	A property of a system that appropriate information about the system is made available to relevant stakeholders. Appropriate information for system transparency can include aspects such as features, performance, limitations, components, procedures, measures, design goals, design choices and assumptions, (training) data sources and testing and evaluation methodologies, benchmarks, test cases and criteria.[108]
Watermarking	Information embedded into digital content, either perceptibly or imperceptibly by humans, that can serve a variety of purposes, such as establishing digital content provenance or informing stakeholders that the contents are AI-generated or significantly modified.[109]

[106] Based on but not identical to the definition from International Organization for Standardisation, '*Artificial intelligence concepts and terminology*', ISO-IEC Standard no. 22989:2022. The standard can be freely accessed here: https://standards.iso.org/ittf/PubliclyAvailableStandards/.

[107] IEEE Computer Society, 'SWEBOK v4: Guide to the Software Engineering Body of Knowledge (public draft),' 2024.

[108] International Organization for Standardisation, 'Artificial intelligence concepts and terminology', ISO-IEC Standard no. 22989:2022. The standard can be freely accessed here: https://standards.iso.org/ittf/PubliclyAvailableStandards/.

[109] Coalition for Content Provenance and Authenticity, '*C2PA Technical Specification*', C2PA Specifications, C2PA, 2023.

Attachment B: Actions by government to support safe and responsible AI

1. Delivering regulatory clarity and certainty

Industry and civil society have been clear that they need clarity and certainty on how AI will be regulated to fully realise the opportunities that AI presents.[110] The initial steps the government is taking to deliver regulatory certainty on AI include:

- considering the establishment of mandatory guardrails for high-risk AI as outlined in this proposals paper
- strengthening and clarifying existing regulation applying to the use of AI.

In the 2024–25 Budget, the Australian Government committed to reviewing and strengthening existing regulations in the areas of health care, consumer and copyright law:

- The Department of Health and Aged Care will conduct a gap analysis of health specific legislative instruments, helping to ensure that AI in health care is clinically safe, ethical and improves equity of health outcomes
- The Treasury will conduct a review of the Australian Consumer Law and investigate how effectively the existing framework may be positioned to respond to business conduct issues arising from the use of AI
- The Attorney General’s Department (AGD) will deliver an intensified program of engagement through the Copyright and AI Reference group to consider how Australia’s copyright law deals with AI.

Other relevant reforms the Australian Government is undertaking include:

- work led by AGD to develop a whole of government legal framework to support use of automated decision-making systems (ADM) for delivery of government services. This may include systems run by AI. This reform work implements the Australian Government’s response to recommendation 17.1 of the Robodebt Royal Commission
- reforms being progressed by AGD to the *Privacy Act 1988*, following the Privacy Act Review. The Australian Government has agreed to progress a range of proposals that would strengthen privacy protections, transparency and accountability. This includes in relation to the handling of personal information by AI
- supporting regulators by issuing statements of expectations on best practice approaches to regulating AI incorporating the insights from this consultation process.[111]

[110] Tech Council of Australia, ‘Supporting safe and responsible AI: Tech Council of Australia submission’, Tech Council, August 2023.

[111] Ministerial Statements of Expectations are issued by responsible Ministers to regulators within their portfolios to provide greater clarity about government policies and objectives relevant to the regulator in line with its statutory objectives, and the priorities the Minister expects it to observe in conducting its operations. Further information on such statements can be found in the Department of Finance’s Guidance Note.

2. Supporting and promoting best-practice governance

The Australian Government acknowledges the importance of supporting industry to uplift governance skills and capabilities to innovate responsibly with AI. It is progressing a range of initiatives, including developing a Voluntary AI Safety Standard for organisations using AI to provide organisation-level and system-level best practice for developers and deployers.

3. Supporting AI capability

The Australian Government is committed to realising the benefits of AI. The Australian Government's investments in the 2024–25 Budget build on investments in the 2023–24 Budget, including:

- $17 million for the AI Adopt Program to establish 3 to 5 centres to provide direct support to SMEs
- $34.5 million continues for the Next Generation Artificial Intelligence and Emerging Technologies Graduates programs.

There are also a range of other initiatives the Australian Government has put in place to support AI industries, including:

- $1 billion in the National Reconstruction Fund for enabling capabilities
- Support through the Industry Growth Program, which supports innovative small and medium-sized organisations by providing commercialisation and growth advice.

The Australian Government will continue to explore opportunities to invest and build AI capability in Australia.

4. Government as an exemplar

The Australian Government recognises the importance of leadership in the use of AI and acting as an exemplar in best-practice approaches to AI governance.

In August 2023, the Australian Government established the AI in Government Taskforce to examine the use and governance of AI by the Australian Public Service. The Taskforce concluded on 30 June 2024. While the Taskforce has concluded, the Digital Transformation Agency (DTA) will continue to develop and implement policies to position government as an exemplar in the use of AI.

On 15 August 2024, the DTA released the policy for the responsible use of AI in government to position government as an exemplar in the use of AI. The Policy outlines mandatory actions required by agencies including identifying an accountable official for AI and publishing a transparency statement about their use of these technologies. The DTA will also soon pilot a draft Commonwealth AI Assurance Framework to support a more consistent approach by agencies to assessing and mitigating the risks of AI use.

Other relevant policies Australian Government agencies are undertaking include:

- The Australian Federal Police (AFP) is committed to implementing the Australia New Zealand Policing Advisory Agency's AI Principles, whereby information about the use of AI systems is made publicly available to the greatest extent possible, without undermining policing objectives. It is also working collaboratively with academia and industry to use AI for policing purposes effectively and responsibly such as through its partnership with Monash University's Artificial Intelligence for Law Enforcement and Community Safety (AiLECS) lab.
- The Australian Government is developing a framework in relation to the use of Automated Decision Making (ADM), which would include but is not limited to ADM systems involving AI. For example, where government uses AI to make administrative decisions, these uses must comply with administrative law principles. This includes providing reasons for a decision, or demonstrating that irrelevant considerations were not taken into account in making a decision.

AI systems operate across all sectors and jurisdictions of the Australian economy, and states and territories have an important role to play in ensuring their safe and responsible governance.

5. International engagement on AI governance

The Australian Government engages internationally to:

- share our regulatory approaches and best-practice policies
- secure efficient and effective international AI governance structures
- learn from other countries' experience.

The Australian Government seeks to ensure that our domestic regulations are interoperable where appropriate and that our interests are advanced and protected in international agreements and principles.

The Australian Government is working with likeminded partners to promote and align regulatory approaches and best-practice policies. Specifically, the Australian Government is:

- strengthening and promoting interoperability between Australia's domestic regulation of AI and data and those of international partners through targeted bilateral and multilateral engagements, and by shaping and participating in key international agreements on AI such as the G7 Hiroshima AI Process and the Seoul Declaration for Safe, Innovative and Inclusive AI
- identifying and proactively responding to AI safety risks of shared concern. Australia works with industry, experts and partners, including through the international AI Safety research network and by contributing to the *International Scientific Report on Advanced AI Safety*, through Dr. Bronwyn Fox, Australia's representative on the Expert Advisory Panel[112]
- working with and through the international AI Safety research network to share research and evaluate risk in AI models and systems.

[112] Department for Science, Innovation and Technology and AI Safety Institute, *International Scientific Report on Safety of Advanced AI: Interim Report*, UK Government, 2024.

Attachment C: Australia’s approach to AI regulation

Australia’s AI regulatory strategy will be developed alongside the government’s consideration of options for mandatory guardrails. It will outline how the range of activities underway fit together to support the safe and responsible use of AI. These include including the development of mandatory guardrails, the Voluntary AI Safety Standard and strengthening existing laws. It is a supporting framework and will not impede the government’s consideration of options for mandatory guardrails.

Core elements that are expected to be included in the AI regulatory strategy to support cohesion across related activities. The proposed core elements included are in the table below.

Key elements of whole-of-government AI regulatory strategy	
Setting clear regulatory objectives and principles	Such as: • Protect people from harm • Innovation for economic benefit • Societal wellbeing
Clarifying the regulatory approach	Such as: • The application of a risk-based approach • Defining harms, specifying the types and levels of AI risk • Grounding in human rights
Ensuring the regulatory approach is fit-for-purpose for Australia’s unique settings	Having regard to: • AI as an identified critical technology impacting Australia’s national interest, including economic prosperity, national security and social cohesion • Australia’s economic interests and comparative advantage
Ensuring consistency in regulatory approach across government	Ensuring alignment in approach and clarifying obligations across the public and private sector including: • Establishment of mandatory guardrails • Strengthening and clarifying existing laws • Voluntary AI Safety Standard • Government as an exemplar
Promoting clarity and consistency of legal terms and rules	Such as • Improving consistency of legal terms and cross-sectoral clarity of existing laws that apply to the development, deployment and use of AI • Clarifying legal liability • Clarifying interaction with state and territory laws

Key elements of whole-of-government AI regulatory strategy	
Clarifying the application of existing and new laws	• Clarify and strengthen existing laws (such as privacy, consumer protection, IP) • Clarify the interaction between existing laws and any new mandatory guardrails • Clarify the role of sectoral regulators
Clarifying the purpose of hard law and soft law measures	Considering the role of: • Hard law in setting minimum expectations for the conduct of developers and deployers of AI systems • Soft law in providing added guidance and supporting transition of industry governance practice
Specifying the approach to stakeholder consultation across regulatory initiatives	Including: • Genuine, ongoing consultation with industry, unions, civil society and academia • Inclusive approaches to engaging potentially impacted communities and their representatives across civil society • Ongoing independent expert advice including legal, policy and technical knowledge
Determining the settings necessary to facilitate an effective and responsive regulatory ecosystem	Considering: • Regulator capabilities, resources and powers necessary to enforce the law • Practical guidance to ensure legal and ethical responsibilities are understood and upheld • Education and resources to support the exercise of individual rights • Support for SMEs to understand and meet obligations
Having regard to international alignment and interoperability	Ensuring Australia's regulatory approach has regard to: • International alignment and interoperability, especially for Australian firms operating across borders • International normative grounding of AI laws in human rights

Attachment D: Examples of high-risk or -impact uses and settings from other jurisdictions

Domain	European Union (see Annex III to the Act)	Canada (draft Artificial Intelligence and Data Act, schedule 2)
Biometrics	Remote biometric identification systems excluding AI systems intended to be used for biometric verification whose sole purpose is to confirm that a specific natural person is the person he or she claims to be. AI systems intended to be used for biometric categorisation, according to sensitive or protected attributes or characteristics based on the inference of those attributes or characteristics. AI systems intended to be used for emotion recognition.	The use of an artificial intelligence system to process biometric information in matters relating to: • the identification of an individual, other than in cases in which the biometric information is processed with the individual's consent to authenticate their identity; or • the assessment of an individual's behaviour or state of mind
Critical Infrastructure	AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.	
Education/ Training	AI systems intended to be used to: • determine access or admission to educational and vocational training institutions • evaluate learning outcomes, including when those outcomes are used to steer the learning process • assessing the appropriate level of education that individual will receive or will be able to access, in the context of/within education and vocational training institution • monitoring and detecting prohibited behaviour of students during tests in the context of/within education and vocational training institutions	

Domain	European Union (see Annex III to the Act)	Canada (draft Artificial Intelligence and Data Act, schedule 2)
Employment	AI systems intended to be used: • for recruitment or selection, notably to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates • to make decisions affecting terms of the work related relationships, promotion and termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics and to monitor and evaluate performance and behaviour of persons in such relationships.	The use of an artificial intelligence system in matters relating to determinations in respect of employment, including recruitment, referral, hiring, remuneration, promotion, training, apprenticeship, transfer or termination.
Access to essential private and public services	AI systems intended to: • be used by or for public authorities to evaluate the eligibility of a person for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services • be used to evaluate the creditworthiness of an individual or establish their credit score , with the exception of AI systems used for the purpose of detecting financial fraud • be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance • evaluate and classify emergency calls.	The use of an artificial intelligence system in matters relating to a. the determination of whether to provide services to an individual b. the determination of the type or cost of services to be provided to an individual c. the prioritization of the services to be provided to individuals. The use of an artificial intelligence system in matters relating to health care or emergency services, excluding a use referred to in any of paragraphs (a) to (e) of the definition device in section 2 of the Food and Drugs Act that is in relation to humans.
Law enforcement	AI systems intended to be used by or on behalf of law enforcement authorities: • to assess the risk of an individual to become a victim of criminal offences • as polygraphs and similar tools	The use of an artificial intelligence system to assist a peace officer, as defined in section 2 of the Criminal Code, in the exercise and performance of their law enforcement powers, duties and functions

Domain	European Union (see Annex III to the Act)	Canada (draft Artificial Intelligence and Data Act, schedule 2)
	• to evaluate the reliability of evidence in the course of investigation or prosecution of criminal offences • for assessing the risk of an individual of offending or re-offending • for profiling of individuals in the course of detection, investigation or prosecution of criminal offence	
Migration/Border Protection	AI systems intended to be used by public authorities: • as polygraphs and similar tools • to assess a risk, including a security risk, a risk of irregular migration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State • for the examination of applications for asylum, visa and residence permits and associated complaints • for the purpose of detecting, recognising or identifying individuals with the exception of verification of travel documents.	
Administration of justice and democratic processes	AI systems intended to be used: • by a judicial authority in researching and interpreting facts and the law and in applying the law • influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referendums.	The use of an artificial intelligence system by a court or administrative body in making a determination in respect of an individual who is a party to proceedings before the court or administrative body.
Online content		The use of an artificial intelligence system in matters relating to:

Domain	European Union (see Annex III to the Act)	Canada (draft Artificial Intelligence and Data Act, schedule 2)
		• the moderation of content that is found on an online communications platform, including a search engine or social media service; or • the prioritization of the presentation of such content.

Attachment E: Applying the guardrails

Guardrail	Developer	Deployer
1. **Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.**	Developers must establish, implement and publish accountability processes, including clearly outlining governance policies and responsibilities. This includes clear reporting structures and roles. They must outline a strategy for regulatory compliance and document details of training provided to staff members.	Deployers must establish, implement and publish accountability processes, including details of who is responsible for overseeing and continuously monitoring the AI system once it is deployed.
2. **Establish and implement a risk management process to identify and mitigate risks.**	Developers must identify and assess specific risks based on the principles for classifying high-risk AI settings and establish risk mitigation strategies appropriate to the context of use. GPAI developers are required to manage underlying risks inherent to the model, such as biases or security vulnerabilities, and any known or foreseeable risks that may arise.	Deployers will be responsible for following instructions for use set by developers and managing risks specific to the use case. Deployers are also responsible for ongoing oversight and observation of the AI system to identify unforeseen risks.
3. **Protect AI systems, and implement data governance measures to manage data quality and provenance.**	Developers must evaluate the quality, track the provenance of and keep records of the data used to train and test high-risk AI systems. They must also ensure data is securely stored and managed.	Deployers who input additional data to a model must comply with this guardrail by evaluating and safeguarding any data used.

Guardrail	Developer	Deployer
4. Test AI models and systems to evaluate model performance and monitor the system once deployed.	Developers must test and evaluate the performance of an AI model by using metrics relevant to the intended purpose or purposes of the model. This could include adhering to standards of accuracy, robustness, cybersecurity or fairness. Developers of GPAI models would be required to use adversarial testing to identify any dangerous capabilities or emergent properties of their models, and undertake evaluations of model weights, training data and model architecture. Developers also have an ongoing obligation to monitor and refine their AI system once deployed based on feedback and data provided by deployers where appropriate.	Deployers must monitor the operation of the AI system based on the information provided under Guardrail 8 for new risks.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight.	Developers of high-risk AI systems must establish measures to ensure that humans are able to exercise oversight at the deployment phase. This involves ensuring that humans are able to effectively understand the system, oversee its operation and intervene where necessary.	Deployers must equip people overseeing the system with the skills to understand the capabilities and limitations of the model and correctly interpret and assess the quality of its output.
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.	Developers must embed measures into an AI system to enable AI-generated content to be labelled as such once deployed, wherever technically feasible.	Organisations must inform individuals when AI is used to make decisions about them and when they are directly interacting with an AI system. They must also apply best efforts to label content as AI-generated using technical methods such as watermarking.

Guardrail	Developer	Deployer
7. Establish processes for people impacted by AI systems to challenge use or outcomes.		Deployers are to establish internal organisational avenues for people negatively affected by AI systems to raise concerns or complaints.
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks.	Developers must help deployers use and interpret an AI system appropriately by sharing relevant information. They must summarise the capabilities, risks and limitations of their models so that downstream deployers and end-users can implement appropriate risk management measures.	Deployers must inform developers of any adverse incidents that occur or risks that emerge while the system is in use.
9. Keep and maintain records to allow third parties to assess compliance with guardrails.	Developers must keep comprehensive technical documentation regarding data, model training and development and testing. They must also keep records regarding design decisions and of measures built into the AI system to enable human oversight and labelling of AI-generated content.	Deployers must keep records relevant to the system's deployment and use.
10. Undertake conformity assessments to demonstrate and certify compliance with guardrails.	Developers must undertake a conformity assessment before the AI system is deployed.	If the AI system is retrained or undergoes any changes that affects compliance with the guardrails, a new conformity assessment may be required.