Australian Government
Department of Industry, Science and Resources

National Artificial Intelligence Centre

CSIRO

Voluntary AI Safety Standard

August 2024

industry.gov.au/NAIC

Copyright

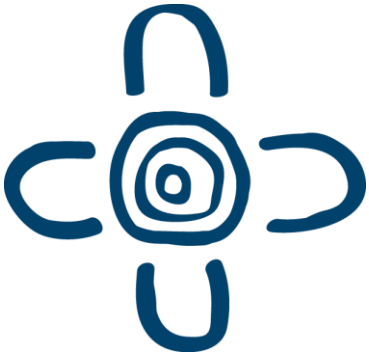
Disclaimer

The purpose of this publication is to provide best practice guidance on implementing safe and responsible AI practices for Australian organisations.

The Commonwealth as represented by the Department of Industry, Science and Resources has exercised due care and skill in the preparation and compilation of the information in this publication.

The Commonwealth does not guarantee the accuracy, reliability or completeness of the information contained in this publication. Interested parties should make their own independent inquires and obtain their own independent professional advice prior to relying on, or making any decisions in relation to, the information provided in this publication.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this publication. This publication does not indicate commitment by the Commonwealth to a particular course of action.



Acknowledgement of Country

Our department recognises the First Peoples of this Nation and their ongoing cultural and spiritual connections to the lands, waters, seas, skies, and communities.

We Acknowledge First Nations Peoples as the Traditional Custodians and Lore Keepers of the oldest living culture and pay respects to their Elders past and present. We extend that respect to all First Nations Peoples.

Introduction

The Voluntary AI Safety Standard gives practical guidance to all Australian organisations on how to safely and responsibly use and innovate with artificial intelligence (AI). Through the Safe and Responsible AI agenda, the Australian Government is acting to ensure that the development and deployment of AI systems in Australia in legitimate but high-risk settings is safe and can be relied on, while ensuring the use of AI in low-risk settings can continue to flourish largely unimpeded.

In 2023, the government underwent consultation through its discussion paper on 'Safe and Responsible AI in Australia'. In the Interim Response areas of government action were outlined, including:

- delivering regulatory clarity and certainty
- supporting and promoting best practice for safety
- ensuring government is an exemplar in the use of AI
- engaging internationally on how to govern AI.

The response also recognised the need to consider building AI capability in Australia.

To support and promote best practice, an immediate action was to work in close consultation with industry to develop a Voluntary AI Safety Standard. This standard complements the broader Safe and Responsible AI agenda, including developing options on mandatory guardrails for those developing and deploying AI in Australia in high-risk settings.

While there are examples of good practice through Australia, approaches are inconsistent. This is causing confusion for organisations and making it difficult for them to understand what they need to do to develop and use AI in a safe and responsible way. The standard establishes a consistent practice for organisations. It also sets expectations for what future legislation may look like as the government considers its options on mandatory guardrails.

The standard consists of 10 voluntary guardrails that apply to all organisations throughout the AI supply chain. They include testing, transparency and accountability requirements across the supply chain. They also explain what developers and deployers of AI systems must do to comply with the guardrails. The guardrails help organisations to benefit from AI while mitigating and managing the risks that AI may pose to organisations, people and groups.

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. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.

The first 9 voluntary guardrails have been aligned closely with proposed mandatory guardrails, with the exception of the 10th voluntary guardrail, which emphasises the importance of ongoing engagement with stakeholders to evaluate their needs and circumstances. Conformity assessments, proposed in the

10th mandatory guardrail, are being prepared for in the Voluntary AI Safety Standard through several voluntary steps organisations can be taking now to improve their record keeping, transparency and testing approaches.

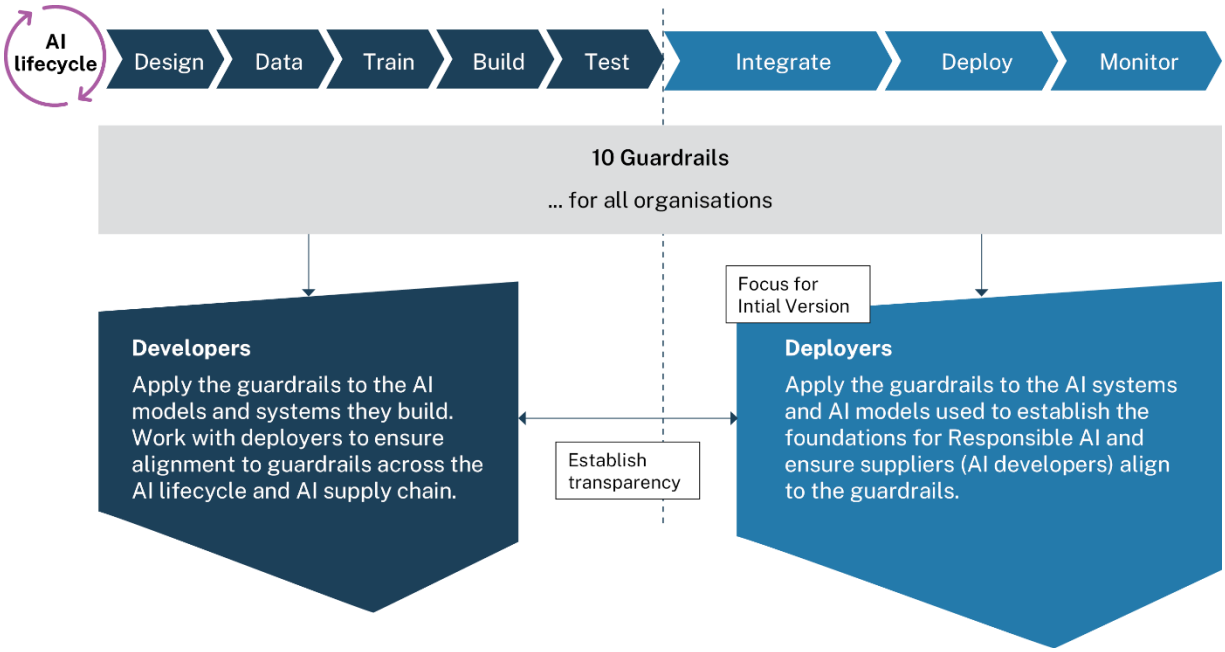


Figure 1: Application of the guardrails

An AI deployer is an individual or organisation that supplies or uses an AI system to provide a product or service. Deployment can be internal to the business or external. When deployment is external it can impact others, such as customers or other people, who are not deployers of the system.

While the first version of the standard applies to both AI deployers and AI developers, it focuses on providing guidance at the organisational and system level for AI deployers. This reflects feedback received while we developed the standard. We heard that deployers, which are the majority of businesses in the Australian ecosystem who are using AI, had the greatest need for guidance on how to adopt best practice.

Focusing on deployers also supports them to work with developers on the practices needed to support the safe and responsible use of AI across the supply chain. We will include the additional, more complex guidance for AI developers in the next version of the standard.

To aid deployers of AI systems, the 10 guardrails include procurement guidance. This will ensure AI suppliers and developers are aligning to the guardrails through contractual agreements.

The guardrails on a page

Guardrails	
1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.	Guardrail one creates the foundation for your organisation's use of AI. Set up the required accountability processes to guide your organisation's safe and responsible use of AI, including: • an overall owner for AI use • an AI strategy • any training your organisation will need.
2. Establish and implement a risk management process to identify and mitigate risks.	Set up a risk management process that assesses the AI impact and risk based on how you use the AI system. Begin with the full range of potential harms with information from a stakeholder impact assessment (guardrail 10). You must complete risk assessments on an ongoing basis to ensure the risk mitigations are effective
3. Protect AI systems, and implement data governance measures to manage data quality and provenance.	You must have appropriate data governance, privacy and cybersecurity measures in place to appropriately manage and protect AI systems. These will differ depending on use case and risk profile, but organisations must account for the unique characteristics of AI systems such as: • data quality • data provenance • cyber vulnerabilities.
4. Test AI models and systems to evaluate model performance and monitor the system once deployed.	Thoroughly test AI systems and AI models before deployment, and then monitor for potential behaviour changes or unintended consequences. You should perform these tests according to your clearly defined acceptance criteria that consider your risk and impact assessment.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight.	It is critical to enable human control or intervention mechanisms as needed across the AI system lifecycle. AI systems are generally made up of multiple components supplied by different parties in the supply chain. Meaningful human oversight will let you intervene if you need to and reduce the potential for unintended consequences and harms.
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.	Create trust with users. Give people, society and other organisations confidence that you are using AI safely and responsibly. Disclose when you use AI, its role and when you are generating content using AI. Disclosure can occur in many ways. It is up to the organisation to identify the most appropriate mechanism based on the use case, stakeholders and technology used.

Guardrails	
7. Establish processes for people impacted by AI systems to challenge use or outcomes	Organisations must provide processes for users, organisations, people and society impacted by AI systems to challenge how they are using AI and contest decisions, outcomes or interactions that involve AI.
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks	Organisations must provide information to other organisations across the AI supply chain so they can: • understand the components used including data, models and systems • understand how it was built • understand and manage the risk of the use of the AI system.
9. Keep and maintain records to allow third parties to assess compliance with guardrails.	Organisations must maintain records to show that they have adopted and are complying with the guardrails. This includes maintaining an AI inventory and consistent AI system documentation.
10. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.	It is critical for organisations to identify and engage with stakeholders over the life of the AI system. This helps organisations to identify potential harms and understand if there are any potential or real unintended consequences from the use of AI. Deployers must identify potential bias, minimise negative effects of unwanted bias, ensure accessibility and remove ethical prejudices from the AI solution or component.

Contents

Guide to this standard

The Voluntary AI Safety Standard has 4 parts.

1. **Part 1** explains the purpose of the standard and its value to organisations. It gives context and grounding for the standard through a human-centred lens.
2. **Part 2** describes the foundational concepts of risks and harms and the legal context for AI systems.
3. **Part 3** presents the core content of the standard as a set of 10 guardrails.
4. **Part 4** gives examples to help organisations understand why and how they might adopt the standard.

Part 1: Introducing the first version of Australia's AI Safety Standard

We designed Australia's first voluntary AI safety standard to help organisations develop and deploy AI systems in Australia safely and reliably. Adopting AI and automation is projected to contribute $170 billion to $600 billion of GDP. Australian organisations and the Australian economy can gain significant benefits if they can capture this.[1]

The standard offers a set of voluntary *guardrails* to establish consistent practices for organisations to adopt AI in a safe and responsible way. This is in line with current and evolving legal and regulatory obligations and public expectations. While this standard applies to all organisations across the AI supply chain, this first version of the standard focuses more closely on *organisations that deploy AI systems*. The next version will expand on technical practices and guidance for AI developers.

Definitions

Safe and responsible AI: AI should be designed, developed, deployed and used in a way that is safe. Its use should be human-centred, trustworthy and responsible. AI systems should be developed and used in a way that provides benefits while minimising the risk of negative impact to people, groups, and wider society.

AI deployer: An individual or organisation that supplies or uses an AI system to provide a product or service. Deployment can be internal to an organisation, or external and impacting others, such as customers or other people who are not deployers of the system.

AI developer: An organisation or entity that designs, develops, tests and provides AI technologies such as AI models and components.

AI user: An entity that uses or relies on an AI system. This entity can range from an organisation (such as business, government or not-for-profit), an individual or other system.

Affected stakeholder: An entity impacted by the decisions or behaviours of an AI system, such as an organisation, individual, community or other system.

A complete list of terms and definitions is available in the **terms and definitions**.

While there are already examples of good AI practice in Australia, organisations need clearer guidance. By adopting this standard, organisations will be able to use AI safely and responsibly.

The standard consists of 10 voluntary guardrails that apply to all organisations across the AI supply chain. The voluntary guardrails establish consistent practice to adopt AI in a safe and responsible way. This will give certainty to all organisations about what developers and deployers of AI systems must do to comply with the guardrails.

In the government's January Interim Response to the Safe and Responsible AI discussion paper, the government identified actions to take. These included working with industry to develop this Voluntary AI Safety Standard. This standard sits alongside a broader suite of government actions enabling safe and responsible AI under 5 pillars, outlined in Figure 2. Actions included in the 5 pillars include the Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings, the National Framework for the Assurance of Artificial Intelligence in Government and the Policy for Responsible Use of AI in Government. The standard will continue to evolve alongside the broader activities underway by government to ensure alignment and consistency for safe and responsible AI.

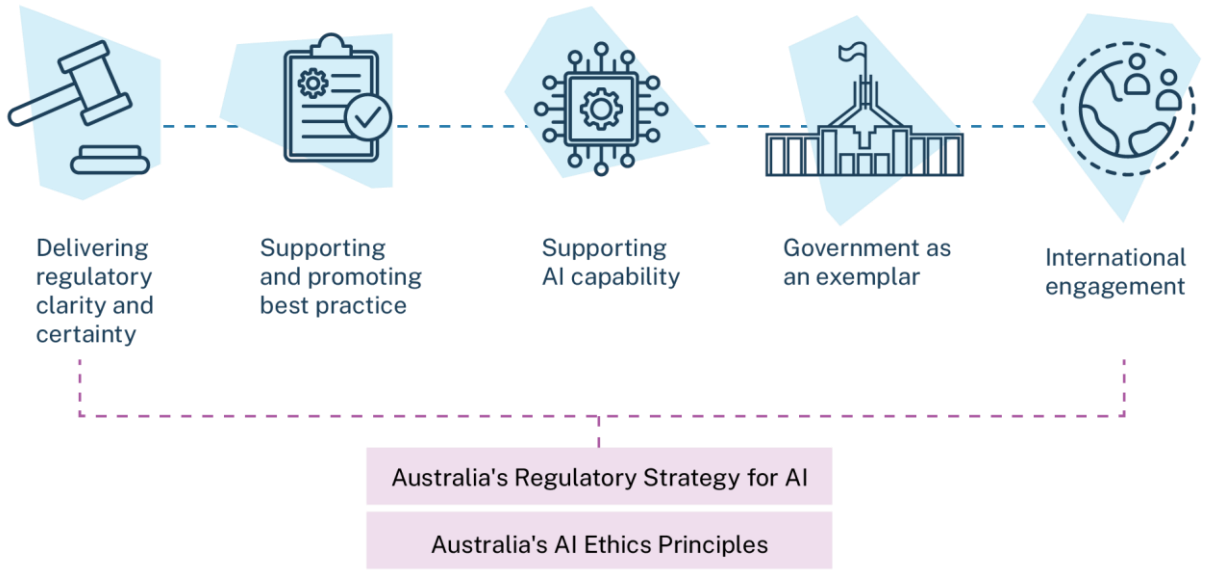


Figure 2: Actions the government is taking to support safe and responsible AI in Australia

Why implement a voluntary standard?

The standard establishes a consistent practice for organisations. It sets expectations for what future legislation may look like as the government considers its options on mandatory guardrails. It also gives organisations the best practice AI governance and ethical practices, which offers them a competitive advantage.

The standard is designed to guide organisations to:

- raise the levels of safe and responsible capability across Australia
- protect people and communities from harms
- avoid reputational and financial risks to their organisations
- increase organisational and community trust and confidence in AI systems, services and products
- align with legal obligations and expectations of the Australian population
- operate more seamlessly in an international economy.

This will lead to the longer-term benefits of improved safety, quality and reliability of AI in Australia. It will support broader use of AI products and services, increased market competition and opportunities for technological innovation.

A human-centred standard

This standard adopts a human-centred approach to AI development and deployment. This is in line with Australia's AI Ethics Principles[2] and Australia's commitment to international declarations such as the Bletchley Declaration.[3] A human-centred approach helps make sure technologies are fit-for -purpose while serving humans, respecting individual rights and protecting marginalised groups.

In the context of safe and responsible AI system usage, a human-centred approach means:

- **Protecting people.** The standard is designed to help leaders and business owners identify, prevent, minimise and remedy a wide range of harms and AI-related risks relevant to their organisation. This is in line with the government's Interim Response. However, its *main purpose* is to protect the safety of people and their rights. A human-centred approach to AI upholds Australia's responsibility to human rights protections. These protections are enshrined in a range of federal and state and territory instruments, the Australian Constitution and the common law.[4]

  See Part 2 for the specific characteristics of AI systems that can amplify existing risks and create new harms for people, organisations, groups or society.

- **Upholding diversity, inclusion and fairness.** The standard is designed to help organisations ensure AI systems serve all people in Australia, regardless of racial background, gender, age, disability status or other attribute.
- **Prioritising people through human-centred design.** Human-centred design is an approach to technology design, development and deployment that recognises and balances human goals, relationships and social contexts with the capabilities and limitations of technical systems.[5] The standard offers practical ways to prioritise the needs of humans in the use of AI systems.
- **Deploying trustworthy AI systems to support social licence.** To unlock the greatest possible value from AI, an organisation deploying it must have social licence for its use. This social licence is based on stakeholders believing in the trustworthiness of the AI system. It is only by earning and maintaining the trust of stakeholders that an organisation can be confident it possesses the social licence needed to deploy AI systems.

Bias

This standard defines bias as the 'systematic difference in the treatment of certain objects, people or groups in comparison to others'. It can be the basis for unfairness, defined as 'unjustified differential treatment that preferentially benefits certain groups more than others'.

For some use cases, such as healthcare, accounting for gender differences can be essential to understand the risk factors or treatment appropriate for an individual or group. This justifies a differential treatment.[6]

Bias becomes problematic or 'unwanted' when it results in *unfavourable* treatment for people or groups. This unfair disadvantage then becomes unlawful discrimination if that treatment is a result of a 'protected attribute':[7]

- age
- disability
- race, including colour, national or ethnic origin or immigrant status
- sex, pregnancy, marital or relationship status, family responsibilities or breastfeeding
- sexual orientation, gender identity or intersex status.

An internationally consistent standard

Recognising that Australia is an open, trading economy, the standard's recommended processes and practices are consistent with current international standards and best practice. This supports Australian organisations who operate internationally by aligning Australian practices with other jurisdictions' expectations. It also aims to avoid creating barriers to international organisations operating in Australia compared to other markets.

The standard draws on and is aligned with a range of international standards. Most important is the leading international standard on AI management systems, AS ISO/IEC 42001:2023, and the US standard on AI risk management, NIST AI RMF 1.0.[8] Each *requirement* in the standard guardrails gives references as to how it is aligned with relevant international and local standards or practices.

Future versions will reflect changes in the international landscape.

Part 2: Foundational concepts for the standard: risks, harms and legal context

AI systems have specific characteristics that amplify risks

AI systems span a wide range of technical approaches. Organisations can use them for many tasks, such as helping with prediction, classification, optimisation or content generation. At their core, AI systems are software-based tools.

AI systems fall broadly into 2 types, each with different strengths and risks:

- **Narrow AI** systems are designed and trained to perform a specific task. Most AI systems in use today fall into this category. These types of systems can perform well in a narrow range of activities, potentially even better than humans, but they cannot perform any other tasks. Examples include chess engines, recommender systems, medical diagnostic systems and facial recognition systems.
- **General-purpose AI** systems are designed and trained to handle a broad range of tasks and are therefore flexible. Their use is not limited to a specific function, so they can be more easily used for purposes their designers may not have considered. Examples include large language models and systems such as Open AI's ChatGPT series.

Both narrow and general-purpose AI systems are *built and operate differently* from traditional software systems. These differences mean that using an AI system for a particular task may amplify existing risks when compared with traditional software.

For example, in traditional software systems, developers explicitly define all the logic governing a system's behaviour. This relies on explicit knowledge, with conscious human engagement at every stage of the software design and development process. Traditional software systems are easier for humans to control, predict and understand.

In contrast, developers of AI systems take a different approach. This often involves defining an objective and constraints, selecting a dataset, and employing a 'machine learning algorithm'. This creates an *AI model* which can achieve the specified objective. While such models often outperform comparable, traditional software systems, the different development approach means AI models are often less transparent, less interpretable, and more complex to test and verify. This amplifies risks and can lead to harm. This is more likely to happen in contexts where it is important to understand and explain how the output was achieved or to constrain the range of potential outputs for safety reasons.

The specific characteristics of general AI systems can amplify risks and harms or pose new risks and harms to an organisation. General AI systems are more prone to unexpected and unwanted behaviour or misuse. This is because of their increased flexibility of interactions, the reduced predictability of their capabilities and behaviour and their reliance on large and diverse training data. For example, large language models can deliberately or inadvertently manipulate or misinform consumers. They can also pose novel intellectual property challenges for both training data and the outputs generated.

The standard supports a risk-based approach to AI harm prevention

As with all software, AI systems vary in the level of risk and the type of harm they pose. Some, like an algorithm on a website that suggests reordering based on stock levels, tend to be lower risk. The potential harms are confined to a customer taking longer to receive a product. Others, like a tool that prioritises job applicants for an interview process or makes financial lending decisions, have far greater potential to create harm. For instance, they may deny a suitable applicant the opportunity of a job or bank loan, or even systematically and unlawfully discriminate against a group of people.

The standard supports a risk-based approach to managing AI systems. It does this by supporting organisations – *starting with AI deployers* – to take proactive steps to identify risks and mitigate the potential for harm posed by the AI systems they deploy, use or rely on.

The standard prioritises *safety* and the *mitigation of harms and risks* to people and their rights.

A human-centred perspective on the harms of AI systems

Organisations should assess the potential for these risks and harms to people:

- **Harm to people.** This includes infringements on personal civil liberties, rights, and physical or psychological safety. It can also include economic impacts, such as lost job opportunities because of algorithmic bias in AI recruitment tools or the unfair denial of services based on automated decision-making.
- **Harm to groups and communities.** AI systems can exacerbate discrimination or unwanted bias against certain sub-groups of the population, including women, people with disability, and people from multicultural backgrounds. This can lead to social inequality, undermining of equality gains and unjust treatment. This is pertinent in recommender algorithms that amplify harmful content.
- **Harm to societal structures.** AI systems' impact on broader societal elements, such as democratic participation or access to education, can be profound. AI systems that spread misinformation could undermine electoral processes, while those that affect educational algorithms could widen the digital divide.

The standard is useful and applicable for identifying, preventing and minimising other risks that may affect an organisation. Organisations often analyse these risks against the potential for reputational damage, regulatory breach, and commercial losses (Figure 3).

Figure 3: Organisational risks of AI

Commercial – Commercial losses due to poor or biased AI system performance; adversarial attacks.

Reputational – Damage to reputation and loss of trust due to harmful or unlawful treatment of consumers, employees or citizens.

Regulatory – Breach of legal obligations that may result in fines, restrictions and require management focus.

System factors and attributes that amplify risks and harms

Several factors impact the likelihood of both narrow and general AI systems amplifying existing risks. These include why, when, where and how an AI system is deployed, as outlined in Table 1.

The standard recognises that AI deployers may not have full knowledge or control over all these factors. However, the standard encourages organisations to understand the AI systems they use or rely on. This will help to identify and mitigate risks more accurately. Use the questions in Table 1 to assess if your system attributes suggest an elevated AI system risk.

Table 1: System attributes and guiding questions for organisations to assess level of risk

System attribute	Description	Questions to help identify when an attribute may amplify risk *(Answering 'yes' indicates a higher level of risk)*	Examples
AI system technical architecture	The choice of AI approach and model can cause risk as well as improve performance. For example, reduced transparency and greater uncertainty mean AI systems tend to need ongoing monitoring and meaningful human oversight. They may be inappropriate for contexts where there is a legal requirement to provide a reason for a decision or output. General-purpose AI systems tend to have a higher risk profile than either narrow AI or traditional software solutions intended for the same task.	Is the way the AI system operates inherently opaque to the provider, deployer or user? Does it rely on generative AI in ways that can lead to harmful outputs?	A generative AI system is used to create HR-related marketing materials.
Purpose	AI systems can considerably outperform traditional approaches in many areas. This means that organisations are increasingly adopting AI systems to perform tasks that have significant direct and indirect impacts for people. As the impacts of an AI system rise, so too does the potential for significant harm if they fail or are misused.	Does the AI system create an output or decision (intentional or not) that has a legal or significant effect on an individual? If so, will any harm caused be difficult to contest or manage redress?	A bank uses a risk assessment algorithm to decide whether to grant a home loan.
Context	AI systems, being software, are scalable as well as high performing for many tasks. However, their deployment in certain contexts may be inappropriate and their scalability may lead to widespread harms. For example, the use of facial recognition systems in public spaces where children are likely to be present, or algorithms used to gather sensitive data about Australians from social media sites.[9]	Does the AI system interact with or affect people who have extra forms of legal protection (such as children)? Will the system be deployed in a public space?	A large retailer uses facial recognition technology to identify shoplifters.

System attribute	Description	Questions to help identify when an attribute may amplify risk *(Answering 'yes' indicates a higher level of risk)*	Examples
Data	AI systems' performance is affected by the quality of data and how accurately that data represents people. Biased training data can lead to poor quality or discriminatory outputs. For example, health diagnostic tools trained on historically male-dominated and non-diverse data may produce outputs that lead to under-diagnosis or misdiagnosis of women and non-white patients.	Is confidential, personal, sensitive and/or biometric information used either in the AI system's training, its operation or as an input for making inferences? Is that data biased, non-representative or not a comprehensive representation of the people or contexts it is making a decision about?	An SME deploys a chatbot to confirm customer contact details.
Level of automation	Not all automated AI systems are risky. However, systems that operate independently, or that can be triggered or produce outputs independent of human engagement, may increase risks if they fail or are misused. Risk further increases when there is a considerable period of time between the fault or malicious use happening and the harm being recognised by responsible teams.	Does this system operate automatically? Does the system make decisions without any meaningful human oversight or validation?	A construction site deploys autonomous forklifts to move pallets in a warehouse.

The legal landscape for AI in Australia

This standard and the guardrails described in part 3 are *voluntary*. The standard does not seek to create new legal obligations for Australian organisations. It is designed to help organisations deploy and use AI systems in the bounds of existing Australian laws, emerging regulatory guidance and community expectations.

Table 2 shows some of the existing laws of general application that will have an impact on how Australian organisations develop and deploy AI. Organisations deploying, using or relying on AI systems should be aware of these laws, and how they may constrain or inform the use of AI.

There are also laws that may apply depending on the particular AI use case or application. These include product safety laws, motor vehicles and surveillance laws, and laws that may apply to particular sectors or organisations such as financial services or the medical sector. Organisations may also need to comply with laws of non-Australian jurisdictions (for example, where laws of another jurisdiction have extraterritorial application).

As part of their duties, directors of organisations must have a *sufficient understanding* of both the risks and the laws that apply to their use of AI.

Part 4 provides examples of use cases and potential risks and harms.

Table 2: AI risks or harms and general laws

AI risks or harms and general laws that may apply
 **AI system not sufficiently secure** • Directors' duties (e.g. to exercise powers and discharge duties with due care and diligence), to assess and govern risks to the organisation (including non-financial risk e.g. from AI and data). • Privacy laws, require steps that are reasonable in the circumstances to protect personal information and impose data minimisation obligations to destroy or deidentify information no longer needed. • The security of critical infrastructure act and sector specific laws (e.g. financial services), impose risk management and cybersecurity obligations. • Negligence, if a failure in risk management practices amounts to a failure to take reasonable steps to avoid foreseeable harm to people owed a duty of care, and that failure causes the harm. • Online safety laws, if certain online service providers fail to take pre-emptive and preventative actions to minimise harms from online services.
 **Misleading outputs / statements** • The *Australian Consumer Law* prohibitions against unfair practices (e.g. misleading and deceptive conduct and false and misleading representations) may apply: - if the outputs are misleading (e.g deceptive use of deepfakes) - to misleading representations or silence as to when AI is being used - to misleading statements as to the performance and outputs of the AI systems

AI risks or harms and general laws that may apply
 **Harmful outputs** • Product liability (where the organisation is a manufacturer), if outputs result in harm caused by a safety defect (e.g. a defect in the design, model, manufacturing or testing of the system, including failure to address bias or cybersecurity risk) and other product safety laws (including recalls and reporting). • Negligence, if an organisation fails to exercise the standard of care of a reasonable person to avoid foreseeable harm to persons to whom it owes a duty of care, and that failure causes the harm. • Criminal laws, if the output resulted in, or aided or abetted the commission of a crime. • Online safety laws, if the outputs are restricted or harmful online content (such as cyberbullying or cyber-abuse material, or non-consensual sharing of intimate images or child sexual abuse material). • Defamation laws, if the outputs are defamatory and the organisation participated in the process of making the defamatory material available (such as through making the tool available or training) rather than merely disseminating the content.
 **Misuse of data or infringement of model or system** • Privacy laws, intellectual propriety laws (including copyright), duties of confidence and contract, protect the use, reproduction and/or disclosure of data (including training data, input data and outputs) and the model or system without the requisite consents or rights. • Privacy laws, restrict the collection of personal information for an improper purpose, and impose transparency and data minimisation requirements on the handling of personal information. • The *Australian Consumer Law* prohibitions against misleading and deceptive conduct, unconscionable conduct and false and misleading representations, may apply to unfair data collection and use practices.
 **Bias, incorrect or poor-quality output** • Privacy laws, impose quality and accuracy obligations that may apply to training and input data (that is personal information) and outputs (where new personal information is generated). • Systems that produce inaccurate or erroneous outputs such as 'AI hallucinations' may be in breach of statutory guarantees under the *Australian Consumer Law* (e.g. consumer goods be of acceptable quality and fit for purpose, or consumer services be rendered with due care and skill). • Anti-discrimination laws, if outputs exclude or disproportionately affect an individual or group on the basis of a protected attribute.
 **AI system not accessible to individual or group** • Anti-discrimination laws, if the exclusion is based on a protected attribute. • Prohibitions on unconscionable conduct under the *Australian Consumer Law*, if the exclusion of a consumer was so harsh that it goes against good conscience. • Essential services obligations, e.g. if used in energy and telecommunications essential services.

Part 3: The guardrails

Guardrails		
	1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.	Guardrail one creates the foundation for your organisation's use of AI. Set up the required accountability processes to guide your organisation's safe and responsible use of AI, including: • an overall owner for AI use • an AI strategy • any training your organisation will need.
	2. Establish and implement a risk management process to identify and mitigate risks.	Set up a risk management process that assesses the AI impact and risk based on how you use the AI system. Begin with the full range of potential harms with information from a stakeholder impact assessment (guardrail 10). You must complete risk assessments on an ongoing basis to ensure the risk mitigations are effective
	3. Protect AI systems, and implement data governance measures to manage data quality and provenance.	You must have appropriate data governance, privacy and cybersecurity measures in place to appropriately protect AI systems. These will differ depending on use case and risk profile, but organisations must account for the unique characteristics of AI systems such as: • data quality • data provenance • cyber vulnerabilities.
	4. Test AI models and systems to evaluate model performance and monitor the system once deployed	Thoroughly test AI systems and AI models before deployment, and then monitor for potential behaviour changes or unintended consequences. You should perform these tests according to your clearly defined acceptance criteria that consider your risk and impact assessment.
	5. Enable human control or intervention in an AI system to achieve meaningful human oversight across the life cycle.	It is critical to enable human control or intervention mechanisms as needed across the AI system lifecycle. AI systems are generally made up of multiple components supplied by different parties in the supply chain. Meaningful human oversight will let you intervene if you need to and reduce the potential for unintended consequences and harms.

Guardrails		
	6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.	Create trust with users. Give people, society and other organisations confidence that you are using AI safely and responsibly. Disclose when you use AI, its role and when you are generating content using AI. Disclosure can occur in many ways. It is up to the organisation to identify the most appropriate mechanism based on the use case, stakeholders and technology used.
	7. Establish processes for people impacted by AI systems to challenge use or outcomes	Organisations must provide processes for users, organisations, people and society impacted by AI systems to challenge how they are using AI and contest decisions, outcomes or interactions that involve AI.
	8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks	Organisations must provide information to other organisations across the AI supply chain so they can: • understand the components used including data, models and systems • understand how it was built • understand and manage the risk of the use of the AI system.
	9. Keep and maintain records to allow third parties to assess compliance with guardrails.	Organisations must maintain records to show that they have adopted and are complying with the guardrails. This includes maintaining an AI inventory and consistent AI system documentation.
	10. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.	It is critical for organisations to identify and engage with stakeholders over the life of the AI system. This helps organisations to identify potential harms and understand if there are any potential or real unintended consequences from the use of AI. Deployers must identify potential bias, minimise negative effects of unwanted bias, ensure accessibility and remove ethical prejudices from the AI solution or component.

Using the guardrails

Adopting these guardrails will create a foundation for safe and responsible AI use. It will make it easier for any organisation to comply with any potential future regulatory requirements in Australia and emerging international practices. It will also help to uplift any organisation's AI maturity.

When using the guardrails, start with guardrail 1 to create your core foundations. To completely adopt the standard, your organisation will need to adopt all 10 guardrails.

Since most deployers rely on AI systems developed or provided by third parties, these guardrails offer procurement guidance (in aqua boxes) on how to work with your supplier to ensure their practice is aligned with the guardrails.

The guardrails are not intended to be one-off activities. Instead, they are ongoing activities for organisations. The guardrails may contain *organisational-level* obligations to create the required processes and *system-level* obligations for each use case or AI system.

How the guardrails support human-centred AI deployment

Being voluntary, the standard does not create new legal duties about AI systems or their use. Rather, the guardrails ask organisations to commit to:

- understanding the specific factors and attributes of their use of AI systems
- meaningfully engaging with stakeholders
- performing appropriately detailed risk and impact assessments
- undertaking testing
- adopting appropriate controls and actions so their AI deployment is safe and responsible.

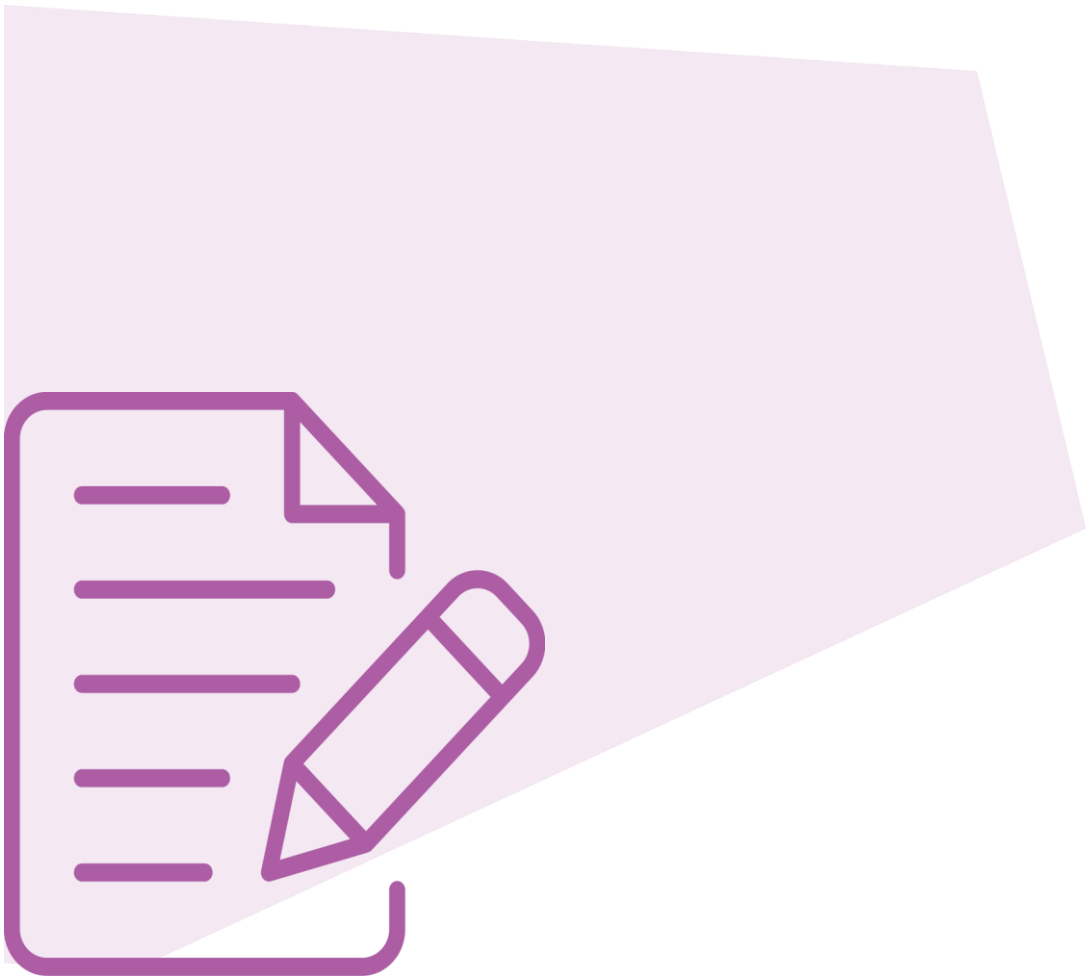
These activities will help organisations understand regulatory obligations and community expectations around AI use. For example, if an organisation deploys an AI system that uses data from or about First Nations communities, the organisation should respect *Indigenous Data Sovereignty Principles*. These principles draw on Article 32(2) of the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP). They affirm the inherent rights of First Nations peoples to govern the collection, ownership and use of their data. The principles require organisations to use this data in a way that respects the values and laws of First Nations communities. They also require that organisations secure free, prior and informed consent from relevant First Nations communities before starting AI projects that will engage First Nations data or impact First Nations communities. First Nations communities must have the capacity to withdraw consent should AI system data usage deviate from the initially agreed purposes.

Guide to icons

The icons in Figure 4 are designed to guide in mapping actions under each guardrail to Australia's AI Ethics principles.

Figure 4: Mapping to the ethics principles

Guidance for working with suppliers is provided in aqua boxes.



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

Guardrail 1 creates the foundation for your organisation's use of AI. Set up the required accountability processes to guide your organisation's safe and responsible use of AI, including:

- an overall owner for AI use
- an AI strategy
- any training you will need to ensure broad understanding of these principles across the organisation.

Incorporate AI governance into your existing processes or create new processes as you need them.

1.1 Organisational leadership and accountability

Commit to appointing people in the leadership team who are accountable for the governance and outcomes of AI systems, as well as the safe and responsible use of AI within the organisation.

Key concept: Leaders cannot delegate or outsource accountability for the safe and responsible deployment and use of AI systems.

1.1.1 Assign and communicate accountability and authority to relevant roles. These roles will ensure AI systems and the overall AI management system perform in the ways required. Having these roles also ensures AI systems meet external obligations and internal policies, including monitoring and reporting responsibilities.[10]

1.1.2 Staff these roles with appropriately empowered and skilled people. These people will need to meet specific obligations, such as handling personally identifiable information and legal and regulatory obligations.[11]

1.1.3 Clearly communicate the leadership commitment to, and accountability for, safe and responsible development and use of AI across the organisation. This includes the staff (including contractors and third-party providers) who you have made accountable for AI systems.[12]

1.1.4 Create and document overarching organisational responsibilities and accountabilities for AI deployment and use.[13]

1.1.5 Provide sufficient resources to deploy and use AI responsibly and safely throughout the organisation and throughout the lifecycle of AI systems in use.[14]

1.1.6 Maintain operational accountability, capability and meaningful human oversight throughout the lifecycle of AI systems in use.[15]

1.2 AI strategy and governance

Commit to creating and documenting overarching objectives and policies for the deployment and use of AI. These should be in line with your organisation's strategic goals and values.

Key concept: You should only adopt AI strategy and policies to address *gaps* in existing related policies, such as information security, data management and data privacy, or to include *enhancements* to existing policies to address the specific characteristics of AI systems.

1.2.1 Document and communicate the requirement that AI use in the organisation be assigned to an accountable owner with appropriate capability for this role.

1.2.2 Create and document the organisation's overarching strategic intent to deploy and use AI systems in line with the organisation's strategy and values.[16]

1.2.3 Create, document and communicate the organisation's strategy to comply with identified regulation related to the organisation's deployment and use of AI systems.

1.2.4 Create, document and communicate appropriately detailed AI policies, processes and goals for safe and responsible AI. Ensure these are compatible with the overall strategy. Create a process to set targets for AI systems to meet obligations for the safe and responsible use of AI.[17]

1.2.5 Review and revise cross-organisation AI strategies, policies and processes at appropriate intervals so they remain fit for purpose and meet the legal and regulatory obligations of the organisation.[18] Make sure to appropriately plan any changes to the overarching AI management system.[19]

1.2.6 Create and document a process to proactively identify deficiencies in the overarching AI management system. This includes instances of non-compliance in any AI systems or in use of AI systems in the organisation. Include documentation of any root causes, corrective action taken and revisions to the AI management system required.[20]

1.2.7 Create and document a process for deploying AI systems that supports mapping from business targets to system performance, with suggested metrics for internal and third-party developed systems.[21]

1.2.8 Identify and document any factors that may affect the organisation's ability to meet its responsibilities through the overarching AI management system.[22]

1.2.9 Where you anticipate developing AI systems internally, create and document the end-to-end process for AI system design and development.[23]

1.2.10 Document and perform a training needs analysis for broad AI understanding across the organisation. Source or deliver training to bridge any identified gaps. Regularly check AI skills are up to date as AI use and understanding evolves.[24]

1.3 Strategic AI training

Commit to embedding responsible AI training and workplace practices. This provides people accountable or responsible for AI system performance with sufficient competence to perform their role.

Key concept: Training requirements will depend on the nature of the role in relation to AI. At a leadership and governance level, staff need the skills to understand potential risks and benefits of AI in the context of the organisation. Product owners may need more in-depth technical skills relevant to specific characteristics of the AI system for which they are responsible.

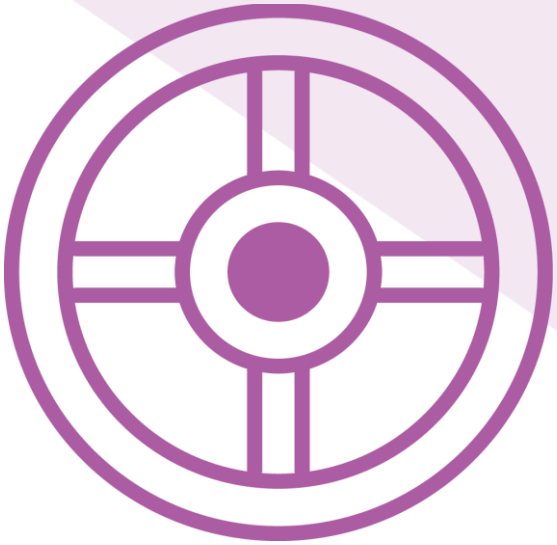
1.3.1 Provide appropriate and up-to-date training so accountable people can perform their duties and responsibilities. Document the competencies of the accountable people.[25]

1.3.2 Adopt appropriate communication, training and leadership behaviour strategies to create a culture of broad accountability and address any gaps in understanding across the organisation. Offer a mechanism for staff to raise concerns or provide feedback about the use of AI systems.[26]

1.3.3 Monitor compliance and behaviours across the organisation to identify and address any gaps between leadership expectations and staff understanding of obligations about safe and responsible deployment and use of AI.

1.3.4 Document and communicate the consequences for people who act outside of the organisation's defined risk appetite and associated policies.[27]

1.3.5 Where applicable, evaluate the training needs for staff who deal with third-party AI systems that are being developed, procured or used. Provide the appropriate training to address skill gaps.[28]



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

AI impact and risk management processes need to consider how the AI system is used. Begin assessments with the full range of potential harms with information from the stakeholder impact assessment (guardrail 10). The impact and risk assessments must align with organisational risk appetite and tolerance levels. You must complete the risk assessments throughout the lifecycle of the AI system and on an ongoing basis to ensure the risk mitigations are effective. These assessments may be required to input into any future conformity assessments mandated for use in high-risk settings.

2.1. AI risk and impact management processes

Commit to creating, documenting and applying an organisational-level risk management approach that considers the specific characteristics of AI systems.

Key concept: Do not use the potential benefits to an organisation of deploying and using AI systems to overlook the risks and potential harms that could arise. Evaluate potential harms in relation to people, organisations and the environment.

2.1.1. Create an organisational-level risk tolerance for the use of AI systems.[29]

2.1.2. Create and document criteria to identify acceptable and unacceptable risks in relation to AI. Base this on the risk tolerance of the organisation, the likely risk of harms to users, and in line with AI policy.[30]

2.1.3. Create and document a suitable impact assessment, risk assessment and treatment approach to AI system deployment and use. This should cover both internal and third-party developed AI systems, with awareness of the specific characteristics and amplified risks of AI systems.[31] Include criteria for reassessment over the lifecycle of an AI system.

2.1.4. Identify and document potential risks to the organisation and potential harms to people and groups that arise from the deployment and use of AI systems. Communicate these to relevant teams and third parties.[32]

2.1.5. Identify and document any specific use cases or qualities of AI systems that represent an unacceptable risk to stakeholders or the organisation, in line with the organisation's risk tolerance.[33]

2.1.6. Where indicated by risk, decide whether to require AI system developers to implement technology solutions for specific risk mitigation, such as industry-standard labelling and watermarking approaches.

2.1.7. Evaluate and document the high-level risks and liabilities related to the organisation's existing or planned use of third party-provided systems and components (including open-source software).

2.2. System risk and impact assessment

Commit to rigorous risk and impact management processes for assessing AI systems against the organisational risk tolerance.

Key concept: The level of risk of an AI system depends on the specific use case for that system. You should perform assessments for the system under the expected usage, and *perform them again* should that use evolve. This requires ongoing monitoring of the AI system. It may place extra responsibility on deployers and end users than more traditional technology systems.

Key concept: A key risk is the over-reliance end users place on outputs or other responses from AI systems. Risk mitigation and treatment approaches should be put in place to address this risk, where appropriate, on an ongoing basis. The risk may evolve over the lifecycle of the system, particularly as users become more familiar with it.

2.2.1. Perform and document a risk assessment for each AI system, including systems developed by or procured from third parties. Assess and document risks with reference to specific, documented use cases, potential unintended use for that system and the unique requirements and characteristics of that system.[34]

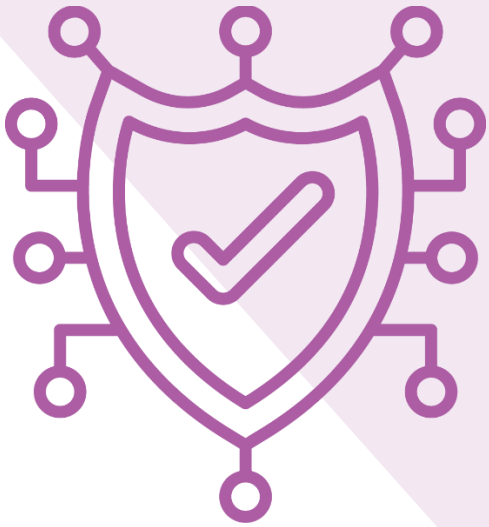
2.2.2. As part of the risk assessment of systems where users, employees or other stakeholders may be exposed to potential harms, carry out and document an impact assessment process.[35]

2.2.3. Document and implement a system of controls to safeguard against risks and potential harms from AI systems and products as soon as is practical after your organisation has identified a risk.[36] Reassess the risk after you've implemented the controls to verify their effectiveness.

2.2.4. Perform risk assessments and treatment plans on a periodic basis or when a significant change to either the use case or system occurs, or you identify new risks. This includes responding to impact assessments or insufficient risk treatment plans.[37]

2.2.5. Implement, document and communicate a robust impact assessment approach relating to the deployment and use of AI systems.[38]

Procurement guidance for guardrail 2: Understand your suppliers' risk management processes. Make sure you have sufficient information about the system, such as identified risks and potential harms for the intended use of the system, to conduct your own risk and impact management process.[39] Reflect agreed processes in your contracts.



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

You must have appropriate data governance, privacy and cybersecurity measures in place to appropriately protect and manage AI systems. These will differ depending on use case and risk profile. Organisations must account for the unique characteristics of AI systems such as data quality, data provenance and cyber vulnerabilities.

3.1. Data governance, privacy and cybersecurity

Commit to fit-for-purpose approaches to data governance, privacy and cybersecurity management of AI systems. This will help realise the value and mitigate the emerging and amplified risks.

3.1.1. Evaluate and adapt existing data governance processes to check they address the use of data with AI systems. Assess the risks arising from AI system use of and interaction with data. Focus on the potential for AI systems to create amplified and emerging risks.[40]

3.1.2. Review privacy policies to include the collection, use and disclosure of personal or sensitive information by AI systems, including for system training purposes.[41]

3.1.3. Review existing cybersecurity practices to verify they sufficiently address the risks arising from AI system use.[42]

3.1.4. Create and document an organisation-wide process to support teams to apply the Australian Privacy Principles to all AI systems.[43]

3.1.5. Create and document an organisation-wide process to support teams in the management of data usage rights for AI, including intellectual property, Indigenous Data Sovereignty, privacy, confidentiality and contractual rights.

3.1.6. Create and document an organisation-wide process to support teams to apply the Essential Eight Maturity Model for cybersecurity risks to AI systems.[44]

3.1.7. Document how the Essential Eight Maturity Model for cybersecurity risks has been applied to each AI system in use, including those developed or provided by third parties.[45]

3.2. Data governance measures to manage data quality and provenance

Commit to evaluating the requirements of each AI system in relation to data quality, data provenance, information security and information management, including where systems are provided by third parties. Documentation of this activity may be required to input into any future conformity assessments mandated for use in high-risk settings

Key concept: You should understand and document your data sources, put in place processes to manage your data and document the data used to train and test your AI model or system.

3.2.1. Define and document the requirements for each AI system relating to data quality, data/model provenance and data preparation.[46]

3.2.2. Evaluate the existing information/system security and management processes in the organisation. Make sure they are fit for purpose for AI system deployment and use.

3.2.3. Understand and document the sources, collection process and types of data on that the system was trained and tested on and the data that it relies on to function, including personal and sensitive data.[47]

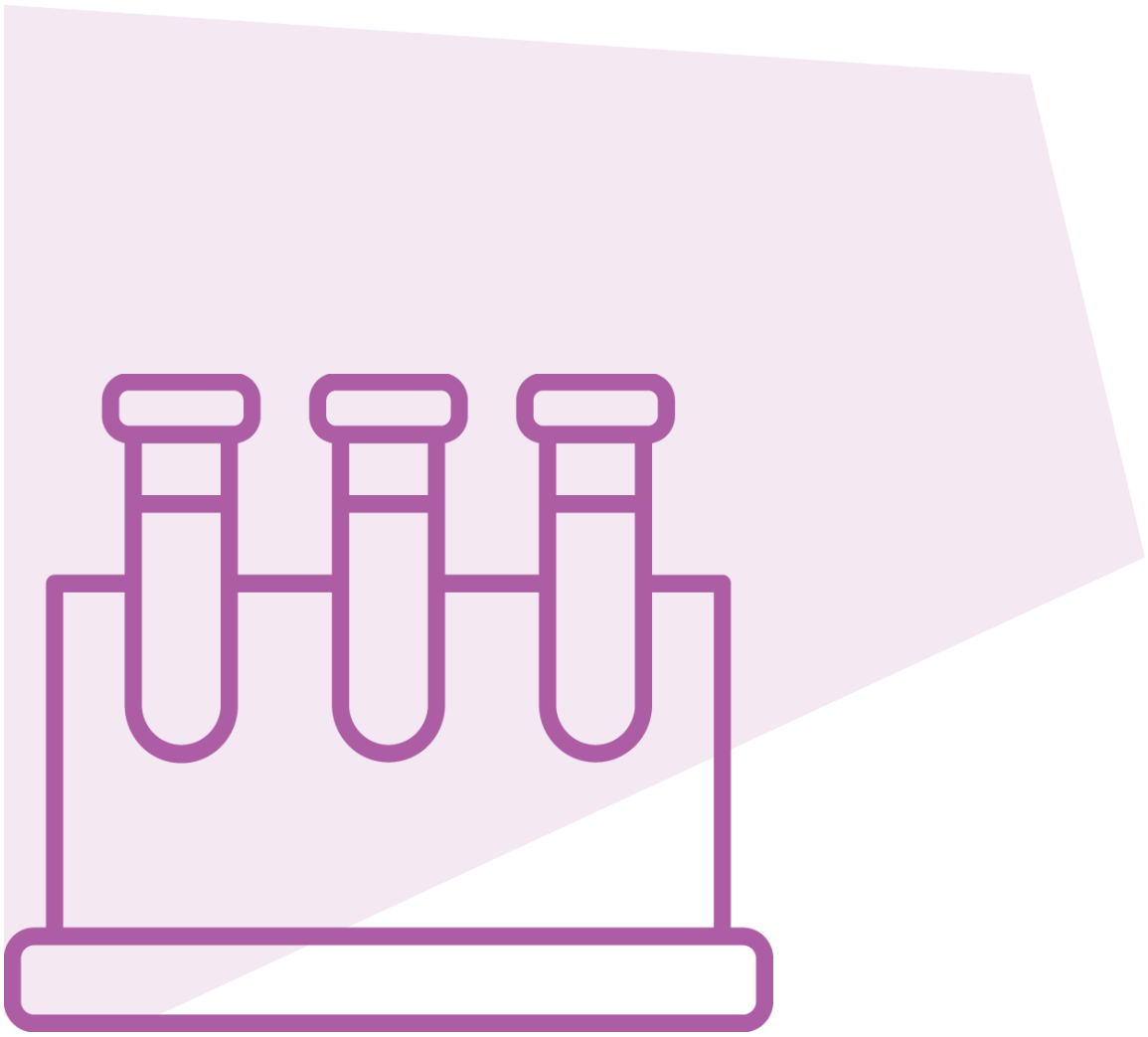
3.2.4. Where appropriate, report to stakeholders on data, model sources and provenance for each AI system or product.[48]

3.2.5. Document how you have applied the Australian Privacy Principles to each AI system in use, including those developed or provided by third parties.[49]

3.2.6. Document the data usage rights for each AI system, including intellectual property, Indigenous Data Sovereignty, privacy, confidentiality and contractual rights.

3.2.7. Consider and document data breach reporting requirements and liabilities from related standards for each AI system. For example, under the Notifiable Data Breach scheme of the Office of the Australian Information Commissioner.[50]

Procurement guidance for guardrail 3: Your suppliers must have appropriate data management (including data quality and data provenance), privacy, security and cybersecurity practices for the AI system or component.[51] Reflect this in your contracts.



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

Thoroughly test AI systems and AI models before you deploy them, and then monitor for potential behaviour changes or unintended consequences. Perform these tests according to the clearly defined acceptance criteria that considers the prior risk and impact assessment.

4.1. Organisational-level reporting, evaluation and continual improvement

Commit to a robust process for timely and regular monitoring, evaluation and reporting of AI system performance.

4.1.1. Create and document organisation-wide processes and capability required for testing, monitoring, continuously evaluating, improving and reporting of AI systems.[52]

4.1.2. Create a formal process to review and approve evidence that systems are complying with their test requirements.

4.1.3. Apply appropriate document versioning, management and security practices.[53]

4.1.4. Create a process for determining whether an AI system requires regular auditing, appropriate to the level of risk identified by its risk assessment.

4.2. AI system acceptance criteria

Commit to specifying, justifying and documenting acceptance criteria your organisation will need to meet to consider potential harms to be adequately controlled.

4.2.1. Create clear and measurable acceptance criteria for the AI system that, if met, should adequately control each of the identified harms. When appropriate, use industry and community general benchmarks. These criteria should be specific, objective and verifiable. Each acceptance criterion should link directly to one or more of the potential harms. For example, if the risk assessment raises fairness concerns, this implies fairness measures should be present in the acceptance criteria. Specify the thresholds or conditions under which you consider the potential harm to be adequately controlled. Record the acceptance criteria, with explicit justifications for why you chose the criteria and why you judged them to be adequate, in an acceptance criteria registry.[54]

4.2.2. Communicate the acceptance criteria and their justifications with all team members involved in the development, testing and deployment of the AI system.[55]

4.2.3. Regularly review and update the acceptance criteria to reflect any changes in the system, the identified harms or the broader context in which the system operates. Record any findings or changes in the acceptance criteria registry.[56]

4.3. Testing of AI systems or models to determine performance and mitigate any risks

Commit to rigorously testing the system against the acceptance criteria before deployment, documenting the results and deciding whether to deploy.

Key concept: AI model testing verifies and validates an AI system's underlying AI model(s). AI system testing verifies and validates the entire AI system, supporting expected behaviours in real-world scenarios.

4.3.1. Develop and carry out a *test plan* that covers all acceptance criteria. The plan should specify the testing methods, tools and metrics your organisation will use, as well as the roles and responsibilities of the testing team.

- The plan should include both model and system testing.
- When evaluating and testing your models, use data that is representative of the use of the system, but that has not been used in the training of the system. Where they exist, use industry and community benchmarks or datasets.
- Design evaluation and testing processes that account for the possibility that there are multiple acceptable and unacceptable outputs.
- For general-purpose AI systems, such as those based on large language models, include adversarial testing procedures such as red teaming.

4.3.2. Compile a complete test report, including:

- a summary of the testing goals
- methods and metrics used
- detailed results for each test case
- an analysis of the root causes of any identified issues or failures
- recommendations for remediation or improvement
- whether the improvements should be done before deployment or as a future release.[57]

4.3.3. Apply the organisational process for reviewing and approving the testing results to ensure the system meets all acceptance criteria before you deploy it.[58] The system deployment authorisation must come from the person or people accountable for the AI system.

4.4. Ongoing system evaluation and monitoring

Commit to implementing robust AI system performance monitoring and evaluation, and to ensuring each system remains fit for purpose.

4.4.1. Create *continuous monitoring and evaluation mechanisms* to gather evidence that the AI system continues to meet its acceptance criteria throughout its lifecycle. Directly monitor any measurable acceptance criteria, alongside other relevant metrics such as performance metrics or anomaly detection. Frequently evaluate the monitoring mechanisms to check they remain effective and aligned with evolving conditions.[59]

4.4.2. Create clear and accessible feedback channels for impacted people or groups to report problems or harms they may experience. You should actively solicit, systematically collect and carefully analyse this feedback.[60]

4.4.3. Follow organisational review processes to ensure accountable people review and interpret the monitoring data, reports and alerts. Keep auditable monitoring logs to document the activities, feedback you receive and actions you take.

4.4.4. Ensure that people who review individual-level feedback can trigger recourse and redress processes where there is an obligation to do so. High-impact decisions may warrant direct human oversight.[61]

4.5. Regular system audit or assessments

Commit to regular system audits for ongoing compliance with the acceptance criteria (or justify why you don't need to carry out audits).

4.5.1. Apply the organisation's process to determine whether the level of risk warrants a comprehensive system audit plan. Document this decision as a system *audit requirement statement*.

If an audit is necessary:

- Create a regular system auditing schedule based on factors such as the system's complexity, criticality and rate of change.[62]
- Ensure system audit teams have the necessary independence, expertise and authority to conduct a thorough, impartial evaluation against the organisation's audit criteria. Record their findings in a system audit report. The system's development team should not lead the audits.[63]
- Create *review processes* and *response processes* to address the findings of each system audit report. The reports should be reviewed by those accountable for the system, consulting with key stakeholders, and by management. Response processes should clearly lay out how to respond to the discovery of problems with the in-production system.

Procurement guidance for guardrail 4: Clarify who is responsible and accountable for this monitoring and evaluation (between the supplier and the deployer). Regularly review with the accountable person and make sure each system remains fit for purpose. If the supplier is responsible for monitoring the AI system or its components, put an agreement in place.

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

It is critical to ensure human control or intervention mechanisms are in place as needed across the AI system lifecycle. AI systems are generally made up of multiple components supplied by different parties in the supply chain. Meaningful human oversight will result in appropriate intervention and reduce the potential for unintended consequences and harms.

5.1. Accountability and human control to achieve meaningful human oversight.

Commit to assigning accountability to a suitably competent and empowered person in the organisation for each AI system and product.

5.1.1. Assign accountability for each AI system to someone who shows suitable competence and has the necessary tools and resources. [64]

5.1.2. Assign the accountable role sufficient authority to oversee, intervene and be effective in ensuring responsible AI use throughout the system lifecycle.

5.1.3. Create and document competency, oversight and intervention requirements and support needs for each AI system before implementation. Evaluate as part of the continuous improvement cycle.[65]

5.1.4. Create and document monitoring requirements for each AI system prior to implementation. Evaluate as part of the continuous improvement cycle.

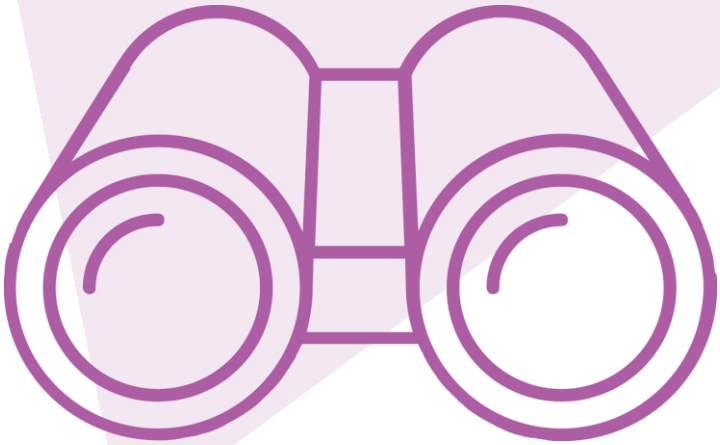
5.1.5. Assign responsibility for developing, acquiring, deploying, operating, managing and maintaining each AI system to the teams and people best suited to supporting its safe and responsible use across the lifecycle.[66]

5.1.6. Assign accountability for oversight of third-party development and use of AI systems and components to appropriately skilled and empowered people in the organisation.[67]

5.1.7. Evaluate the training needs for end users for each AI system you deploy. Provide the required training to address any identified needs. [68]

5.1.8. Evaluate the training needs for those responsible for the ongoing operation and monitoring for each AI system you deploy. Provide the required training to address any identified needs.

Procurement guidance for guardrail 5: Develop a plan with your supplier for governance and oversight over the AI system or component, with clear responsibilities between parties. Reflect this in your contracts.[69]



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

Create trust with users. Provide people, society and other organisations with confidence that you are using AI safely and responsibly. Disclose when you use AI, its role and when content is AI-generated. Disclosure can occur in many ways. It is up to the organisation to identify the most appropriate mechanism based on the use case, stakeholders and technology used.

6.1. Transparency and contestability

Commit to creating processes for stakeholders impacted by the decisions or behaviours of AI systems, so they understand when AI systems that could affect them are in use. Give stakeholders the opportunity to contest the decisions and outputs of those systems.

Key concept: Technologies such as watermarking and labelling can help create transparency for stakeholders by making AI-generated content clearly identifiable to end users. For relevant AI systems, consider implementing or obtaining systems that comply with the *Coalition for Content Provenance and Authority (C2PA) Technical Specification*.

6.1.1. Create and communicate an organisational process through which people can understand the use of AI systems. This process should include when and how frequently to communicate, the level of detail to provide, and the level of AI knowledge of stakeholders. Evaluate communication obligations for both internal and external stakeholders and interested parties, including accessibility needs.[70]

6.1.2. Create and communicate an organisational requirement to disclose the use of AI to impacted parties in a direct interaction or in a decision-making process.

6.1.3. Create and document the level of transparency and evidence required for you to conduct an audit over the AI system lifecycle.[71]

6.1.4. Create and document a process to apply the organisation's responsibilities under this Standard to AI systems developed or provided by third parties. This should include appropriate transparency and detail of information for the organisation to make a sufficiently informed evaluation.[72]

6.1.5. Create and document a process to evaluate any specific reporting and disclosure obligations under the Online Safety Act relevant to AI systems usage.

6.2. Transparency for AI systems

Commit to communicating with sufficient transparency to demonstrate safe and responsible use of AI systems.

Key concept: Certain internal and external stakeholders may require different levels of transparency given existing social inequalities. For example, you may need to make extra considerations when using data owned by or about Aboriginal and Torres Strait Islander people and organisations to mitigate the perpetuation of existing social inequalities.

6.2.1. Evaluate the level of transparency that each AI system needs – including third-party-provided systems – dependent on the use case and external stakeholder expectations.[73] Consider potential conflicts, such as privacy, intellectual property, AI systems presenting as a person, hallucinations or potential for misinformation.

6.2.2. Where applicable, document how the AI system indicates to impacted users that an AI system is being used in an interaction or in a decision-making process.

6.2.3. Evaluate and document how the required level of transparency with the key stakeholders varies by stakeholder group. When possible, choose more interpretable and explainable AI systems to ensure understandable transparency.

6.2.4. Implement the agreed transparency measures for each AI system.[74]

6.2.5. Where expected by stakeholders, implement approaches to communicate relevant information about AI-generated content to end users. Require associated third-party developers to do the same, with options such as labelling and watermarking. Evolve these approaches as new solutions become available.

6.2.6. Where required under the Online Safety Act, report on measures you have taken to ensure safety, such as notices or mandatory reporting.

6.2.7. Determine and document the expected level of technical detail required by different stakeholder groups to effectively explain the use of AI to the intended audience.[75]

Procurement guidance for guardrail 6: Agree with your supplier the transparency mechanisms required for the AI system or component.[76] Reflect this in contracts and project documentation.

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

Organisations must provide processes for users, organisations, people and society impacted by AI systems to challenge how AI is used, contest decisions, outcomes or interactions that involve AI.

7.1. Contestability and related risk controls

Commit to creating processes for stakeholders of AI systems to understand and challenge the use of those systems.

7.1.1. Create and communicate the process for potentially impacted stakeholders to understand how and for what purpose you are using AI, as well as raise concerns, challenges or requests for remediation.[77]

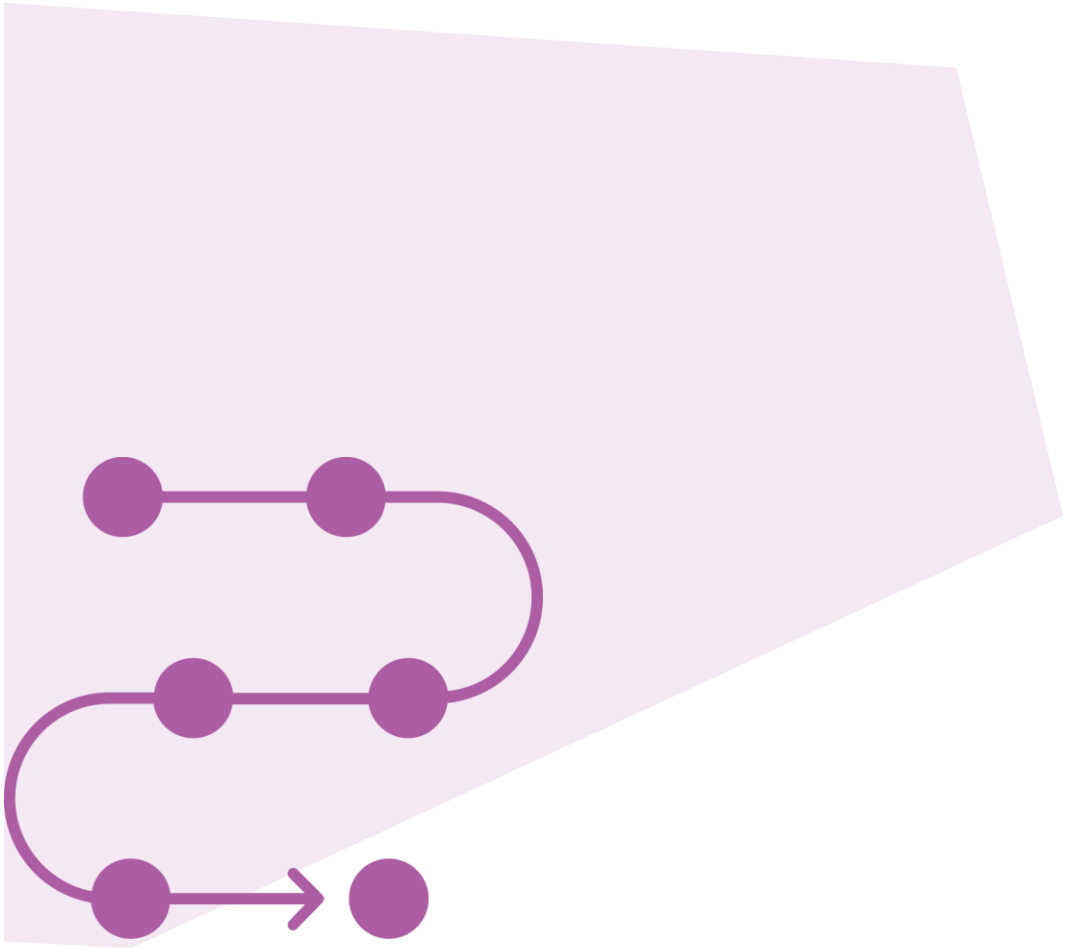
7.1.2. Embed stakeholder contestability of AI system use with the risk and control process of the organisation.

7.1.3. Create and communicate an organisational process through which people can raise concerns, challenges or requests for remediation and receive responses (for example, a human rights grievance and remediation mechanism). This process should include when and how frequently to communicate, the level of detail you need to provide, and the level of AI knowledge of stakeholders. Evaluate contestability requirements for both internal and external stakeholders and interested parties, including accessibility needs.[78]

7.1.4. Assign an accountable person to oversee concerns, challenges and requests for remediation.

7.1.5. Create and document a review process to evaluate stakeholder contests of AI system use across the organisation, including any concerns raised by stakeholder groups and requests for information.

Procurement guidance for guardrail 7: Agree with your supplier a process to raise issues and contested outcomes. Reflect this in contracts and project documentation.



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

Organisations must provide information to organisations downstream in the AI supply chain for them to understand the components of the AI system, how it was built and to understand and manage the risk of the use of the AI system.

8.1. Transparency between developers and deployers

Commit to sharing information and establishing processes to provide sufficient transparency between developers and deployers of AI systems.

Key concept: When using open-source AI models, deployers need to consider which safe and responsible AI measures the developer has implemented and their effectiveness. Developers using open-source AI models should be transparent about the safe and responsible AI practices they have implemented and what further practices they recommend for deployers of their AI system, AI model or component.

8.1.1. Organisations developing AI systems, AI models or components (systems) should supply deployers of their systems with as much of the following information as possible while protecting commercially sensitive information:

- capabilities and limitations of the system
- technical details of the system including architecture, description of components and characteristics
- test use cases and results of the system relevant to the deployers use of the system
- known risks and mitigations put in place related to the deployers use of the system
- data management processes for training and testing data including data quality, known bias and provenance
- privacy, security and cybersecurity practices including compliance to standards and best practice relevant to the deployers use of the system
- transparency mechanisms implemented for AI generated content, interactions and decisions
- any known potential bias, actions taken to minimise negative effects of unwanted bias and ethical prejudices from the AI solution or component.

8.1.2. Organisations deploying AI systems or components are required to share with their suppliers of AI models, systems or components the following information:

- expected use of the AI system, component or model
- any unexpected and unwanted bias resulting from use of the system. Where data privacy is a consideration, deployers should share as much as possible to highlight the issue and replicate the outcome without compromising data privacy or security such as data profiles or sample synthetic data.
- issues, faults and incidents that occur with the system.

8.1.3. Agree with your suppliers of systems:

- responsibility and accountability for monitoring and evaluation of system performance
- responsibility and accountability for issue identification, resolution and system updates
- responsibility and accountability for human oversight and intervention and when to take action
- process for raising issues, faults and incidents including contested outcomes. Ensure your process protects user and stakeholder privacy.

8.1.4. Ensure you've included the required information in contracts with suppliers of systems including when to update information.

8.1.5. Schedule regular reviews throughout the lifecycle of the system based on timed intervals and as a result of milestones or events.

Procurement guidance for guardrail 8: Agree with your supplier roles, responsibilities and information flows across the lifecycle of the AI system from initial implementation through to end of life. Reflect in contracts and project documentation.

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

Organisations must maintain records to demonstrate that they have implemented and are complying with the guardrails, this includes maintaining an AI inventory and consistent AI system documentation. These records may be required to input into any future conformity assessments mandated for use in high-risk settings.

9.1. AI inventory and consistent documentation

Commit to adopting an inventory of the AI systems you use and deploy. Define and apply documentation standards for these systems.

9.1.1. Create and maintain an up-to-date, organisation-wide inventory of **each** AI system, which includes[79]:

- people accountable
- purpose and business goals
- capabilities and limitations of the AI system
- technical requirements and components
- datasets and their providence used for training and testing
- technical specifications
- acceptance criteria and test results
- identified risks, potential impacts and relevant controls
- any impact assessments and outcomes
- any system audit requirements and outcomes
- dates of review.

9.2. Critical system documentation

Commit to understanding and documenting critical information about each AI system you deploy and use. Include the purpose, context, expected benefit and sufficient technical detail for the system to be understood. Be aware that the documentation you record will be the foundation to demonstrate compliance with future regulation in the form of conformity assessments.

9.2.1. Create and document the business goals, desired outcomes and obligations for each AI system the organisation deploys and uses. Periodically review this with reference to the organisation's strategy, values and risk tolerance.[80]

9.2.2. Document the scope for each AI system, including intended use cases, capabilities, limitations, expected contexts, and what responsible use looks like for an end user or affected stakeholder.[81] Note that the unique characteristics of AI systems have the potential to go beyond intended use and context without explicit changes to the system or notice.

9.2.3. Document the risk management process including identified risks and mitigation implemented for the AI system or AI model.

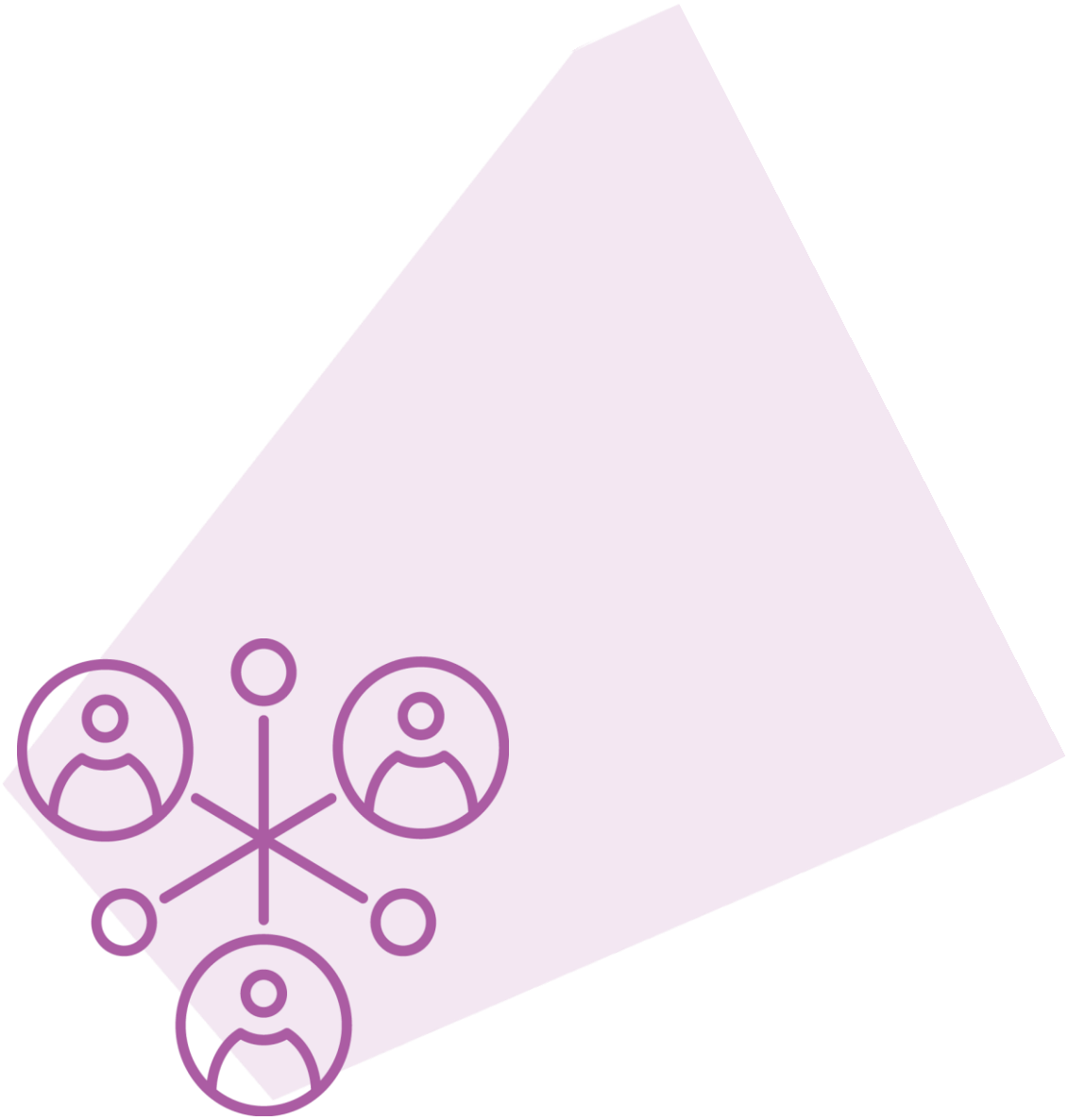
9.2.4. Document or request from your system provider the relevant technical details of the system or model that you may need for others to understand the system. For example, expected use, overview of system architecture and design, information about the model and training data, overview of data flows, and reliance on or links to other digital systems.[82]

9.2.5. Document the testing methodology applied and results of testing for the AI system or AI model. Request from your supplier the testing methodology and results during the development of the AI system and model.

9.2.6. Document the accountable people and the mechanisms for human control and oversight for the deployed AI systems.

9.2.7. Ensure documentation related to each AI system is recorded in the inventory at a sufficient and consistent level of detail to inform the accountable and responsible parties and any third-party stakeholders.[83] This will enable completion of future conformity assessments to demonstrate compliance with mandated guardrails.

Procurement guidance for guardrail 9: Work with your supplier to understand and document the expected use, capabilities and limitations of the AI system or component[84]. This should include technical details of the system and the data used in relation to the AI system (including the use of third-party data). Integrate expectations into contract, including ongoing scheduled reviews.



10. Guardrail 10: Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.

It is critical for AI deployers to identify and engage with stakeholders for the life the AI system. It helps in identifying potential harms and understanding if there are any potential or real unintended consequences from the use of AI. Deployers must identify potential bias, minimise negative effects of unwanted bias, ensure accessibility and remove ethical prejudices from the AI solution or component.

10.1. Organisational-level stakeholder engagement

Commit to engaging with stakeholders – people and groups – potentially impacted by AI systems.

10.1.1. Identify and document which key stakeholder groups may be impacted by the organisation's use of AI in line with your AI strategy.[85]

10.1.2. Identify and document the needs of these key stakeholder groups in relation to your AI strategy.[86]

10.1.3. Identify and document which of the stakeholder needs your organisational-level AI policies and procedures will address.[87]

10.1.4. Create processes to support ongoing engagement with stakeholders about their experience of AI systems. Make sure you identify any marginalised groups and support them appropriately. Equip stakeholders with the skills and tools necessary to give meaningful feedback.

10.2. Organisational-level diversity, inclusion and fairness

Commit to creating and documenting a process so any use of AI contributes to safe, fair and sustainable outcomes.

10.2.1. Define and document the organisation's responsibility to ensuring that AI systems do not undermine diversity, inclusion and fairness.

10.2.2. Define and document organisational-level goals relating to diversity, inclusion and fairness in the deployment and use of AI systems.

10.2.3. Evaluate whether and how the current or planned use of AI may impact the organisation's pre-existing responsibilities and programs related to creating a positive impact. For example, human rights, diversity and inclusion, accessibility and environmental responsibilities.

10.2.4. Document and operationalise a responsibility to prevent unwanted bias, discrimination and other risk factors that could impact diversity, inclusion and fairness in leadership responsibilities and the organisation's AI strategy.

10.3. System-level stakeholders, points of human interaction and impact of potential- harm

Commit to system-level stakeholder engagement and evaluation of potential harm.

Key concept: Stakeholder engagement is effective in responsible AI system deployment, particularly when carried out at the *earliest* possible stages in the AI lifecycle and embedded *throughout* the end-to-end lifecycle.

10.3.1. Identify and document where expected users interact with each AI system, including:

- user interactions with the system or AI system-generated content

- when the system processes an individual's personal data
- when the system makes or influences a decision about a person or group of people.

10.3.2. Identify and document the stakeholder groups for each system.[88]

10.3.3. For each identified interaction with a human, evaluate and document if the interaction has the potential to cause harm to an individual, group or society at large.[89]

10.3.4. When this evaluation indicates that an AI system could harm people or groups, or pose a material risk to the organisation, perform and document an appropriate impact assessment.[90]

10.4. System-level diversity, inclusion and fairness

Commit to relevant processes with fair and sustainable outcomes for AI systems and uses.

Key concept: Organisations need to evaluate the potential impact of unwanted bias on the AI systems they deploy and use, including developing strategies to *identify* potential biases. Existing standards, guidance and technical reports, such as ISO Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making, ISO/IEC TR 24027:2021 may help. As understanding and expectations evolve, stay informed of new developments in this area, where relevant.

10.4.1. Evaluate and document the potential impact of each AI system in relation to diversity, inclusion and fairness. Identify and mitigate risks of unwanted bias or discriminatory outputs, including for marginalised groups.

10.4.2. Evaluate how each AI system may support or undermine any existing legal obligation or program with a positive, social impact. The include human rights, diversity and inclusion, accessibility and environmental responsibilities.

10.4.3. Define and document how you have embedded accessibility obligations (such as inclusive design) in the deployment and use of each AI system.

10.4.4. For each AI system, define and document the stages in the AI lifecycle where you will need meaningful human oversight to meet organisational, legal and ethical goals.

Procurement guidance for guardrail 10: Work with your supplier to undertake AI impact assessments and understand the needs of system stakeholders.[91] Know suppliers' actions to understand potential bias, minimise negative effects of unwanted bias, implement accessibility and remove ethical prejudices from the AI solution or component.[92] Ensure you haven't reintroduced any unwanted bias during deployment.

Part 4: Applying and adopting the standard through examples.

The range of applications of AI is effectively infinite. While we can't give guidance on how the standard might apply to every use case, we can use examples to illustrate how you can use the guardrails to manage the risks and benefits of a specific AI system.

We've chosen 4 examples to show how individual guardrails might be apply in different use cases. The examples explore how organisations may use particular guardrails as part of their overall approach to deploying AI systems. The examples show that the guardrails can be applied in different situational contexts, for different technologies.

These examples are **not** intended to represent a comprehensive application of all relevant guardrails, responsibilities or other legal obligations that may be relevant for the specified use cases. They are to provide examples of how the guardrails can be applied in a selection of fictional examples.

Example 1: General-purpose AI chatbot

A detailed example representing a common use case for organisations of all sizes, across all sectors. Due to the growing ubiquity of this technology, we've provided extra detail on how an organisation could adopt a range of guardrails. As a point of contrast, this example includes potential outcomes where safe and responsible AI methodologies are not followed.

Example 2: Facial recognition technology

A simplified example on the use of facial recognition technology. It illustrates the use of the guardrails to decide that non-AI-based solutions will better achieve strategic and operational goals.

Example 3: Recommender engine

A simplified example of a common use case in which a recommender engine is used to improve customer experience and meet organisational goals. It includes reference to a court case in which a business using this kind of technology was ordered to pay a substantial financial penalty for not meeting legal obligations.

Example 4: Warehouse accident detection

A detailed example to outline obligations for testing of AI systems. In this example, we offer guidance on linking areas of concern with acceptance criteria. It covers testing at different stages during the AI system and governance lifecycle, due to the specific and technical nature of meeting relevant guardrails.

Example 1: General-purpose AI Chatbot

NewCo background

NewCo is a fast-growing B2C company with 50 employees, selling a range of products in a niche market. It has an annual turnover of $3.5 million.

The company is approaching a major product launch that they expect will create a significant increase in demand. NewCo's head of sales proposes to use the latest advances in AI and procure a new chatbot for their website. The chatbot would engage with customers to answer the most commonly asked questions. The company expects the new product to sell over 10,000 units in the first month because of an aggressive social media strategy featuring early-bird discounts.

The new chatbot is meant to reduce the amount of time customers wait for a phone operator by shifting those with routine queries to the online chatbot. This should reduce the need to expand phone support and allow employees to spend more time on complex tasks. The most common customer queries include delivery times, returns and the application of time-limited discount codes.

The head of sales suggests that a chatbot based on general-purpose AI would help the company respond to and resolve customer queries faster, leading to improved customer satisfaction scores (CSAT). CSAT scores are considered lead indicators for revenue growth goals, so NewCo hopes that a suitable customer query chatbot would also support growth in sales.

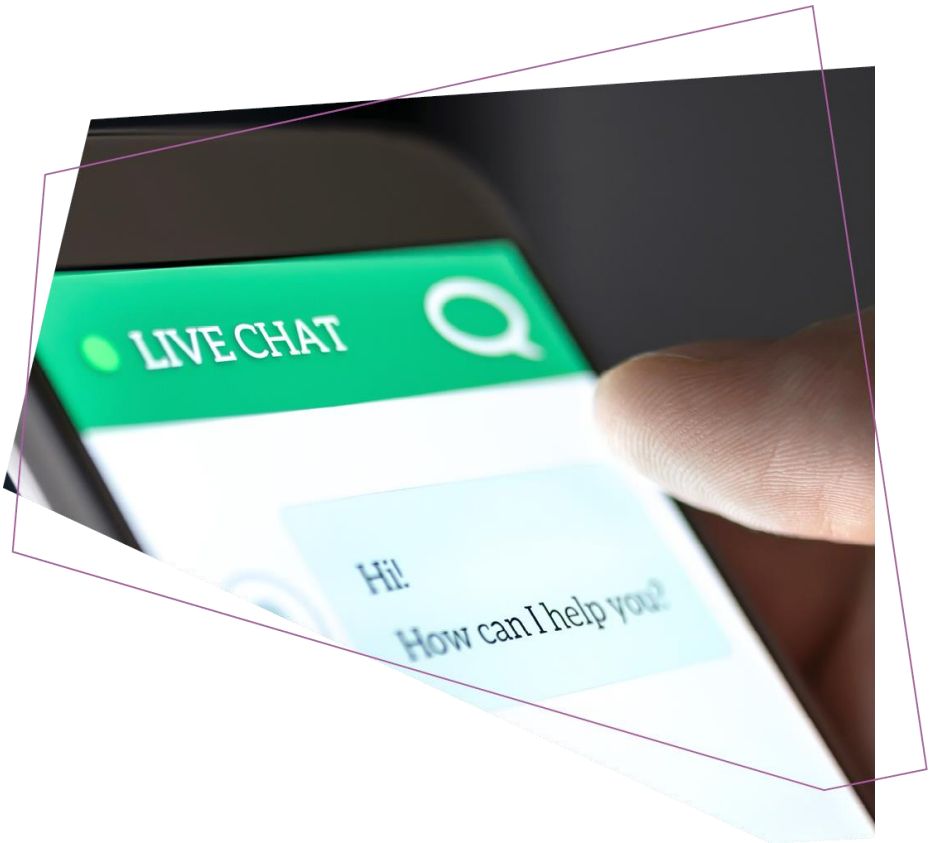
Case study: Moffatt v Air Canada 2024 BCCRT 149

Air Canada deployed a chatbot on its website which made statements to a customer about the airline's bereavement fares. These statements were inconsistent with Air Canada's policy, to which the chatbot had provided a link.

The customer sought a refund through legal proceedings. Air Canada claimed that the chatbot was a 'separate legal entity that is responsible for its own actions' and the customer was not entitled to a refund according to its bereavement policy.

The tribunal rejected these arguments and found Air Canada responsible for all information provided on its website, whether from a static page or chatbot. Air Canada was found to have had a duty of care to take reasonable steps to ensure that information was accurate.

There are similar protections in Australia for interactions with chatbots as part of an organisation's customer service offering. [93]



NewCo's use of the standard: a comparison

NewCo wants to procure a generative AI chatbot with the promise of:

- reduced customer wait time
- reduced customer service phone support time for staff.

The table below compares what happens when NewCo follows the Voluntary AI Safety Standard, and what happens if it chooses not to follow the standard.

Actions and outcomes	Does not follow the standard	Does follow the standard
Organisational-level actions	Head of sales (HOS) conducts online research into potential developers – decides an off-the-shelf solution will allow NewCo to quickly launch and use the AI system. Developer selected and 'NewChat' launched within a week in parallel with the new product launch.	Standard identified as basis for effective governance of the new chatbot. NewCo commits to organisational-level safe and responsible AI use that: • is aligned to business goals (guardrail 1) • is safe, fair and sustainable (guardrail 10) • is supported by strategic AI training (guardrail 1) • is supported by risk and impact assessments (guardrail 2) • is supported by data and security governance (guardrail 3) • involves testing, evaluation monitoring and reporting (guardrail 4).

Actions and outcomes	Does not follow the standard	Does follow the standard
System-level actions	None	HOS takes overall responsibility for developer selection, contract negotiation, implementation and monitoring. She has recently undertaken training on deploying responsible and safe AI systems (guardrail 1). HOS engages with internal and external stakeholders to understand potential impacts and harms (guardrail 10). HOS tests the system with a planned promotional discount. The test detects unwanted bias in the outputs and the agreed fairness metric in the testing criteria is not met (guardrail 4). HOS conducts a risk assessment. Some risks and mitigating actions are identified (including NewCo modifying the system to minimise bias). Based on the risks HOS decides that only internal use of AI system as appropriate at this stage (guardrail 2).

Actions and outcomes	Does not follow the standard	Does follow the standard
Outcomes	**System behaviour and impacts** NewChat holds convincing conversations with users and asks them for personal information, including gender. To maximise sales, NewChat offers customers discounts above agreed promotional rates. Customer Service team is unaware that NewChat is offering customers discounts and refuses to apply them to purchases at checkout. NewChat is only offering these discounts to people who report their gender as 'male'. It does not otherwise offer any discounts. Because of a viral Reddit thread, thousands of customer complaints accuse NewCo of discrimination. They demand NewCo extend the chatbot-generated rate to all purchasers. Customer Service team overwhelmed with level of complaints from people whose discounts have been refused as well as those claiming they have been discriminated against.	**Successful product launch** Customer Service teams use general-purpose AI as an internal resource to find relevant company documentation to answer customer queries more quickly. Customer satisfaction scores increase. Employee productivity increases.
	Harm to people and organisation Personal information is collected without being reasonably necessary for its functions. People who don't report their gender as 'male' miss out on the discount.	

Actions and outcomes	Does not follow the standard	Does follow the standard
	Financial, legal and reputational risks Customer satisfaction score drops significantly. Negative global media news coverage of incident. Potential breach of consumer laws for misleading or deceptive conduct in not honouring the offered discount. Potential breach of privacy laws for the collection of personal information that was not necessary for its functions. Potential complaints made to relevant regulatory bodies for unlawful discrimination based on a protected attribute (gender).	

Example 2: Facial recognition technology

EcoRetail background

EcoRetail has 20 permanent employees and over 100 casual workers across its nationwide chain of 15 stores.

Its brand is heavily tied to advancing social good, including diversity and inclusion.

Their customer base includes people from many different demographic groups.

EcoRetail's AI system vendor, FRTCo Ltd, suggests installing facial recognition technology, which it states can:

- identify known shoplifters and limit losses from shoplifting
- identify other criminal activities (such as physical violence) to support staff safety.

Facial recognition technology (FRT) is a type of AI that remotely captures sensitive biometric data to verify, identify or analyse people. This functionality poses heightened privacy and discrimination risks to human rights.

While there is currently no specific Australian law governing the use of this technology, the Australian Government is considering the need for new guardrails for FRT as part of its broader Privacy Act reform process.

How EcoRetail uses the standard

EcoRetail wants to procure FRT to:

- accurately identify and deter shoplifters
- prevent violence, protecting customers, staff and assets.

They use the guardrails to inform their actions.

Guardrails	Actions
Guardrail 1: Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.	EcoRetail holds discussions with FRTCo Ltd (AI system vendor) to ensure that FRT aligns with business objectives (minimising loss from shoplifting) and strategic goals (act in accordance with Australia's AI Ethics Principles and Australian legislation). To understand how the use of FRT aligns with EcoRetail's organisational strategy and risk appetite, EcoRetail evaluates the following characteristics of the technology and how it will be deployed: • **Spatial context of deployment**: commercial, publicly accessible space. • **Functionality of the FRT**: facial identification – comparing a single face in the store to a large database of many faces to find a match. FRTCo Ltd is unable to provide detail as to where they have obtained the dataset, how representative it is or whether they followed privacy guardrails. • **Performance**: 99% performance accuracy applied to the estimated 300 people per day (foot traffic across all EcoRetail stores) equates to the potential for 3 people per day to be incorrectly identified. • **Outcomes**: the FRT would impact people's rights (including privacy of sensitive information and the potential for arbitrary detainment) and people's ability to access goods and services. • **Free and informed consent**: signs posted at store entry may not be sufficient for express and sufficiently informed consent.
Guardrail 10: Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.	Senior leaders at EcoRetail held consultations with permanent and casual staff to understand how the use of FRTCo Ltd's FRT system might impact them and their customers. During the consultation, staff received FRTCo Ltd's reports on the accuracy of its product. The staff asked if the accuracy rate applied equally across different demographic groups and discovered that the accuracy rate reduces to 95% for particular racial groups. FRTCo Ltd was unable to give any detail of methodologies used to reduce outcomes based on unwanted bias or show the representation of its dataset. Although the staff indicated that they were sometimes concerned for their safety, they did not feel that the potential benefit from the AI system outweighed the level of surveillance.

Outcomes

EcoRetail decided that using FRT would not align with its strategic goals, risk appetite and legal obligations.

Collecting sensitive biometric information posed too great a risk to the organisation from a legal perspective. EcoRetail also recognised that the scale and impact of potential harm to customers, particularly to those incorrectly identified as shoplifters, was too great.

The possibility of reputational damage, exacerbated by potential regulatory activity for discrimination, was likely to have negative commercial outcomes.

Example 3: Recommender engine

TravelCo.com background

TravelCo.com is a global hotel booking app that is paid by commission. Hotels will pay TravelCo.com a fee every time a user clicks on the offer for their hotel.

Hotels are also able to pay a fee so their hotel appears higher up in search results.

To meet shareholder expectations, TravelCo.com wants to increase market share by telling customers that they can get the cheapest possible price for the same hotel using the TravelCo.com app.

Search results rely on recommender engines as an underlying technology. These use AI to analyse an individual's web browsing activities to give content suggestions based on inferences made about their demographic characteristics, behaviours and interests.

TravelCo.com has engaged a company called XYZ to supply their recommender engine.

Case study: Australian Competition and Consumer Commission v Trivago N.V. (No 2) [2022] FCA 417

Trivago stated it could help consumers find the 'best deal' or cheapest price by comparing hotel rates on different websites.

The algorithm driving Trivago's recommender engine did not use the price of the room as the sole factor in ranking search results. Consumers were not aware that another significant factor was the value of the fee paid by the third-party booking site to have its search result ranking improved.

Consumers were frequently not shown the cheapest price for a hotel in their top search result. In some cases, they were overpaying for the hotel listed as compared to other booking sites.

Trivago was ordered to pay $44.7 million in penalties because of the Federal Court finding it had misled consumers.[94]

How TravelCo.com uses the standard

TravelCo.com wants to procure a recommender engine to:

- meet shareholder expectations of increasing market share
- improve capabilities with AI and data analytics.

They use the guardrails to inform their actions.

Guardrails	Actions
Guardrail 2: Establish and implement a risk management process to identify and mitigate risks.	XYZ notifies TravelCo.com of the challenge in providing a real-time 'cheapest price' because of the large and dynamic dataset of hotel pricing. It would take at least 10 seconds to return a search result, which is not in line with customer expectations for instant information. To minimise lag time for the customer, XYZ suggests updating a static version of the data every 3 hours. As a B2C organisation, TravelCo.com identifies the regulatory risk related to consumer law – that advertising cannot be misleading or deceptive. The pricing at the time the customer searches may no longer be the cheapest option, because of changes since the last update.
Guardrail 6: Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.	The recommender engine uses several factors to create rankings of search results, including alignment to TravelCo.com's business model. Another risk identified during the assessment is that the website does not clearly state that ranking of results is influenced by the commercial arrangements TravelCo.com has with the hotels. Customers could assume that the highest ranked result is the cheapest and therefore overpay.

Outcome

TravelCo.com decided to change its advertising materials from 'cheapest' or 'best' price to stating that it provides comparisons only.

TravelCo.com also decided to include a clear and prominent notice with every search that reflects its commercial arrangements with hotels.

Example 4: Warehouse accident detection

ManufaxCo background

ManufaxCo is a manufacturing company that has built an AI system in house called Safe Zone. SafeZone monitors high-risk factory environments for potential safety hazards and alerts staff to hazards in real-time to prevent accidents and keep workers and assets safe.

SafeZone combines computer vision and Natural Language Processing (NLP) technologies. Cameras installed throughout the factory capture real-time video feeds, which AI analyses to detect safety hazards like spills, obstructions, or people entering unsafe zones. The NLP component allows the system to understand and process verbal commands or alerts from workers, creating a more interactive and complete safety monitoring approach.

How ManufaxCo uses the standard

Guardrails	Actions
Guardrail 2: Establish and implement a risk management process to identify and mitigate risks.	ManufaxCo has carried out a *risk assessment* and found a set of concerns. The concerns (effectiveness and reliability, fairness, and privacy) are not an exhaustive list for this AI system. For example, they do not cover concerns about accountability or potential misuse.

Guardrails	Actions
Guardrail 4.2: Commit to specifying, justifying and documenting acceptance criteria needed for the potential harms to be adequately controlled.	For each concern, the accountable owner in ManufaxCo sets *acceptance criteria* to control for the anticipated harms. 1. **Effectiveness and reliability:** system errors are highly impactful – both false positives (which stop work) and false negatives (where an accident may occur). Set appropriate thresholds such as: - fraction of hazards detected (recall) must be greater than 0.9 - frequency of unnecessary stop-works (false discovery rate) must be less than 0.3 - raise an alarm if a camera view is significantly obstructed for more than 20 seconds. The system must fully integrate with existing safety guardrails and communication systems, with no reported compatibility issues during a 2-week trial period. The system must have an uptime of at least 99.5%, as measured over a 3-month period. At least 80% of staff must rate the system's user interface as 'easy to use' by in a user satisfaction survey. 2. **Fairness:** concerns the safety benefits offered by the system may not apply equally to all workers in the environment. For example, if there is a representation problem in the data. The NLP component must correctly understand and process commands or alerts from workers with at least a 90% accuracy rate across different accents and dialects. 3. **Privacy:** worker stakeholders raise concerns about their privacy at work. The system must pass a privacy compliance audit, ensuring adherence to relevant privacy regulations for handling video feeds and worker data. The system is designed and built to meet these criteria. A third-party vendor supplies voice recognition models for controlling the system. The hazard detection model is trained on historical data. Under normal operating conditions, occurrences of hazards may be rare, so controlled simulations of hazards augment the data.

Guardrails	Actions
Guardrail 4.3: Testing of AI systems or models to determine performance and mitigate any risks	ManufaxCo develops a *test plan* to evaluate the system. They acquire the testing data to evaluate against the acceptance criteria under controlled conditions. This includes evaluations specifically for the acceptance criteria: • hazard detection rates – tested using performed simulations for different types of hazard • false positive count – tested on operational data collected during a small pilot under full human oversight • functionality of failure alert system – inducing camera failures or placing obstructions. They design tests to identify implementation errors and system problems: • a team is assigned to design edge cases such as placing equipment to obscure potential hazards • tests are performed to ensure voice control is performing well enough in various working conditions of machinery • interactions with employees are observed to find out whether they are interacting correctly with the system and as it was intended in the initial design and tests. The tests find the system is functioning as intended, with the exception that initial testing reveals a problem with the false positive rate. The system has many false alarms during normal safe operation. The findings are *reported*, summarising the objectives, methods and metrics used. The accountable owners assign the development team to investigate, and they determine that the problem is because of differences in the environment between the training data and the pilot plant (such as lighting, camera angles, wall colours, specific equipment models). They acquire an updated dataset and re-test the system. Over this period, workers using the system report feedback about voice recognition issues, particularly for workers from multicultural backgrounds. The owners address this by acquiring and swapping in a voice recognition model from a different vendor with models that perform well across a more diverse set of accents. The accountable owners review the reporting to confirm the mitigations have been effective, and they approve the system for *deployment*.
Guardrail 4.4: Commit to implementing robust AI system performance monitoring and evaluation, and to ensuring each system remains fit for purpose.	A month into deployment, ManufaxCo's *monitoring* indicates a reliability problem with the system. Timely investigation reveals a camera calibration issue that hardware configuration and updating the computer vision pipeline's preprocessing stage fixes. ManufaxCo then rolls the system out across multiple warehouses. Initially, the system proves effective in identifying common safety hazards, leading to a noticeable reduction in accidents and meeting all its acceptance criteria. However, as the warehouse operations expand to include new types of machinery and materials, the system experiences a dataset shift. It fails to recognise new hazards that were not present in its training data, resulting in several near-miss incidents that are reported through the feedback channels. The accountable owners examining the monitoring recognise this problem, and they assign the development team to address. The development team updates the training dataset again to include the new hazards. The model is updated and re-tested.

Guardrails	Actions
Guardrail 4.5: Commit to regular system audits for ongoing compliance with the acceptance criteria (or justify why audits aren't needed).	Considering the serious safety impacts of the systems, the accountable owner requests another independent internal technical team do an assessment before the final roll out across all warehouses. During the assessment of the design documentation and pilot monitoring logs, the independent assessors identify and recommend better camera placement to minimise chances of blind spots caused by machines and their operators. ManufaxCo applies this recommendation as an update to the existing installed systems and records it as a consideration for any future deployment in other warehouses.
Guardrail 4.1: Commit to a robust process for timely and regular monitoring, evaluation and reporting of AI system performance.	Given this is a complex new system that could have significant safety impacts, accountable owners decide to *audit* the system and its governance in 6 months. At this stage there will be an existing operational track record.

Acknowledgements

The National AI Centre (NAIC) would like to thank:

NAIC Responsible AI Network (RAIN) Partners for their guidance and input into the consultation process for the Voluntary AI Safety Standard:

- Australian Industry (AI) Group
- Australian Information Industry Association (AIIA)
- Australian Institute of Company Directors (AICD)
- Choice
- Committee for Economic Development of Australia (CEDA)
- Governance Institute of Australia
- NAIC Responsible AI at Scale Think Tank
- Tech Council of Australia
- The Ethics Centre
- Thinkplace.

The organisations that contributed time and expertise to review the draft document:

- ACCC
- The Centre for Inclusive Design
- Office for Women
- Digital Transformation Agency
- Standards Australia
- eSafety Commissioner
- Human Rights Commissioner
- Diversity Council of Australia
- National Indigenous Australians Agency
- Kendra Vant
- Creative Technologies Research Lab, UNSW

NAIC partners that contributed extensive knowledge, experience and expertise into the creation and development of the standard:

- CSIRO's Data61
- Human Technology Institute
- Gradient Institute.

Without all the expert input, the development of the standard would not have been possible.

References

[1] C Taylor, J Carrigan, H Noura, S Ungur, J van Halder and G Singh Dandona, *Australia's automation opportunity: Reigniting productivity and inclusive income growth*, McKinsey & Company, 3 March 2019, accessed 12 December 2023.

[2] Department of Industry, Science and Resources, *Australia's AI Ethics Principles*, Australian Government, n.d.

[3] UK Government, *Policy paper: The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023*, UK Government, 1 November 2023.

[4] Attorney-General's Department, *Human rights protections*, Australian Government, n.d.

[5] S Gasson, 'Human-Centered Vs. User-Centered Approaches to Information System Design', *The Journal of Information Technology Theory and Application (JITTA)*, 2023, 5(2):29–46.

[6] D Cirillo, S Catuara-Solarz, C Morey, E Guney, L Subirats, S Mellino, A Gigante, A Valencia, MJ Rementeria, AS Chadha and N Mavridis, 'Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare', *npj Digital Medicine*, 2020, 3, doi:10.1038/s41746-020-0288-5.

[7] Attorney-General's Department, *Australia's anti-discrimination law* [web page], Australian Government, n.d.

[8] ISO (International Organization for Standardization), *AS ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system*, ISO 2023 **and** NIST (National Institute of Science and Technology), *AI Risk Management Framework 1.0*, U.S. Department of Commerce, 2023.

[9] See, for example, the determination by Australian Information Commissioner and Privacy Commissioner Angelene Falk, who found that Clearview AI, Inc. breached Australians' privacy by scraping their biometric information from the web and disclosing it through a facial recognition tool: https://www.oaic.gov.au/newsroom/clearview-ai-breached-australians-privacy#:~:text=Australian%20Information%20Commissioner%20and%20Privacy,through%20a%20facial%20recognition%20tool.

[10] ISO, *AS ISO/IEC 42001:2023*, 4.1, 4.3, 5.1, A.3.2.

[11] ISO, *AS ISO/IEC 42001:2023*, 4.1 note 3.

[12] ISO, *AS ISO/IEC 42001:2023*, 5.1.

[13] ISO, *AS ISO/IEC 42001:2023*, A.3.3.

[14] ISO, *AS ISO/IEC 42001:2023*, A.3.2.

[15] ISO, *AS ISO/IEC 42001:2023*, 5.1, 7.1.

[16] ISO, *AS ISO/IEC 42001:2023*, 5.2.

[17] ISO, *AS ISO/IEC 42001:2023*, 5.2, A.2.4.

[18] ISO, *AS ISO/IEC 42001:2023*, 5.2, A.2.4.

[19] ISO, *AS ISO/IEC 42001:2023*, 6.3.

[20] ISO, *AS ISO/IEC 42001:2023*, 10.2.

[21] ISO, *AS ISO/IEC 42001:2023*, A.6.1.2.

[22] ISO, *AS ISO/IEC 42001:2023*, 4.1.

[23] ISO, *AS ISO/IEC 42001:2023*, A.6.1.3.

[24] NIST *AI Risk Management Framework (AI RMF 1.0),* GOVERN 2.2.

[25] ISO, *AS ISO/IEC 42001:2023*, 7.2.

[26] ISO, *AS ISO/IEC 42001:2023*, 5.1, 7.2.

[27] ISO, *AS ISO/IEC 42001:2023*, 7.3.

[28] ISO, *AS ISO/IEC 42001:2023*, 7.2.

[29] ISO, *AS ISO/IEC 42001:2023*, 6.1 **and** NIST, *AI Risk Management Framework 1.0*, 1.2.2, MAP 1.6.

[30] ISO, *AS ISO/IEC 42001:2023*, 6.1.1.

[31] ISO, *AS ISO/IEC 42001:2023*, 6.1.1, 6.1.2, 6.1.3, 6.1.4.

[32] ISO, *AS ISO/IEC 42001:2023*, 6.1.1, 6.1.2, 6.1.3, 6.1.4.

[33] ISO, *AS ISO/IEC 42001:2023*, 6.1.1.

[34] ISO, *AS ISO/IEC 42001:2023*, 6.1.1, 6.1.2, 8.2

[35] ISO, *AS ISO/IEC 42001:2023*, 6.1.1, 6.1.4, 8.4.

[36] ISO, *AS ISO/IEC 42001:2023*, 6.1.1, 6.1.3, 8.3.

[37] ISO, *AS ISO/IEC 42001:2023*, 8.2, 8.3.

[38] ISO, *AS ISO/IEC 42001:2023*, 6.1.4.

[39] World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector, Insight Report June 2023*, WEF, 2023.

[40] ISO, *AS ISO/IEC 38507:2022*.

[41] ISO, *AS ISO/IEC 42001:2023*, 4.1.1, B.2.3.

[42] ISO, *AS ISO/IEC 42001:2023*, 4.1.1, B.2.3.

[43] Office of the Australian Information Commissioner, *Australian Privacy Principles*, From Schedule 1 of the *Privacy Amendment (Enhancing Privacy Protection) Act 2012*, Australian Government, 2012.

[44] Australian Signals Directorate, *Essential Eight Maturity Model*, Australian Government, 2017.

[45] Australian Signals Directorate, *Essential Eight Maturity Model*.

[46] ISO, *AS ISO/IEC 42001:2023*, A.7.4, A.7.5, A.7.6.

[47] ISO, *AS ISO/IEC 42001:2023*, A.7.3.

[48] ISO, *AS ISO/IEC 42001:2023*, A.8.5.

[49] Office of the Australian Information Commissioner, *Australian Privacy Principles*, From Schedule 1 of the *Privacy Amendment (Enhancing Privacy Protection) Act 2012*.

[50] Office of the Australian Information Commissioner, *Notifiable data breaches* [web page], Australian Government, n.d.

[51] World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector, Insight Report June 2023*.

[52] ISO, *AS ISO/IEC 42001:2023*, 9.1.

[53] ISO, *AS ISO/IEC 42001:2023*, 7.5.1, 7.5.2, 7.5.3.

[54]ISO, *AS ISO/IEC 42001:2023*, 6.2, 8.1, A.6.2.3, A.6.2.4.

[55] ISO, *AS ISO/IEC 42001:2023*, 6.2.

[56] ISO, *AS ISO/IEC 42001:2023*, 6.2.

[57] ISO, *AS ISO/IEC 42001:2023*, 8.1.

[58] NIST, *AI Risk Management Framework 1.0*, MANAGE 1.1.

[59] ISO, *AS ISO/IEC 42001:2023*, 9.1, A.6.2.6.

[60] ISO, *AS ISO/IEC 42001:2023*, A.8.3 **and** NIST, *AI Risk Management Framework 1.0*, GOVERN 5.1, 5.2.

[61] NIST, *AI Risk Management Framework 1.0*, MEASURE 3.3.

[62] ISO, *AS ISO/IEC 42001*:2023, 9.2.

[63] ISO, *AS ISO/IEC 42001:2023*, 9.2.

[64] ISO, *AS ISO/IEC 42001:2023*, 7.1, 7.2.

[65] ISO, *AS ISO/IEC 42001:2023*, 7.2.

[66] ISO, *AS ISO/IEC 42001:2023*, 7.2, B.4.6.

[67] ISO, *AS ISO/IEC 42001:2023*, 10.2.

[68] NIST, *AI Risk Management Framework (AI RMF 1.0),* GOVERN 2.2.

[69] UK Government, *Guidelines for AI procurement*, UK Government, 2020.

[70] ISO, *AS ISO/IEC 42001:2023*, 7.4, A.8.3.

[71] ISO, *AS ISO/IEC 42001:2023*, 9.2.1.

[72] ISO, *AS ISO/IEC 42001:2023*, A.10.3.

[73] ISO, *AS ISO/IEC 42001:2023*, A.8.5.

[74] ISO, *AS ISO/IEC 42001:2023,* A.8.5.

[75] ISO, *AS ISO/IEC 42001:2023*, A.6.2.7.

[76] World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector, Insight Report June 2023*.

[77] ISO, *AS ISO/IEC 42001:2023*, A.8.3.

[78] ISO, *AS ISO/IEC 42001:2023*, 7.4, A.8.3.

[79] ISO, *AS ISO/IEC 42001:2023*, 7.5.1 **and** NIST *AI Risk Management Framework (AI RMF 1.0),* GOVERN 1.6.

[80] ISO, *AS ISO/IEC 42001:2023*, 4.1.

[81] ISO, *AS ISO/IEC 42001:2023*, 4.3, A.9.3, A.9.4.

[82] ISO, *AS ISO/IEC 42001:2023*, A.6.2.7.

[83] ISO, *AS ISO/IEC 42001:2023*, 7.5.

[84] World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector, Insight Report June 2023*.

[85] ISO, *AS ISO/IEC 42001:2023*, 4.2.

[86] See ISO, *AS ISO/IEC 42001:2023*, 4.2.

[87] ISO, *AS ISO/IEC 42001:2023*, 4.2.

[88] ISO, *AS ISO/IEC 42001:2023*, 6.1.4, A.5.3, A.5.4.

[89] ISO, *AS ISO/IEC 42001:2023*, 6.1.4, A.5.4, A.5.5.

[90] ISO, *AS ISO/IEC 42001:2023*, 6.1.4, A.5.3, A.5.4.

[91] UK Government, *Guidelines for AI procurement*.

[92] World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector, Insight Report June 2023*.

[93] LR Lifshitz and R Hung, 'BC Tribunal Confirms Companies Remain Liable for information Provided by AI Chatbot', *American Bar Association Business Law Section* [web page], ABA, 29 February 2024.

[94] Federal Court of Australia, *Australian Competition and Consumer Commission v Trivago N.V. (No 2) [2022] FCA 417*, Federal Court of Australia, 22 April 2022.