Australian Government
Department of Industry, Science and Resources

National Artificial Intelligence Centre

Guidance for AI Adoption: Implementation practices

Detailed guidance for responsible AI governance

v1.0, October 2025

industry.gov.au/NAIC

Introduction

The Australian Government is committed to supporting industry adopt AI responsibly to secure significant benefits for our economy and community. Best-practice industry guidance plays an important role in building the confidence and capability of Australian workers and organisations to adopt and use AI in ways that makes our lives better.

The *Guidance for AI Adoption* forms the core guidance for the responsible adoption of AI across Australia's economy. The guidance details 6 key responsible AI practices. These practices align with Australia's AI Ethics Principles, as well as international standards and governance approaches.

The guidance advises both non-technical and technical audiences:

- ***Foundations***: a high-level, accessible guide to establish the foundations of the 6 responsible AI adoption practices, for organisations such as small and medium-sized enterprises and not-for-profits.
- ***Implementation practices***: comprehensive guidance for governance professionals and technical experts to implement the 6 responsible AI adoption practices. This resource aligns with international standards and incorporates all VAISS practices, making them more streamlined and accessible. It also extends best practices to AI developers.

How to use this guidance

This guidance applies to both developers and deployers of AI systems. Where practices are particularly relevant to either developers or deployers, this is marked with a corresponding (DEP) or (DEV).

Find all the definitions for this guidance in the definitions table.

Six practices for responsible AI adoption

1. Decide who is accountable	Establish end-to-end accountability and robust AI governance
2. Understand impacts and plan accordingly	Ensure stakeholder rights and fair treatment
3. Measure and manage risks	Implement AI specific risk management
4. Share essential information	Ensure appropriate transparency and explainability
5. Test and monitor	Ensure quality, reliability and protection through evaluation and monitoring of AI systems
6. Maintain human control	Integrate meaningful human oversight and control of AI

Why implement this guidance?

Build trust with customers and stakeholders when using AI

Customers and the community want to know that organisations are using AI ethically and responsibly. Implementing good governance practices can help to build trust with stakeholders.

Secure the intended benefits of AI while mitigating the risks

The implementation practices support organisations to focus on the purpose of AI adoption, align activities to strategic goals and integrate responsible AI practices with existing governance mechanisms.

Build the confidence of decision makers and leaders to adopt AI at greater scale

By establishing good governance foundations and practices, organisations adopting AI can more confidently experiment and take risks with new AI-powered initiatives.

Follow a roadmap that can help to navigate a complex governance landscape

The implementation practices align to international standards and regulation, supporting organisations to adopt leading practices globally in responsible AI.

Our approach

Human-centered

We adopt a human-centred approach to AI development and deployment. This is in line with Australia's AI Ethics Principles and Australia's commitment to international declarations such as the Bletchley Declaration. 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 development and/or deployment, a human-centred approach means:

- **Protecting people.** The implementation practices are designed to help leaders and business owners identify, prevent, minimise and remedy a wide range of AI-related risk of harm to their organisation and stakeholders, including consumers, employees and the Australian community. In this guidance, the approach towards protecting the safety of people is grounded in respecting human 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.
- **Upholding diversity, inclusion and fairness.** The implementation practices are 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/or deployment that recognises and balances human goals, relationships and social contexts with the capabilities and limitations of technical systems (Gasson 2023). This guidance offers practical ways to prioritise the needs of humans in the development and/or deployment of AI systems.
- **Developing and deploying trustworthy AI systems to support social licence.** To unlock the greatest possible value from AI, an organisation developing and 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 develop and/or deploy AI systems.

Bias

This guidance 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 (Cirillo et al.2020)

Bias becomes problematic or ‘unwanted’ when it results in *unfavourable* treatment for people or groups. Unfair treatment will also constitute unlawful discrimination in certain areas of public life if that treatment is based on a ‘protected attribute’:

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

Internationally consistent

Recognising that Australia is an open, trading economy, recommended processes and practices in this guidance 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 implementation practices draw on and are aligned with a range of international standards and equivalent best practices. Key examples include the ISO standard on AI management systems, AS ISO/IEC 42001:2023, the US NIST AI Risk Management Framework (RMF) 1.0 and its Generative AI Profile.

To see how this guidance aligns to international standards please refer to the crosswalk. Future versions will reflect changes in the international landscape.

How the implementation practices help to mitigate AI-specific risks

AI systems have specific characteristics that amplify risks

AI systems are composed of AI models and non-AI components, with AI models playing a key role in influencing their characteristics. In this guidance, the term 'AI system' is used to include AI models when the distinction between the two is not critical. However, 'AI system' and 'AI models' will be explicitly distinguished when the difference or emphasis on both is important.

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

- **Narrow AI** systems are developed to perform a specific task. Many AI systems in use today fall into this category. These types of systems can perform well in a narrow range of tasks, 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 (GPAI)** systems are developed to handle a broad range of tasks and are therefore flexible. Their use is not limited to a specific task, so they can be more easily used for purposes their developers may not have considered. Examples include large language models and systems such as Open AI’s ChatGPT series.

Both narrow and GPAI systems are developed and operated differently from traditional software systems. These differences mean that deploying an AI system for a particular task may amplify existing risks or create new 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, and together with other non-AI components, forms an AI system that can perform a variety of tasks. While such AI systems often outperform comparable, traditional software systems, the different development approach means AI systems, in particular the AI models within them, 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 GPAI systems, especially frontier AI, can further amplify risks and pose new risks and harms to an organisation. This is because they are highly complex and not fully understood, even by their developers. They may possess advanced capabilities that are unknown or emergent. GPAI systems have the capability to understand and use software tools and can access other systems and knowledge, enhancing their capabilities in specific deployment contexts. GPAI systems are also highly general, supporting an unlimited number of downstream planned and unplanned use cases, including deliberate and inadvertent misuse. It is impossible to evaluate all possible use cases, making pre-deployment evaluation and testing highly challenging.

For example, a GPAI chatbot system that can generate code could potentially produce malware and autonomously hack into critical systems. Similarly, a GPAI chatbot that can generate realistic images could be used to create deepfakes for impersonation and fabricating non-existent real-world events. While these systems were not designed for such specific purposes, and some guardrails can be implemented to refuse certain tasks, it is difficult to cover all potential misuses.

A proportionate approach to AI harm prevention and mitigation

As with all software, AI systems vary in the level of risk and the type of harm they pose. Some, like an algorithm 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 or the financial impact of over- or under-ordering. Others, like a tool that prioritises job applicants for an interview process or makes financial lending decisions, have potential to create far greater 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.

This guidance supports a risk-based approach to managing AI systems. It does this by supporting organisations – both AI developers and AI deployers – to take proactive steps to identify risk of harms posed by the AI systems they develop, deploy, or rely on.

The implementation practices prioritise safety and the prevention, identification and mitigation of risk of harm to people. This is grounded in an approach that seeks to protect, respect and remedy human rights. By adopting this approach, AI developers and AI deployers, in turn, also prevent and mitigate the risk of harm to their own organisations.

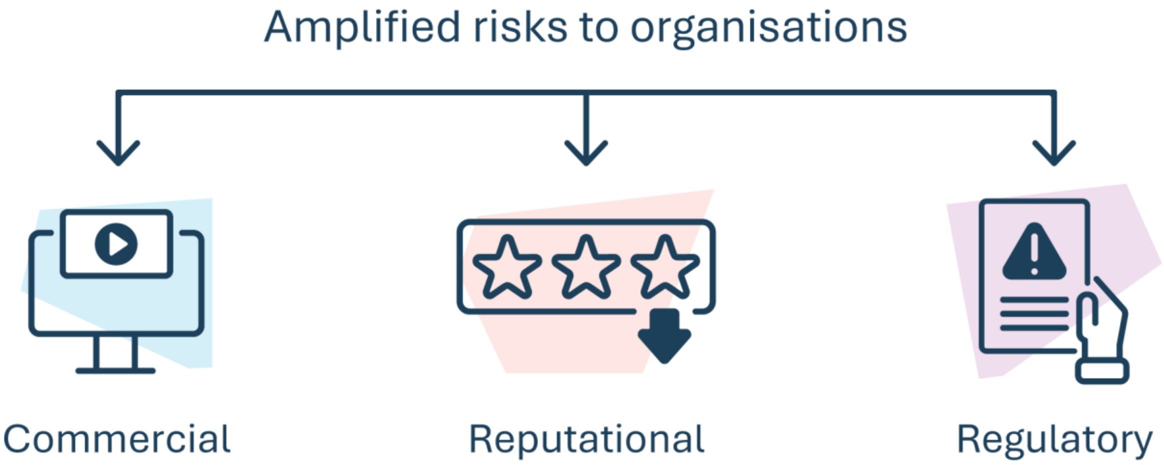
A human-centred perspective on the harms of AI systems

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 job augmentation or 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 generate and spread misinformation could undermine electoral processes, while those that affect educational algorithms could widen the digital divide.

Implementing this guidance can help with identifying, preventing and minimising other risks that may affect an organisation and its stakeholders. Organisations often analyse these risks against the potential for reputational damage, regulatory breach, and commercial losses.

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 stakeholders such as 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 GPAI systems amplifying existing risks. These include why, when, where and how an AI system is deployed. The next section gives examples of important factors to consider when you are designing your approach to high-level risk assessment. For a practical example of how to translate this into a simple process, please refer to the AI screening tool.

We recognise that a single organisation in the AI supply chain may not have full knowledge or control over all these factors. However, the implementation practices encourage organisations to understand the AI systems they develop, deploy, or rely on, and to share relevant information across the supply chain. This will help to identify and mitigate risks more effectively.

AI system attributes and their levels of risk

> This section contains system attribute descriptions and questions to help identify when an attribute may amplify risk. Answering 'yes' to a guiding question indicates a higher level of risk.

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 more careful monitoring and meaningful human oversight. They may be inappropriate for contexts where there is a legal requirement to provide a reason for an output, outcome or decision.

GPAI systems can have higher risks than either narrow AI or traditional software solutions intended for the same task.

Guiding questions (answering 'yes' indicates a higher level of risk)

- Is the way the AI system operates inherently opaque to the developer, deployer, user or affected stakeholder?
- Does it rely on generative AI in ways that can lead to harmful outputs?

Example

A generative AI system is used to create HR related marketing materials.

Purpose

AI systems can considerably outperform traditional software 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.

Guiding questions (answering 'yes' indicates a higher level of risk)

- 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?

Example

A bank uses a risk assessment AI system 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 AI systems used to gather sensitive data about Australians from social media sites.

Guiding questions (answering 'yes' indicates a higher level of risk)

- 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?

Example

A large retailer uses a facial recognition system to identify shoplifters.

Data

AI systems' performance is affected by the quality of data and how accurately that data represents people. Biased 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.

Guiding questions (answering 'yes' indicates a higher level of risk)

- 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 non representative of the people or contexts it is making a decision about?
- Does the dataset produce decisions or outputs which could cause unwanted bias?

Example

An SME deploys a chatbot to confirm customer contact details.

Level of autonomy

Not all automated AI systems are risky. However, systems that operate autonomously, i.e. independent of meaningful human engagement or oversight, may increase risks if they fail or are misused. Risk further increases when there is a considerable period of time between the failure or malicious use happening and the harm being recognised by responsible parties.

Guiding questions (answering 'yes' indicates a higher level of risk)

- Does this system operate autonomously?
- Does the system make decisions without any meaningful human oversight or validation?

Example

A construction site deploys autonomous forklifts to move pallets in a warehouse.

System design

System based on general-purpose LLMs such as GPT5 where decision-making processes cannot be explained or understood, or highly adaptable AI tools that accept open-ended natural language instructions.

Guiding questions (answering 'yes' indicates a higher level of risk)

- Is the AI system designed for multiple purposes or easily adaptable beyond its intended use through interfaces that are not tightly controlled?

AI risks or harms and general laws that may apply

AI risks / harms	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 / harms	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. • Work, health and safety laws where outputs introduce physical or psychosocial risks or harms to workers. • 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	• Intellectual property laws (including copyright), privacy laws, 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 regulate the collection, use and disclosure of, personal information and impose transparency (with specific provisions for some automated decision making to apply from 10 December 2026) and data minimisation requirements on the handling of personal information, and provide for a statutory tort for serious invasions of privacy, which commenced 10 June 2025. • 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

AI risks / harms	General laws that may apply
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, including the Fair Work Act if outputs negatively exclude or disproportionately affect an individual or group on the basis of a protected attribute. Organisations should also ensure they meet obligations in enterprise agreements where applicable.
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.

AI risks / harms	General laws that may apply
Engagement with others in the AI supply chain	• Privacy laws, to be open and transparent in managing personal information, including privacy policies setting out where personal information is collected from or disclosed to third parties. • The *Australian Consumer Law* prohibitions on unfair practices (e.g. misleading and deceptive conduct) and unfair contract terms in how an organisation engages with consumers and other businesses. • The *Australian Consumer Law* statutory guarantees, (e.g. that consumer goods be of acceptable quality and fit for purpose, or that consumer services be rendered with due care and skill) apply to business to business relationships where a party meets the test of a consumer. • Anti-competitive and restrictive trade practices under competition laws, apply to how organisations engage in trade or commerce, including using AI systems to engage in anti-competitive conduct • Product liability, may require manufacturers to indemnify suppliers under the statutory guarantees, and proportional liability laws can restrict the liability of concurrent wrongdoers to their proportionate contribution.

How we developed the guidance

October 2025

In September 2024, we released the Voluntary AI Safety Standard (VAISS). This guidance, published in October 2025, is the first update of the VAISS.

In this update, we have:

- condensed 10 guardrails into 6 essential practices
- removed redundant language
- expanded our audience to developers as well as deployers.

The VAISS was the government's first comprehensive AI governance resource to help organisations develop and deploy AI systems in Australia safely and reliably. It set out best-practice AI governance, focused on AI deployers, that aligned with international standards and regulation.

The technology and governance landscape has shifted rapidly over the past year. The VAISS was intended to be iterative to ensure that this guidance remains fit-for-purpose. At time of release, we flagged that we would update the VAISS to extend best practices to AI developers. The National AI Centre started work on this next version of VAISS in late 2024.

NAIC received extensive feedback throughout 2024–25, as part of consultation extending VAISS practices to developers, including:

- Industry professionals more advanced in AI adoption, and technical experts from a wide range of organisations, valued the VAISS as a useful framework to compare and guide practices and procedures.
- Most industry stakeholders were seeking more accessible, actionable and streamlined guidance which can be tailored to both technical and non-technical audiences, particularly SMEs.
- Specific additional guidance was requested on procuring AI systems as well as transparency mechanisms for AI-generated content, such as watermarking.

In addition, the 2025 Responsible AI Index, released 26 August 2025, surveyed the state of responsible AI across a range of organisations and sectors. The report found that:

- **Responsible AI practice adoption is progressing**: 12% of organisations are now in the Leading category for implementing responsible AI practices, up 4% from 2024.

- **A ‘saying-doing’ gap remains**: while 78% of respondents agreed with ethical AI performance statements, only 29% had implemented relevant responsible AI practices.
- **Smaller organisations face challenges implementing more resource-intensive governance practices**: confidence levels in responsible AI declined for those organisations with 20–99 employees.

This guidance will underpin an expansion of NAIC’s tools and resources. We will roll this out the next 12 months. This integrated approach will build greater coherence and consistency in the delivery of advice to industry.

For organisations that used the VAISS, all practices have now been integrated into the *Implementation practices*. Please refer to the VAISS x Implementation practices crosswalk.

Implementation practices for responsible AI adoption

1. Decide who is accountable

AI systems can make automated decisions that significantly impact people, communities and businesses. Overall, your organisation is ultimately accountable for how and where AI is used, AI complexity can create gaps where no one takes clear responsibility for outcomes.

Accountability is the first step to using AI responsibly.

1.1 Accountable people

> Understanding your role in the supply chain and identifying clear roles for how AI is governed, developed and deployed in the organisation supports accountability and effective oversight.

1.1.1 To ensure AI systems perform as required and obligations are met, assign, document and clearly communicate who is accountable across the organisation (including contractors and third-party providers/systems) for the operation of the AI management system, including:

a) safe and responsible policies, practices and procedures

b) the development and deployment of every AI system, including ongoing human control and oversight

c) oversight of the development and use of AI systems by third parties

d) testing of AI systems across the organisation

e) oversight of concerns, challenges and requests for redress

f) the performance and continual improvement of the AI management system.

1.1.2 For each accountable person, define and communicate the required competencies and their authority. Ensure they are staffed with appropriately skilled people and have the necessary resources.

1.2 Supply chain accountabilities

> Understanding your role in the AI supply chain and identifying which parties are responsible for maintaining the performance, safety and integrity of AI systems throughout their lifecycle is key to effective accountability.

1.2.1 Identify, document and communicate accountability for shared responsibility across the AI supply chain (model developers, system developers and system deployers) for:

a) monitoring and evaluation of model and system performance, quality and safety

b) human oversight and intervention

c) processes to raise issues, faults, failures incidents, contested outcomes, issue resolution and system updates.

1.2.2 Clearly document and communicate the accountability and obligations that developers have towards downstream organisations when integrating, customising, enhancing developer provided AI models or systems. This includes transparency of AI model and system risks, expected behaviours, outcomes under expected use cases and changes to the model or system, paying particular attention to any specific contractual obligations, which could vary by customer (DEV).

1.3 AI literacy and training

> Delivering effective training in AI across the organisation can build confidence, support AI adoption and ensure accountable people have the right capabilities to perform their roles.

1.3.1 Evaluate and document the training needed to build broad AI understanding and a culture of accountability across the organisation. Source or deliver training to bridge any identified gaps. Regularly check skills are up-to-date as AI development and deployment evolves.

1.3.2 Evaluate the training needs of accountable people and provide appropriate up-to-date training to address gaps, such as those responsible for:

a) meeting legal and regulatory obligations

b) handling personally identifiable information

c) operation, control, intervention or termination of each AI system

d) oversight and monitoring of each AI system

e) procurement, development or deployment of third-party AI systems

f) safe and responsible development of AI systems (DEV).

1.4 AI governance framework

> Implementing policies, processes and an overall management system for the development and deployment of AI across the organisation is fundamental to effective and responsible governance of AI.

1.4.1 Document and communicate:

a) the organisation's strategic intent to develop and deploy AI systems in line with organisational strategy and values

b) the regulations relevant to the development and deployment of AI systems and how the organisation will comply

c) appropriately detailed policies, processes and goals for the safe and responsible development and deployment of AI systems which align to the strategy, including:

- an end-to-end process for AI system design and development (DEV)
- goals for AI systems to meet organisational policies for the safe and responsible use of AI
- the consequences for people who act outside of the organisation's policies and defined risk appetite.

1.4.2 Ensure effective operation of the AI management system by:

a) documenting and implementing a process to proactively identify deficiencies in the AI management system. This includes instances of non-compliance in AI systems or in their development or deployment, documenting root causes, corrective action and revisions to the AI management system.

b) appropriately planning changes to the AI management system

c) identifying and documenting the internal and external factors (such as infrastructure or the deployment context) that may affect the organisation's ability to meet its responsibilities through the overarching AI management system

d) providing sufficient resources such as human effort and compute to deploy AI systems safely and responsibly over the lifecycle.

1.4.3 Monitor compliance with organisational policies to identify and address any gaps between leadership expectations and staff understanding of how to develop and deploy AI safely and responsibly.

2. Understand impacts and plan accordingly

Because AI systems can operate at speed and scale, their potential impacts are often magnified. Without careful planning, a single AI system can lead to widespread negative outcomes, such as unfair decisions or the provision of inaccurate information.
For example, AI systems can learn from and amplify existing issues such as unwanted bias in data. This can lead to unfair decisions or inappropriate generated content that could affect many people. If an AI system used for shortlisting in hiring has a bias problem, it could unfairly reject hundreds of qualified candidates before anyone notices.

To use AI responsibly, organisations need to understand, plan for and monitor potential impacts of AI systems. Those affected should be able to raise complaints and get help.

2.1 Identify and engage stakeholders

> Engaging potentially impacted stakeholders is an important way to identify and understand the impacts of AI systems.

2.1.1 Identify and document key types of stakeholders (such as employees and end users) that may be impacted by the organisation's development and deployment of AI, and their needs.

2.1.2 Prioritise, select and document which stakeholder needs will be addressed in organisational policies and procedures.

2.1.3 Document and communicate the organisation's commitment to preventing harms to people from AI models and systems and upholding diversity, inclusion and fairness.

2.1.4 Document the scope for each AI system, including intended use cases, foreseeable misuse, capabilities, limitations and expected context.

2.1.5 For each AI system, engage stakeholders to identify and document the potential benefits and harms to different types of stakeholders, including:

a) impacts to vulnerable groups

b) risks of unwanted bias or discriminatory outputs

c) use of an individual's personal information

d) where the system makes or influences a decision about a person or group of people.

2.1.6 For every documented risk of harm to affected stakeholders, conduct appropriate stakeholder impact analysis.

2.1.7 Monitor for potential harms by engaging affected stakeholders for each AI system on an ongoing basis to identify new stakeholders, including end users throughout the AI lifecycle.

2.1.8 Create processes to support ongoing engagement with stakeholders about their experience of AI systems. Identify vulnerable groups and support appropriately. Equip stakeholders with the skills and tools necessary to give meaningful feedback.

2.2 Establish feedback and redress processes

> Establishing processes for people affected by AI systems to give feedback, ask questions, and challenge decisions easily and safely can ensure issues are identified and resolved.

2.2.1 Create, document and communicate a process for:

a) Potentially affected stakeholders to raise concerns, challenges, or requests for remediation and receive responses (for example, a human rights grievance and remediation mechanism). This includes when and how frequently to communicate, the level of detail to provide and the communication needs of stakeholders, considering the level of AI knowledge and any regulatory requirements.

b) Evaluation of contestability requirements of both internal and external stakeholders and interested parties including accessibility needs.

2.2.2 Implement and document system-level mechanisms to enable contestability of AI use and decisions, enabling stakeholders to understand, challenge and appeal AI use and decisions. These mechanisms must be accessible, understandable and available to users at the appropriate time during interaction with an AI system. Consider mechanisms to share information regarding end user contests and any redress with deployers of AI systems and models (DEV).

2.2.3 Implement and document mechanisms to enable deployers to escalate feedback, report unexpected behaviours, performance concerns or realised harms and support improvements to models or systems (DEV).

2.2.4 Ensure people who monitor and review affected stakeholder feedback can trigger recourse and redress processes where there is an obligation to do so.

2.3 Monitor for systemic issues

> Ongoing monitoring of stakeholder feedback and redress processes can ensure that systemic issues are identified and addressed.

2.3.1 Monitor and evaluate contestability and redress processes to identify and address systemic risks as well as improve effectiveness of these processes.

2.3.2 Create document and communicate a process to review and evaluate stakeholder contests of AI system use across the organisation including any concerns raised by affected stakeholders and requests for information.

3. Measure and manage risks: implement AI-specific risk management

AI risks fundamentally change depending on the type and complexity of your AI systems. Risks often emerge from how the AI system behaves in different situations and use-cases, rather than only from software updates. They can rapidly amplify smaller issues into significant problems.

For example, an AI chatbot that answers simple questions during business hours, when it can be monitored by a staff member, is a low-risk use of AI. The risks expand, however, if that chatbot operates 24/7, without human oversight, and answers more complex questions.

To use AI responsibly, organisations need to be able to identify and manage its risks.

3.1 Establish a fit-for-purpose risk management framework

> An effective risk management framework supports organisations to identify and manage the risks of using AI, set clear rules about what risks are acceptable, and regularly check how AI systems are working over the lifecycle.

3.1.1 Create and document:

a) a risk management framework that addresses the specific characteristics and risks of AI systems

b) organisational-level risk tolerance and criteria to determine acceptable / unacceptable risks for the development and deployment of AI systems. This should include the significance and likelihood of potential harms to affected stakeholders in line with the AI policy and objectives.

c) AI impact assessment, risk assessment and risk treatment processes, including criteria for reassessment over the lifecycle of an AI system. 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.

3.1.2 Ensure that risk management processes include steps to identify, assess and treat risks arising from other parties in the AI supply chain, such as third-party developers and third party deployers. Specific risks relating to open-source AI models, systems and components should be considered by both providers and consumers of these technologies.

3.1.3 Adopt or develop clear and consistent reporting formats, such as data sheets, model cards, or system cards, to communicate appropriate risk management outcomes, including residual risks, to relevant stakeholders (DEV).

3.2 Assess AI system risks

> Using proportionate, robust methods to assess AI system risks is a key part of the operation of the risk management framework.

3.2.1 Establish a triage system to determine which AI systems may pose an enhanced or unacceptable risk, aligned to the organisation's context and risk tolerance (see Foundations Triage template).

3.2.2 Perform and document a risk assessment and evaluation for the specific requirements, characteristics and documented use cases of each AI system, including systems developed or procured from third party suppliers.

3.2.3 In undertaking AI system risk assessments, take the following steps to evaluate the likelihood and consequence of each risk as well as the consequence of not deploying the AI system:

a) Identify potential severity and likelihood of harms to stakeholders, drawing on the Stakeholder Impact Assessment (see 2.1.1 – 2.1.6).

b) Identify legal, commercial and reputational risks such as failing to meet legal obligations, organisational commitments to ESG, diversity, inclusion and accessibility or programs supporting diversity, equity and fairness.

c) Consider the potential amplified and emerging data governance risks across each phase of the AI system lifecycle including before and after model training.

d) Analyze risks systemically using risk models to identify the sources and pathways through which AI systems could produce the identified risks.

e) Compare the estimated value or level of identified risks to pre-determined organisational risk criteria (see 3.1.1) or those defined by regulatory bodies or stakeholders.

f) Document any specific use cases or qualities that represent an unacceptable level of risk to stakeholders or the organisation.

g) Communicate risk assessments in clear reporting formats to relevant stakeholders.

3.3 Implement controls for AI system risks

> Where risks are identified, risk treatment plans make it clear how risks will be mitigated.

3.3.1 Create, document and implement a risk treatment plan to prioritise, select and implement treatment options (e.g. risk avoidance, transfer, acceptance, reduction) and controls to mitigate identified risks. Reassess risks after controls are implemented to verify their effectiveness.

3.3.2 Communicate risk treatment plans in clear reporting formats to relevant stakeholders.

3.3.3 Create and document a deployment plan which includes the response, recovery and communications for the realization of residual risks.

3.3.4 Research, document and implement leading practices in safety measures as safeguards, as appropriate for identified risks (DEV).

3.4 Monitor and report incidents

> Reporting incidents when they happen and communicating the steps you've taken is essential to build trust with stakeholders and meet regulatory obligations.

3.4.1 Track, document and report relevant information about serious incidents and possible corrective measures to relevant regulators and/or the public in a reasonable timeframe. Reporting near-misses and corrective measures is good practice. Communication of corrective measures should consider privacy and cybersecurity risks.

3.4.2 Create and document a process to evaluate and fulfil reporting and disclosure obligations such as those under the Online Safety Act relevant to AI systems usage, including documentation of safety measures implemented such as notices and incident reporting.

3.4.3 Conform to and document data breach reporting requirements and liabilities from related standards. For example, under the Notifiable Data Breaches scheme of the Office of the Australian Information Commissioner.

3.4.4 Maintain two-way communication between developers and deployers for incident reporting, sharing performance insights and coordinating responses to identified issues.

3.4.5 Monitor and evaluate risk assessments and treatment plans on a regular, periodic basis or when a significant change to the use case or the system occurs, or new risks are identified. This includes responding to impact assessments or insufficient risk treatment plans.

3.4.6 Monitor and evaluate the overall effectiveness of risk management processes and continually improve them.

4. Share essential information

People should know when they're interacting with AI and understand when AI decisions affect them. For example, when a customer is receiving services and guidance from a chatbot, they should know this is not a human specialist.

To use AI responsibly, organisations need to tell users and stakeholders when and how they're interacting with AI.

4.1 Maintain an AI register

> An AI register is a central place that records important details about all of the AI systems across the organisation.

4.1.1 Create and maintain an up-to-date, organisation-wide inventory of each AI model and system, with sufficient detail to inform key stakeholders and support future conformance assessments, including:

a) accountable people

b) purpose and business goals

c) capabilities and limitations of the AI model and/or system

d) origin, fine-tuning and updates where applicable

e) technical requirements and components

f) datasets and their provenance used for training and testing

g) acceptance criteria and test results

h) any impact and risk assessments and outcomes

i) identified risks, potential impacts and the risk treatment plan

j) any system audit requirements and outcomes

k) dates of review.

See the AI register template for additional guidance.

4.2 AI system transparency and explainability

> Clearly communicating when and how AI is being developed or deployed by the organisation and explaining the impacts to users and stakeholders is important to build accountability and trust.

4.2.1 Create, document, implement and communicate a policy and process for how to transparently communicate:

a) to AI users, and affected stakeholders that engage directly with AI systems, AI-enabled decisions or AI-generated content about when and how they will be informed about AI development, deployment and use in the organisation.

b) the capabilities, limitations and potential risks of the AI systems that users and affected stakeholders may engage with. This should include when and how frequently to communicate, the level of detail and the level of AI knowledge of AI users and affected stakeholders. It should also address communication obligations and the accessibility needs of AI users and affected stakeholders and incorporate feedback mechanisms where appropriate.

4.2.2 For each AI system, evaluate and document:

a) the transparency requirements for each user and affected stakeholder group (see 2.1.1 and 2.2.1)

b) the transparency and explainability system requirements and measures – including for third-party -provided systems – dependent on use case, stakeholder requirements and risks arising from disclosure.

c) how accessibility obligations and commitments are met by implementing human-centered design.

4.2.3 Create, document and implement organisational processes and transparency mechanisms proportionate to the risks arising from the diverse, evolving, and alternative uses of GPAI beyond predefined applications, including their potential for unexpected and hard-to-explain behaviours. (GPAI DEV/DEP).

4.2.4 Wherever possible choose more interpretable and explainable systems.

4.2.5 Wherever possible, provide reasonably interpretable and explainable AI systems and models to accountable people within the organisation or downstream deployers to enable them to meet their own regulatory obligations (DEV).

4.2.6 Conduct internal testing of the AI model and/or system’s capabilities and limitations. Clearly communicate results to deployers prior to deployment. (DEV).

4.3 Supply chain transparency

> Developers and deployers need to work together to share information and build mechanisms that can clearly communicate information about AI systems to all parties.

4.3.1 Document or request from upstream providers the technical details of the system or model that may be required to meet the needs of users within the organisation or stakeholders.

4.3.2 Share as much of the following information as possible about AI models and systems with downstream deployers (while protecting commercially sensitive information and meeting legal compliance) (DEV):

a) Technical details such as model architecture, description of data, components and their characteristics

b) Test methods, use cases and testing resulting

c) Known limitations, risks and mitigations (such as potential bias and corrective actions) and external audit findings

d) Data management processes for training and testing data including data quality, meta data, and provenance

e) Privacy and cybersecurity practices including conformance to standards and best practice

f) Transparency mechanisms implemented for AI-generated content, interactions and decisions

g) Document and share the following key information for GPAI systems with downstream organisations, stakeholders, researchers and regulators (GPAI DEV):

- Training data sources and compliance details with relevant privacy, intellectual property and copyright laws
- Model cards and system cards, including risk assessment results particularly evaluation of dangerous and emerging capabilities in the deployment and scaffolding context of tool access and agent design.

h) Restricted and managed access to model weights and other associated artefacts.

4.3.3 Share as much of the following information as possible about AI systems with upstream developers (while protecting commercially sensitive information and meeting privacy obligations) (DEP) (See also incident reporting 3.4.1 and 3.4.4).

a) Issues, faults, failures, incidents and any other observed risks that can be addressed by developers

b) Any unexpected and unwanted bias resulting from use of the system.

4.3.4 Ensure you’ve included the required information in contracts with suppliers of systems, including when to update information.

4.4 AI-generated content transparency

> Being clear about when and how content is AI-generated or modified is important to build trust in digital content with stakeholders.

4.4.1 Implement fit-for-purpose and proportionate transparency mechanisms for AI generated content as set out in the *Being clear about AI-generated content* guidance (forthcoming).

5. Test and monitor

AI systems can change their behaviour over time or act in ways that are less predictable than conventional software. For example, an AI system that worked well last month might start giving different answers today if it is trained on additional data.

To use AI safely, organisations should test and monitor their AI systems.

5.1 Pre-deployment testing

> Conducting testing before an AI system is deployed and documenting the outcomes supports ongoing risk mitigation.

5.1.1 Establish oversight mechanisms to review and approve testing methodologies and results, and to monitor system performance, user feedback and operational impacts post-deployment.

5.1.2 Define and communicate clear acceptance criteria and test methodologies that reflect the intended use, context and potential risks (as identified in Essential Practice 3). Conduct predeployment testing.

5.1.3 Clearly document tests and outcomes to support external audits and oversight

5.1.4 When testing is conducted by upstream providers during the development of the AI system and model, request test methodologies and results from the provider and ensure its alignment with your acceptance criteria.

5.1.5 Obtain documented deployment authorisation and rationale from the accountable person for the AI system based on test results.

5.2 Monitor system performance

> Setting performance metrics, closely monitoring and reviewing the performance of AI systems ensures that they operate as intended.

5.2.1 Establish monitoring systems for each AI system to track key performance metrics and indicators relevant to the identified risks.

5.2.2 Implement a deployment process for AI systems that maps business targets to system performance metrics for both internal and third-party developed systems.

5.2.3 Establish and document response processes for addressing all foreseeable issues and harms during system operation.

5.2.4 Establish regular system performance review cycles with stakeholders and subject matter experts to evaluate testing criteria, effectiveness and outcomes.

5.2.5 For each AI system, create and document monitoring requirements including human oversight prior to deployment and evaluate as part of continuous improvement cycle.

5.3 Conduct additional testing proportionate to risk

> Determine whether AI systems require further safety evaluations, independent testing or auditing which are proportionate to their risks.

5.3.1 Conduct safety evaluations that scale with model capabilities (GPAI DEV), for example: assessment for cyber-offensive capabilities and vulnerabilities; testing for potential chemical, biological, radiological and nuclear information risks; evaluation of model behaviours beyond intended use cases; testing for jailbreaking or prompt manipulation; data privacy risks; or comprehensive red teaming to identify vulnerabilities.

5.3.2 For use cases requiring enhanced practices and GPAI systems, conduct independent (internal or external) and thorough review of testing and evaluation methodologies and results. Document and report any issues to the accountable person for the system.

5.3.3 Create a process for and determine whether an AI system requires regular auditing, appropriate to the level of risk identified by its risk assessment. Conduct audits when required.

5.4 Implement robust data and cybersecurity measures

> Effective data governance, privacy and cybersecurity practices are fundamental to supporting the responsible operation of AI systems.

5.4.1 Implement and evaluate the effectiveness of policies and procedures in addressing AI-specific risks and adapt as necessary:

a) Data governance processes covering the use of data with AI models and systems. This includes the management of data usage rights for AI including intellectual property (including copyright), Indigenous Data Sovereignty, privacy, confidentiality and contractual rights.

b) Privacy policies covering the collection, use and disclosure of personal or sensitive information by AI models and systems, including for model training purposes. This needs to support teams to comply with the Australian Privacy Principles for all AI systems.

c) Cybersecurity processes to cover the emerging and amplified risks of AI systems interaction with existing systems and data, such as AI systems unintentionally exposing sensitive information or bypassing security controls. This includes application of the Essential Eight Maturity Model for cybersecurity risks to AI systems.

5.4.2 For each AI use case:

a) Define and document the data quality, data/model provenance and data preparation requirements.

b) Understand and document the data sources and collection processes each AI model / system relies on to function including personal and sensitive data. Put in place systems to manage the data and document the data used to train and test each AI model or systems and data used for inference.

c) Define and document processes for protecting AI models and systems to address emerging cybersecurity and privacy risks (DEV).

d) Where appropriate, report to relevant stakeholders on data, model and system provenance.

e) Document how the Australian Privacy Principles have been applied including in models and systems developed by third parties.

f) Document data usage rights including intellectual property (including copyright), indigenous data sovereignty privacy confidentiality and contractual rights.

g) Monitor for and detect any leakage of personal and sensitive information from AI models and systems.

6. Maintain human control

Unlike traditional software that follows explicit instructions, AI systems learn patterns from data and make their own opaque decision logic. This means they need human oversight to make sure they operate safely. For example, while regular software does exactly what you program it to do, AI might interpret your instructions differently than you intended.

To responsibly use AI, organisations need to make sure a human appropriately oversees any AI systems in use. The person overseeing your AI systems should know how to do so appropriately, and what they need to do to override the system if something goes wrong.

6.1 Maintain human oversight and control

> Ensuring that people in the organisation retain oversight of AI systems, with the ability to intervene where necessary is important to the safe ongoing operation of the system.

6.1.1 Maintain operational accountability, capability and human oversight throughout the lifecycle of AI systems.

6.1.2 Implement mechanisms to enable human control and intervention during the operation of the AI system (DEV).

6.1.3 Implement mechanisms to enable human oversight and intervention to address systemic risks and emerging capabilities such as capability evaluation, training, pause, independent oversight, dynamic guardrails and tool/system access controls (DEV).

6.1.4 Ensure appropriate training is provided to anyone overseeing or using AI systems to understand each system's capabilities, limitations and failure modes and when human intervention is needed.

6.2 Decommission when appropriate

> Establishing processes to decommission AI systems when they are no longer needed or performing as intended can protect ongoing service delivery and data assets.

6.2.1 Define and determine:

a) the criteria or reasons that termination of an AI model or system might need to occur, and at what point intervention should take place

b) the most appropriate role or person to oversee the intervention and decommissioning process

c) whether the model or system is essential to any critical infrastructure or service delivery

- Assess the risks and impacts of shutting down the AI model or system, including impacts on end-users and interdependencies with other integrated systems on which the model, data or outputs rely to function.
- Develop a timeframe and treatment plan to minimise impacts or disruption caused by the decommissioning process.
- Determine a method to extract data and for the return or deletion of assets. Establish which information should be preserved for record keeping purposes.

6.2.2 Create a process for how your organisation will inform relevant parties (such as employees, customers, and upstream or downstream parties) within a reasonable timeframe of the retirement or shutdown of an AI model or system. Establish a channel for people to raise concerns, request support and receive responses.

6.2.3 Determine whether alternative systems or processes will need to be provided to address any issues or gaps.

6.2.4 Maintain alternative pathways for critical functions so operations can continue if AI systems malfunction and/or are taken offline.

Appendix 1 – Terms and definitions

Term	Definition	Source
Accountable / Accountability	Accountable: answerable for actions, decisions and performance. Accountability: state of being accountable.	Aligned ISO 22989:2022 3.5.1 3.5.2
AI agent	Automated entity that senses and responds to its environment and takes actions to achieve goals.	Aligned ISO 22989:2022 3.1.1
AI audit	An internal or (independent) external evaluation of an AI system to determine whether the given system meets the requirements set by a normative framework.	Whittenberg
AI developer / AI development	AI Developer: An organisation or individual that is concerned with the development of AI models, systems, and associated applications, products, and services. AI Model Developer: An organisation or individual that is concerned with activities such as preparing training data, feature engineering, training and fine-tuning, testing, validating AI models. AI System Developer: An organisation or individual that is concerned with activities such as designing, building, testing, training or adapting the overall AI system. This includes integrating AI models with other components such as knowledge or databases, input/output filters, user interfaces, tools, and other systems. Note: A single organisation may play multiple roles, such as AI model developer, AI system developer, and AI system deployer. Note: Organisations or individuals who, design, build, train, adapt, or combine AI systems and applications and distributes or otherwise places it on the market as a service for others to use, whether for payment or free of charge are referred to as an AI provider under the EU AI Act.	Adapted from ISO 22989:2022 5.19.3.2

Term	Definition	Source
AI deployer	An organisation or individual that 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 affect other stakeholders, such as customers. AI deployment is concerned with making an AI model or system available in a specific production environment tailored to particular use cases. Deployment activities often involve customising and integrating AI systems with existing systems and workflows, preparing infrastructure to support operational demands, conducting environment or use case-specific testing, ensuring compliance with security and regulatory standards, creating policies for AI users, and setting up monitoring mechanisms for operations and AI usage. Note: The technical and legal nature, as well as the amount of customisation and integration, may affect whether the activity is system deployment or system development.	
AI lifecycle	The sequence of phases that an AI system goes through, from its conception, all the way through its development, testing, deployment, use, and eventual retirement.	
AI management system	Overarching organisational governance framework for the development and deployment of AI systems, including the policies, objectives and processes to meet objectives. Includes structure, roles and responsibilities, planning and operation.	ISO42001

Term	Definition	Source
AI model	Representation of an entity, phenomenon, process or data, employing various algorithms to interpret, predict, or generate responses based on input. Note: A machine learning model is a type of AI model and is a mathematical construct that generates inferences or predictions based on input data or information. Note: AI models, together with other components, are combined to form AI systems. The inference capability in the AI system which is the key difference with conventional software, comes from its models.	Adapted from ISO 22989:2022 3.1.23
AI safety	Principles and practices to ensure AI is designed, developed, deployed, and used in ways which are human-centric, trustworthy and responsible. This is to realise the potential of AI to help and not harm people; to protect human rights; as well as to promote inclusive economic growth, sustainable development and innovation.	Bletchley Declaration
AI supply chain	Sequence of activities or parties that provides AI products or services to an organisation or individual. Note: In some cases, AI supply chain is used interchangeably with AI value chain.	ISO 26000:2010

Term	Definition	Source
AI system	A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. Note: AI systems integrate AI models with other components such as knowledge or databases, input/output data processes, user interfaces, tools, and other systems. Note: AI models typically need to be converted to an AI system to be deployed and used. For example, by addition of at least some minimal interfaces or input/output data processes.	OECD.AI Policy Observatory
AI user/Use of AI	See AI deployer. Note: In this standard, we use the term AI users refer to organisational users, which largely equates to AI deployers, as the scope of the standard addresses organisational, not individual responsibilities. We note that, in typical software standards, user is often defined as a "person who interacts with a system, product, or service" (ISO 25066:2016) or an "individual or group that interacts with a system or benefits from it during its utilization" (ISO 25010:2011).	Adapted from ISO 25066:2016, ISO 25010:2011
Affected stakeholder	Anyone impacted by the decisions or behaviours of an AI system. These can include organisations, individuals, communities or other systems. For example, consumers, employees and Unions.	
Algorithm	A set of instructions that guide a computer in performing specific tasks or solving problems. Note: A machine learning algorithm is an algorithm used to determine parameters of a machine learning model from data according to given criteria.	ISO 22989:2022

Term	Definition	Source
Bias	Systematic difference in treatment of certain objects, people or groups in comparison to others. Note: From a technical perspective, bias is necessary for AI to identify systematic differences between groups of objects or people and to treat them differently when justified. However, bias becomes problematic or "unwanted" when it leads to unfairness, i.e. unjustified differential treatment that preferentially benefits or harms certain individuals or groups over others.	Aligned ISO 22989:2022 3.5.4
AI end user	Any intended or actual individual or organisation that consumes an AI-based product or service, interacts with it, or is impacted by it after deployment. Affected stakeholder. Note: The term end user is often defined as a “person who directly uses the system for its intended purpose” (ISO 25010:2023) to emphasize direct interaction, or as an “individual person who ultimately benefits from the outcomes of the system or software” (ISO/IEC 25000:2014), highlighting the derived benefits. Since this standard encompasses both benefits and risks, with a focus on affected impacts on a wide range of stakeholders, we adjusted the definition to reflect this.	TGA 2024: Clarifying and strengthening the regulation of AI

Term	Definition	Source
Evaluation	The process of assessing against specific criteria with or without executing the artifacts, including model/system evaluation, capability evaluation, benchmarking, testing, verification, validation, as well as broader risk assessment and impact assessment against criteria or thresholds. AI model evaluation: the process of assessing an AI model against predefined specific criteria or general benchmarks. AI system evaluation: the process of assessing an AI system against predefined specific criteria or general benchmarks. AI capability evaluation: a comprehensive assessment of an AI model or system's overall capabilities, including both planned capabilities and unplanned, emerging, or dangerous capabilities.	Xia et.al 2024 https://arxiv.org/abs/2404.05388v1
Explainability	Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.	Aligned ISO 22989:2022 3.5.7
Fairness	Treatment, behaviour or outcomes that respect established facts, beliefs or norms and are not determined or affected by favouritism or unjust discrimination. Unfairness: unjustified differential treatment that preferentially benefits certain groups more than others.	Aligned ISO TR 24368:2022, ISO TR 24027:2021
General-purpose AI/General AI	AI models or systems developed to handle a broad range of tasks and integrate into a variety of downstream systems or applications.	Adapted from ISO22928 3.1.14

Term	Definition	Source
Generative AI (GenAI)	A type of AI models or systems with the capability to generate synthetic content such as text, images, videos, and other media. Note: Many current GenAI is based on GPAI. However, GenAI can be developed to perform a narrow set of tasks either by restricting GPAI capabilities or via other development approaches. Note: ChatGPT can be considered both a GenAI system and a GPAI system. It is based on the GPT series of models, which are GPAI models (further tuned from foundation models to follow instructions and align with human).	Qinghua et al.
Impact Assessment	A process by which an organisation developing, deploying, or using AI systems identifies, analyses, and evaluates the broader economic, social, and environmental effects of the AI systems on individuals, groups, and societies. Note: Compared to AI risk assessment, AI impact assessment typically considers broader effects beyond the immediate consequences of an AI system. It usually does not incorporate detailed likelihood or probability analysis and focuses directly on affected stakeholders and society. In contrast, risk assessment emphasises the financial, reputational, and legal consequences for the organisation, which are only indirectly linked to	ISO 42001:2022 ISO CD 42005: 2024

Term	Definition	Source
Labelling	Labelling (data): the process of attaching meaningful information (called labels) to pieces of data so that an AI system can learn from them or be tested on them. Datasets are labelled where samples are associated with target variables. Labelling (content): techniques which vary by modality to alert stakeholders to the presence of AI-generated content and its provenance. Note: labelling may take the form of overt watermarks (such as icons overlaid on content, audible disclosures), labels within content (such as warnings, pre-roll or interstitial labels in video and/or audio, or font differences), or user interfaces (such as disclaimers, warnings or symbols to indicate provenance data).	Aligned ISO22989:2022 5.10 Aligned NIST AI 100-4 p30
Measurement (of AI systems)	Employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyse, assess, benchmark, and monitor AI system performance, risk and related impacts.	Aligned NISTL AI RMF Core
Metrics	A qualitative or quantitative measure used to assess, compare, and track the performance or quality of a system, process, product or service. • Internal metrics measure the AI model or system itself (such as model complexity, explainability, training compute resources). • External metrics measure the behaviour and quality of the AI model or system (such as accuracy, response time, scalability). • Risk metrics measure the negative outcomes of using the AI system in a specific context (such as impacts on bias, privacy, security and compliance). • Impact metrics measure the broader effects of AI systems on users, groups and society.	

Term	Definition	Source
Narrow AI	Type of AI system that is focused on defined tasks and uses to address a specific problem.	Aligned ISO 22989:2022 5.2
Performance	Measurable results Note: this can relate to quantitative or qualitative findings, actions or behaviours.	Aligned ISO 22989:2022 3.1.25
Provenance	The logical concept of understanding the history of an asset and its interaction with actors and other assets, as represented by the provenance data.	Aligned C2PA 2.3.8
Red teaming / adversarial testing	An exercise, reflecting real-world conditions, that is conducted as a simulated adversarial attempt to provide a comprehensive assessment of the security capability of the AI system and organisation.	Aligned NIST: CSRC Term
Responsible AI	The practice of developing and using AI systems in a way that provides benefits to individuals, groups, and wider society, while minimising the risk of negative consequences. This includes implementing appropriate governance, oversight and compliance mechanisms.	Adapted Qinghua et al
Risk (of AI system)	Composite measure of an event's probability of occurring and the magnitude of the impacts or consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats.	NIST: AI RMF Core ISO 31000:2018
Risk analysis (of AI system)	The systematic use of risk or threat models to identify sources and pathways through which AI systems could produce risks, and to estimate the level of risks quantitatively or qualitatively.	EU AI Act: Code of Practice
Risk assessment (of AI system)	The systematic process of of risk identification, risk analysis and risk evaluation.	ISO 31073:2022 EU AI Act: Code of Practice
Risk control (of AI system)	Measure that maintains and/or modifies risk. Controls include but are not limited to any process, policy, device, practice or other actions.	ISO 31000: 2018

Term	Definition	Source
Risk evaluation (of AI system)	The process of comparing the estimated value or level of risks from risk analysis to predefined criteria, such as risk thresholds, or risk leves/tiers defined by regulatory bodies and stakeholders or an organisation's risk tolerance.	EU AI Act: Code of Practice
Risk identification (of AI system)	The process of finding, recognising and describing risks. Risk identification involves the identification of hazards, events, and their potential consequences.	ISO 31073:2022 EU AI Act: Code of Practice
Risk mitigation (of AI system)	The process of prioritising, selecting, and implementing appropriate risk-reduction controls. Note: Risk mitigation focuses on risk-reduction controls, while risk treatment includes additional options as well as recovery, response, and communication plans for the realisation of risks.	NIST: CSRC Terms
Risk threshold (of AI system)	The values establishing concrete decision points and operational limits that trigger a response, action, or escalation. They can involve technical indicators (e.g., error rates, scale, training compute) and human values (e.g., social or legal norms) in determining when AI systems present unacceptable risks or risks that demand enhanced scrutiny and mitigation measures.	NIST: AI RMF Core OECD.AI
Risk tolerance	An organisation's or individual's readiness to bear the risk in order to achieve their objectives. Note: It is sometimes used interchangeably with risk appetite—referring to a justified or unjustified attitude as opposed to a readiness to bear risks.	NIST: AI RMF Core

Term	Definition	Source
Risk treatment	The systematic process of prioritising, selecting, and implementing options (e.g., avoidance, transfer, acceptance, reduction) and risk controls to manage and address identified risks. Note: Risk treatment is broader than risk mitigation, as it often involves detailed prioritisation based on impact, probibility, and available resources, along with response, recovery, and communication plans for the realisation of risks.	ISO 23894:2023 NIST: AI RMF Core
Systemic risk (of AI system)	A risk that is specific to the high-impact capabilities of AI, having a significant impact due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.	EU AI Act
Testing	The process of executing an AI model or system to verify and validate that it exhibits expected behaviours across a set of appropriately selected test cases.	IEEE SEBoK
Transparency / transparent	<organisation> Property of an organisation that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner. <system> Property of a system that appropriate information about the system is made available to relevant stakeholders. <mechanism> Process of making information about an AI system or AI-generated content or its provenance available to users and stakeholders.	Adapted ISO 22989:2022 3.5.14 and 3.5.15
Trust (of AI system)	The extent to which a stakeholder is persuaded that the AI will behave as intended.	ISO 25010

Term	Definition	Source
Trustworthiness (of AI system)	An AI system which is deserving of trust due to its ability to meet stakeholder expectations (e.g. reliability, fairness, privacy and security) in a verifiable way.	ISO 22989:2022 3.5.16 Australia’s AI Ethics Principles
Validation/ Validate	Confirmation, through the provision of objective evidence, that the needs of the user have been fulfilled.	IEEE SEBoK
Verification / Verify	Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled.	IEEE SEBoK
Watermark	Information embedded into digital content, either perceptibly or imperceptibly by humans, that can serve a variety of purposes, such as establishing digital content provenance or informing stakeholders that the contents are AI-generated or significantly modified. AI-generated content watermarking: a procedure by which watermarks are embedded into AI-generated content.	Adapted from C2PA 2.4.2

Appendix 2 – Crosswalk VAISS x Guidance for AI adoption: Implementation practices

VAISS V1	Guidance for AI adoption: Implementation Practices
1.1.1	1.1.1 a-f
1.1.2	1.1.2
1.1.3	1.1.1 a
1.1.4	1.1.1 a – f
1.1.5	1.4.2 d
1.1.6	6.1.1
1.2.1	1.1.1 b
1.2.2	1.4.1 a
1.2.3	1.4.1. b
1.2.4	1.4.1 c
1.2.5	1.4.1 c, 1.4.2 b
1.2.6	1.4.2 a
1.2.7	1.4.1 c
1.2.8	1.4.2 c
1.2.9	1.4.1 c
1.2.10	1.2.1
1.3.1	1.2.1
1.3.2	1.2.1, 2.2.1 a
1.3.3	1.4.3
1.3.4	1.4.1 c
1.3.5	1.2.1, 1.2.2
2.1.1	1.3.1 b
2.1.2	3.1.1 b
2.1.3	3.1.1 c
2.1.4	3.1.1 b
2.1.5	3.1.2

VAISS V1	Guidance for AI adoption: Implementation Practices
2.1.6	4.4.1 – 4.4.4
2.1.7	3.1.3
2.2.1	3.2.2
2.2.2	3.2.3 a
2.2.3	3.3.1
2.2.4	3.4.5
2.2.5	2.1.5, 2.1.6, 3.2.3 a - g
3.1.1	5.4.1 a
3.1.2	5.4.1 b
3.1.3	5.4.1 c
3.1.4	5.4.1 b
3.1.5	5.4.1 a
3.1.6	5.4.1 c
3.1.7	5.4.1 c
3.2.1	5.4.2 a
3.2.2	5.4.1 c
3.2.3	5.4.2 b
3.2.4	5.4.2 d
3.2.5	5.4.2 e
3.2.6	5.4.2 f
3.2.7	3.4.3
4.1.1	1.1.1 d, 1.1.2, 5.1.1
4.1.2	5.1.1
4.1.3	See introduction – Documentation and record keeping
4.1.4	5.3.3
4.2.1	5.1.2
4.2.1	5.1.2
4.2.2	5.1.2
4.2.3	5.2.4
4.3.1	5.1.2
4.3.2	5.1.3
4.3.3	5.1.5

VAISS V1	Guidance for AI adoption: Implementation Practices
4.4.1	5.2.1
4.4.2	2.3.1 a
4.4.3	See introduction – Documentation and record keeping
4.4.4	2.3.4
4.5.1	4.1.1 j, 5.3.3
5.1.1	1.1.1 b, 1.1.2
5.1.2	1.1.1 b, 1.1.2
5.1.3	5.2.5
5.1.4	5.2.5
5.1.5	1.1.1 a, 1.1.1 b
5.1.6	1.1.1 c, 1.2.2 e
5.1.7	1.2.2 c
5.1.8	1.2.2 c, 1.2.2 d
6.1.1	4.2.1 a, 4.2.1 b
6.1.2	4.2.1 a
6.1.3	4.2.2 b
6.1.4	4.2.2 b
6.1.5	3.5.2
6.2.1	4.2.2 b
6.2.2	4.2.1 a
6.2.3	4.2.1 a, 4.2.1 b
6.2.4	4.2.2 a, 4.2.2 b
6.2.5	4.6.1 – 4
6.2.6	3.5.2
6.2.7	4.2.1 a, 4.2.1 b
7.1.1	2.3.1 a
7.1.2	2.3.1 a
7.1.3	2.3.1 a, 4.2.1 b
7.1.4	1.1.1 e
7.1.5	2.4.2
8.1.1	4.5.2
8.1.2	4.5.4

VAISS V1	Guidance for AI adoption: Implementation Practices
8.1.3	1.3.1
8.1.4	4.3.4
8.1.5	4.1.1 k
9.1.1	4.1.1
9.2.1	4.1.1
9.2.2	4.1.1
9.2.3	3.1.1 - 4, 3.4.1
9.2.4	4.5.1
9.2.5	5.1.2, 5.1.4
9.2.6	6.1.1, 1.1.1 b
9.2.7	4.1.1
10.1.1	2.1.1
10.1.2	2.1.1
10.1.3	2.1.2
10.1.4	2.1.8
10.2.1	2.1.3
10.2.2	2.1.3
10.2.3	3.2.3 b
10.2.4	2.1.3
10.3.1	2.1.5
10.3.2	2.1.5
10.3.3	2.1.6
10.3.4	4.2.2 c
10.4.1	2.1.4, 3.4.1
10.4.2	3.2.3 b
10.4.3	4.2.2 c
10.4.4	5.2.5