Australian Government
Digital Transformation Agency

dta

Australian Government AI technical standard

Version 1

Contents

Key terms

Term	Definition
AI incident	An event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: a. injury or harm to the health of a person or groups of people; b. disruption of the management and operation of critical infrastructure; c. violations of human rights or a significant breach of obligations under applicable laws, including intellectual property, privacy and Indigenous Cultural and Intellectual Property; d. harm to property, communities or the environment
AI model	'A model is defined as a "physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data" in the ISO/IEC 22989 standard. AI models include, among others, statistical models and various kinds of input-output functions (such as decision trees and neural networks). An AI model can represent the transition dynamics of the environment, allowing an AI system to select actions by examining their possible consequences using the model. AI models can be built manually by human programmers or automatically through, for example, unsupervised, supervised, or reinforcement machine learning techniques.' OECD definition page 8 | oecd.org
AI system	'An Artificial Intelligence (AI) system is 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.' OECD definition | oecd.org
AI watermarking	Information embedded into digital content, either perceptibly or imperceptibly by humans, that can serve a variety of purposes, such as establishing digital content provenance or informing stakeholders that the contents are AI-generated or significantly modified. AI-generated content watermarking: a procedure by which watermarks are embedded into AI-generated content. This embedding can occur at 2 distinct stages: during generation by altering a GenAI model's inference procedure or post-generation, as the content is distributed along the data and information distribution chain.
Algorithm	'A clearly specified mathematical process for computation; a set of rules that, if followed, will give a prescribed result' NIST definition | csrc.nist.gov
Application programming interface (API)	'A system access point or library function that has a well-defined syntax and is accessible from application programs or user code to provide well-defined functionality.' NIST definition | csrc.nist.gov

Term	Definition
Artificial general intelligence (AGI)	‘Artificial general intelligence (AGI), also known as strong AI, is the (currently hypothetical) intelligence of a machine that can accomplish any intellectual task that a human can perform. AGI is a trait attributed to future autonomous AI systems that can achieve goals in a wide range of real or virtual environments at least as effectively as humans can.’ Gartner | gartner.com
Bias	‘systematic difference in treatment of certain objects, people, or groups in comparison to others’ – ISO/IEC 24027 | iso.org
C2PA	The Coalition for Content Provenance and Authenticity, or C2PA | c2pa.org, provides an open technical standard for publishers, creators and consumers to establish the origin and edits of digital content.
Classification model	‘Machine learning model whose expected output for a given input is one or more classes’ ISO/IEC 23053 | iso.org
Data labelling	‘data labelling, in which datasets are labelled, which means that samples are associated with target variables.’ ISO/IEC 22989 | iso.org
Dataset	‘collection of data with a shared format’ ISO/IEC 22989 | iso.org
Explainability	‘property of an AI system to express important factors influencing the AI system results in a way that humans can understand’ ISO/IEC 22989 | iso.org
Fairness	See Guidance 4. Fairness | digital.gov.au
Fine-tuning	‘Model fine-tuning involves adjusting the parameters of foundation models or training models with small datasets for a specific task. This process adapts and enhances the model's performance for particular business needs.’ ITU-T F.748.43 publication | itu.int
Generative AI (GenAI)	‘The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.’ NIST definition | csrc.nist.gov
Grounding	Providing context or relevant knowledge to an AI model by connecting it to trusted data sources at inference time. This does not update the model itself.
Ground truth	‘Value of the target variable for a particular item of labelled input data. The term ground truth does not imply that the labelled input data consistently corresponds to the real-world value of the target variables.’ ISO/IEC 22989 | iso.org
Hallucination	‘Outputs generated by an AI system may not always be accurate or factually correct. Generative AI systems are known to hallucinate information that is not factually correct. Organisational functions that rely on the accuracy of generative AI outputs could be negatively impacted by hallucinations, unless appropriate mitigations are implemented.’ Source: Engaging with artificial intelligence | cyber.gov.au

Term	Definition
Harm	Any adverse effects that would be experienced by an individual (i.e., that may be socially, physically, or financially damaging) or an organization if the confidentiality of PII were breached.' NIST Definition | csrc.nist.gov
Hyperparameters	'characteristic of a machine learning algorithm that affects its learning process Note 1 to entry: Hyperparameters are selected prior to training and can be used in processes to help estimate model parameters.' ISO/IEC 22989 | iso.org
Infrastructure as a service (IaaS)	'The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).' NIST definition | csrc.nist.gov
Large language model (LLM)	Based on artificial neural network technology to take natural language text as input, process it and generate text as output e.g. code generation and content creation.
Machine learning	'Process of optimizing model parameters through computational techniques, such that the model's behavior reflects the data or experience' ISO/IEC 22989 | iso.org
Model dataset	The dataset is used to train an AI model. It is made up of smaller datasets - train dataset, validation dataset and test dataset.
Model explainability	Ability of the model to provide clear and understandable reasons for model outputs to authorised humans.
Model refresh	Update or replace an existing model with a new model
Offline training	'The system is trained during the development process before the system is put into production. This is similar in nature to standard software development, where the system is built and tested fully before it is put into production.' ISO/IEC 22989 | iso.org
Online training	'Online learning / continuous learning – involve the incremental update of the model in the system as it operates during production. The data input to the system during operation is not only analysed to produce an output from the system, but also simultaneously used to adjust the model in the system, with the aim of improving the model on the basis of the production data. Depending on the design of the continuous learning AI system, there can be human actions required in the process, for example data labelling, validating the application of a specific incremental update or monitoring the AI system performance.' ISO/IEC 22989 | iso.org

Term	Definition
Platform as a service (PaaS)	‘The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.’ NIST definition | csrc.nist.gov
Poisoning	‘Adversarial attacks in which an adversary interferes with a model during its training stage, such as by inserting malicious training data (data poisoning) or modifying the training process itself (model poisoning).’ NIST definition | csrc.nist.gov
Prompt engineering	‘Prompt engineering is the discipline of providing inputs, in the form of text or images, to generative AI models to specify and confine the set of responses the model can produce. The inputs prompt a set that produces a desired outcome without updating the actual weights of the model (as done with fine-tuning).’ Gartner | gartner.com
Pre-trained model	‘a component of the training stage in which a model learns general patterns, features, and relationships from vast amounts of unlabeled data, such as through self-supervised learning. Pre-training can equip models with knowledge of general features or patterns which may be useful in downstream tasks, and can be followed with additional training or fine-tuning that specializes the model for a specific downstream task.’ NIST definition | csrc.nist.gov
Regression model	‘machine learning model whose expected output for a given input is a continuous variable’ ISO/IEC 23053 | iso.org
Reliability	‘AI systems reliably operate in accordance with their intended purpose throughout their lifecycle’ – AI Ethics principle | industry.gov.au
Retrieval augmented generation (RAG)	‘RAG enhances LLMs by retrieving relevant information from an external knowledge base and incorporating it into the LLM's generation process’. Computers and Education: Artificial Intelligence | sciencedirect.com
Safety	‘Expectation that a system does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.’ ISO/IEC/IEEE 12207 | iso.org
Semantic versioning	‘Version numbers and the way they change convey meaning about the underlying code and what has been modified from one version to the next.’ Semantic versioning standard | semver.org
Software as a service (SaaS)	‘Software as a service (SaaS) is software that is owned, delivered and managed remotely by one or more providers. The provider delivers software based on one set of common code and data definitions that is consumed in a one-to-many model by all contracted customers at anytime on a pay-for-use basis or as a subscription based on use metrics.’ Gartner | gartner.com
Test dataset	‘data used to assess the performance of a final model’ ISO/IEC 22989 | iso.org.

Term	Definition
Train dataset	‘data used to train a machine learning model’ ISO/IEC 22989 | iso.org.
Transparency	‘property of a system that appropriate information about the system is made available to relevant stakeholders’ ISO/IEC 22989 | iso.org
Validation	‘confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled’ ISO/IEC 22989 | iso.org
Validation dataset	‘data used to compare the performance of different candidate models’ ISO/IEC 22989 | iso.org.
Verification	‘confirmation, through the provision of objective evidence, that specified requirements have been fulfilled’ ISO/IEC 22989 | iso.org
WCAG (Web Content Accessibility Guidelines)	WCAG explains how to make web content more accessible to people with disabilities. Web ‘content’ generally refers to the information in a web page including natural information such as text, images, and sounds, or code or markup that defines structure or presentation. WCAG | w3.org

Introduction

The AI technical standard (the standard) sets consistent practices for government agencies adopting artificial intelligence (AI) systems across the AI lifecycle.

The standard brings together a set of practices for procuring, designing, developing, deploying, and using AI systems. The standard reinforces the Australian Government's AI Ethics Principles | industry.gov.au into a set of technical requirements and guidelines. It complements the Policy for the responsible use of AI in government | digital.gov.au, the AI assurance framework | digital.gov.au, and the Voluntary AI Safety Standard | industry.gov.au.

The standard adopts the OECD definition of an Artificial Intelligence (AI) system | oecd.ai:

'An AI system is 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.'

The standard adopts an agency-first approach. Rather than introducing new processes or duplication, it emphasises the reuse of agency policies, frameworks and practices. Agencies may choose to combine the standard with existing frameworks, such as project governance or data, to include AI-related activities. The standard complements existing frameworks and legislation to help ensure agencies meet their obligations in the use of AI.

The challenges for government use of AI are complex and linked with other governance considerations, such as:

- The Australian Public Service (APS) Code of Conduct
- data governance
- cyber security
- ICT infrastructure
- privacy
- sourcing and procurement
- copyright
- ethics practices.

While not exhaustive, a list of related existing frameworks and related resources | digital.gov.au is provided by the Policy for the responsible use of AI in government.

The practices outlined in this document take the form of standard statements, criteria, and explanatory notes.

1. **Statement**: The standard statements describe ‘what’ needs to be done.
2. **Criteria**: Each statement has at least one criterion to satisfy the statement. Each criterion is ‘required’ or ‘recommended’.
    a. **Required**: Agencies must satisfy criterion marked as required to meet the standard. Required criterion are driven by Australian legislation, regulation and policies, and ethics principles.
    b. **Recommended**: Agencies should implement any criterion marked as recommended.
3. **Explanatory notes**: Explanatory notes are provided for each criterion. These notes are intended to offer guidance rather than serve as a comprehensive checklist. Unless specified, explanatory notes are not mandatory but are intended to support understanding, offer ways to implement the criterion, and provides examples, scenarios, and concepts to guide implementation.

The level of detail and implementation of each statement will vary across use cases. Practical use case guidance has been provided in the Use case assessment section of the standard.

The standard is applicable regardless of whether an agency develops an AI system in-house or contracts an external provider to build or supply it. Engaging external providers does not prevent agencies from implementing each of the criteria in the statements. Agencies that adopt the standard are accountable for ensuring it is met in line with the required and recommended criterion. Transparency documents can be utilised to support assessments, including open-source software.

For early experimentation, proofs of concept, and pilots of AI products and services, the standard should be used to provide guidance for building responsible and safe AI systems, ensuring a clear pathway to production.

The standard helps government:

- contribute to the ethical use of AI to ensure public trust
- stay compliant with regulatory requirements and alignment with AI strategic frameworks
- align with cybersecurity guidelines and the AI assurance framework
- support innovation and whole of economy growth
- support AI sourcing and adoption processes
- provide alignment with international AI best practices across government.

Scope

In scope

The standard applies to:

- AI services and products for administrative decision-making in government
- AI systems that may produce discriminatory, unfair, or harmful outcomes
- platform, data, and software for AI services and products
- a product or service with at least one AI model, hosted internally or externally
- reuse of AI assets, including applying to new or changed use cases
- systems with embedded AI services and products
- publicly available AI tools, such as ChatGPT.

Examples of the types of AI considered for the standard include machine learning, computer vision, deep learning, artificial neural networks, generative AI (GenAI) or any combinations of these.

Out of scope

While the below list is out of scope, agencies can adapt and apply the standard at their own discretion:

- automated decision-making | architecture.digital.gov.au
- robotic process automation
- human-repeatable scripts or processes
- artificial general intelligence
- incidental use of AI.

The standard does not define, but works in conjunction with, the following:

- procurement processes and guidance
- project management methodologies
- risk identification and impact assessment
- incident, problem, and change management.

Target audience

The standard impacts both roles and responsibilities at varying organisational levels across government, and impacts Australians more broadly. The following functions and communities may be impacted or assisted by the standard. Noting individuals may perform multiple roles.

Entities external to government agencies includes the following:

1. **Civil society** can be the owners of the data used by AI systems, or users who are directly or indirectly impacted by AI systems. They can investigate the use of AI in the public sector and publish their findings. Civil society helps shape government's AI policy, programs and strategy to safeguard human rights. They include the public, academia and research, advocacy and media groups.
2. **Oversight bodies** review the extent to which agencies have implemented the standard. They enforce laws, assess compliance, and provide stakeholder confidence. They include the regulators, assurance teams, ethics officers and auditors. They can be internal or external to an agency.
3. **Industry partners** implement the standard for the products or services they provide. They may need to adopt the standard to conform with responsible AI principles and policies for government: They include government suppliers, managed services providers, consultants and contractors. They can range from startups to large corporations, local and international.
4. **International organisations** are interested in global collaboration and advancing interoperability. They include standards bodies, international governments and intergovernmental organisations.

The standard will impact roles and responsibilities at varying organisational levels. The following functions may be impacted and assisted by the standard, noting that individuals may perform multiple roles:

1. **AI leaders in government** shape the future of the public service and the safe and responsible use of AI across Australia, promoting public trust. They include government leaders, AI accountable officials (AO) | digital.gov.au, executive boards, chief technology officers, chief data officers, chief information officers.
2. **Business leadership teams** identify, prioritise and schedule product features considering the requirements in the standard. Accountable and responsible for AI system deployment. They understand the problem that needs to be solved, the end users and the wider operating environment of AI-enabled systems. They include senior

responsible officers, business owners, product managers, project managers and delivery leads.

3. **Technical leadership teams** translate the standard into system-specific requirements and technical procurement requirements. They are the technical system owners. They design technology solutions for business problems. They ensure alignment with enterprise architecture principles and patterns. They include the technical leads, enterprise architects and system analysts. They design technology solutions for business problems.
4. **Development teams** apply the standard to solution from concept and prototyping, design, implementation, integration and testing of AI systems. They include the AI scientists, data engineers, data labellers, software developers, application developers, infrastructure engineers, user and customer experience specialists, test specialists, integrators, and cybersecurity.
5. **Operations teams** apply the standard on continuous deployment, testing, and monitoring of an AI system. They operate AI systems and ensure reliable delivery of services. They include service delivery representatives, technical support, security operations, system administrators, hosting engineers, network engineers, DevOps engineers and maintenance personnel. They need to understand the capability and limitations of the AI systems they are deploying, operating, or using.

AI system lifecycle

The practices described in the standard use a reference AI lifecycle model to ensure holistic coverage of an AI system from inception to retirement, as shown in Figure 1.

The statements and criteria outlined in this standard are structured according to the relevant lifecycle stages and are intended to be implemented through an iterative process.

The AI system lifecycle shown in Figure 1, is a structured process that occurs in stages, ensuring the holistic coverage of the AI system from discovery to retirement.

The AI lifecycle stages include:

1. **Discover:** Design, data, train and evaluate.
2. **Operate:** Integrate, deploy and monitor.
3. **Retire:** Decommission.

This lifecycle model is based on the Voluntary AI Safety Standard | industry.gov.au.

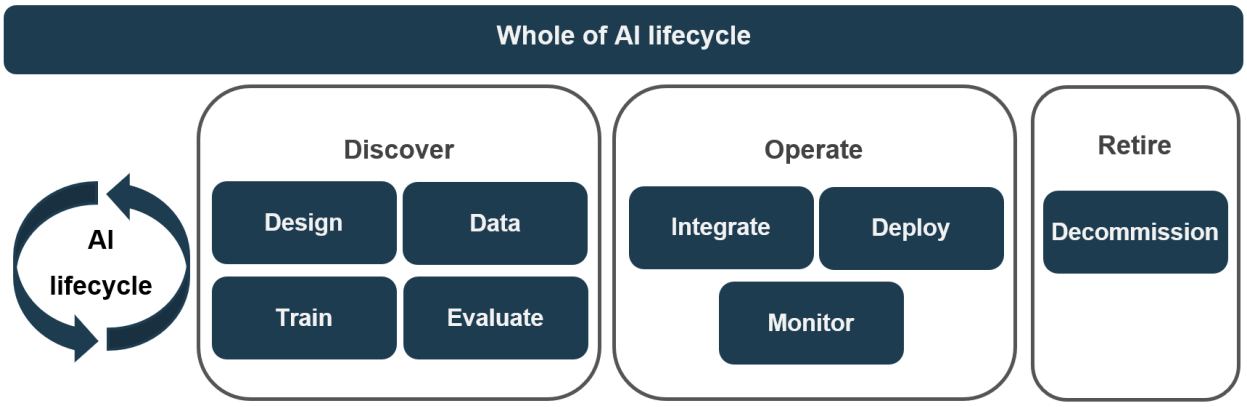

Figure 1: AI system lifecycle model

AI system development is generally an iterative approach. At any point of the lifecycle, issues, risks, or opportunities may be discovered for improvement that could prompt changes to system requirements, design, data, model, or test cases. After deployment, feedback and issues could prompt changes to the requirements.

Each agency may have existing architecture and processes relating to the adoption and implementation of AI systems. The standard complements existing architecture and processes.

The Policy for the responsible use of AI in government | digital.gov.au encourages continuous improvement to enable AI capability uplift.

Summary of requirements in the standard

The statements and criteria in the standard are organised by AI lifecycle stage, including those that apply across all lifecycle stages.

Whole of AI lifecycle

Statement 1: Define an operational model

Recommended

- **Criterion 1**: Identify a suitable operational model to design, develop, and deliver the system securely and efficiently.
- **Criterion 2**: Consider the technology impacts of the operating model.
- **Criterion 3**: Consider technology hosting strategies.

Statement 2: Define the reference architecture

Required

- **Criterion 4**: Evaluate existing reference architectures.

Recommended

- **Criterion 5**: Monitor emerging reference architectures to evaluate and update the AI system.

Statement 3: Identify and build people capabilities

Required

- **Criterion 6**: Identify and assign AI roles to ensure a diverse team of business and technology professionals with specialised skills.
- **Criterion 7**: Build and maintain AI capabilities by undertaking regular training and education of end users, staff, and stakeholders.

Recommended

- **Criterion 8**: Mitigate staff over reliance, under reliance, and aversion of AI.

Statement 4: Enable AI auditing

Required

- **Criterion 9**: Provide end-to-end auditability.
- **Criterion 10**: Perform ongoing data-specific checks across the AI lifecycle.
- **Criterion 11**: Perform ongoing model-specific checks across the AI lifecycle.

Statement 5: Provide explainability based on the use case

Required

- **Criterion 12**: Explain the AI system and technology used, including the limitations and capabilities of the system.

Recommended

- **Criterion 13**: Explain outputs made by the AI system to end users.
- **Criterion 14**: Explain how data is used and shared by the AI system.

Statement 6: Manage system bias

Required

- **Criterion 15**: Identify how bias could affect people, processes, data, and technologies involved in the AI system lifecycle.
- **Criterion 16**: Assess the impact of bias on your use case.
- **Criterion 17**: Manage identified bias across the AI system lifecycle.

Statement 7: Apply version control practices

Required

- **Criterion 18**: Apply version management practices to the end-to-end development lifecycle.

Recommended

- **Criterion 19**: Use metadata in version control to distinguish between production and non-production data, models, and code.
- **Criterion 20**: Use a version control toolset to improve useability for users.
- **Criterion 21**: Record version control information in audit logs.

Statement 8: Apply watermarking techniques

Required

- **Criterion 22**: Apply visual watermarks and metadata to generated media content to provide transparency and provenance, including authorship.
- **Criterion 23**: Apply watermarks that are WCAG compatible where relevant.
- **Criterion 24**: Apply visual and accessible content to indicate when a user is interacting with an AI system.

Recommended

- **Criterion 25**: For hidden watermarks, use watermarking tools based on the use case and content risk.
- **Criterion 26**: Assess watermarking risks and limitations.

Design

Statement 9: Conduct pre-work

Required

- **Criterion 27**: Define the problem to be solved, its context, intended use, and impacted stakeholders.
- **Criterion 28**: Assess AI and non-AI alternatives.
- **Criterion 29**: Assess environmental impact and sustainability.
- **Criterion 30**: Perform cost analysis across all aspects of the AI system.
- **Criterion 31**: Analyse how the use of AI will impact the solution and its delivery.

Statement 10: Adopt a human-centred approach

Required

- **Criterion 32**: Identify human values requirements.
- **Criterion 33**: Establish a mechanism to inform users of AI interactions and output, as part of transparency.
- **Criterion 34**: Design AI systems to be inclusive, ethical, and meets accessibility standards using appropriate mechanisms.
- **Criterion 35**: Design feedback mechanisms.

- **Criterion 36**: Define human oversight and control mechanisms.

Recommended

- **Criterion 37**: Involve users in the design process.

Statement 11: Design safety systemically

Required

- **Criterion 38**: Analyse and assess harms.
- **Criterion 39**: Mitigate harms by embedding mechanisms for prevention, detection, and intervention.

Recommended

- **Criterion 40**: Design the system to allow calibration at deployment.

Statement 12: Define success criteria

Required

- **Criterion 41**: Identify, assess, and select metrics appropriate to the AI system.

Recommended

- **Criterion 42**: Reevaluate the selection of appropriate success metrics as the AI system moves through the AI lifecycle.
- **Criterion 43**: Continuously verify correctness of the metrics.

Data

Statement 13: Establish data supply chain management processes

Required

- **Criterion 44**: Create and collect data for the AI system and identify the purpose for its use.
- **Criterion 45**: Plan for data archival and destruction.

Recommended

- **Criterion 46**: Analyse data for use by mapping the data supply chain and ensuring traceability.
- **Criterion 47**: Implement practices to maintain and reuse data.

Statement 14: Implement data orchestration processes

Required

- **Criterion 48**: Implement processes to enable data access and retrieval, encompassing the sharing, archiving, and deletion of data.

Recommended

- **Criterion 49**: Establish standard operating procedures for data orchestration.
- **Criterion 50**: Configure integration processes to integrate data in increments.
- **Criterion 51**: Implement automation processes to orchestrate the reliable flow of data between systems and platforms.
- **Criterion 52**: Perform oversight and regular testing of task dependencies.
- **Criterion 53**: Establish and maintain data exchange processes.

Statement 15: Implement data transformation and feature engineering practices

Recommended

- **Criterion 54**: Establish data cleaning procedures to manage any data issues.
- **Criterion 55**: Define data transformation processes to convert and optimise data for the AI system.
- **Criterion 56**: Map the points where transformation occurs between datasets and across the AI system.
- **Criterion 57**: Identify fit-for-purpose feature engineering techniques.
- **Criterion 58**: Apply consistent data transformation and feature engineering methods to support data reuse and extensibility.

Statement 16: Ensure data quality is acceptable

Required

- **Criterion 59**: Define quality assessment criteria for the data used in the AI system.

Recommended

- **Criterion 60**: Implement data profiling activities and remediate any data quality issues.
- **Criterion 61**: Define processes for labelling data and managing the quality of data labels.

Statement 17: Validate and select data

Required

- **Criterion 62**: Perform data validation activities to ensure data meets the requirements for the system's purpose.
- **Criterion 63**: Select data for use that is aligned with the purpose of the AI system.

Statement 18: Enable data fusion, integration and sharing

Recommended

- **Criterion 64**: Analyse data fusion and integration requirements.
- **Criterion 65**: Establish an approach to data fusion and integration.
- **Criterion 66**: Identify data sharing arrangements and processes to maintain consistency.

Statement 19: Establish the model and context dataset

Required

- **Criterion 67**: Measure how representative the model dataset is.
- **Criterion 68**: Separate the model training dataset from the validation and testing datasets.
- **Criterion 69**: Manage bias in the data.

Recommended

- **Criterion 70**: For generative AI, build reference or contextual datasets to improve the quality of AI outputs.

Train

Statement 20: Plan the model architecture

Required

- **Criterion 71**: Establish success criteria that cover any AI training and operational limitations for infrastructure and costs.
- **Criterion 72**: Define a model architecture for the use case suitable to the data and AI system operation.

- **Criterion 73**: Select algorithms aligned with the purpose of the AI system and the available data.
- **Criterion 74**: Set training boundaries in relation to any infrastructure, performance, and cost limitations.

Recommended

- **Criterion 75**: Start small, scale gradually.

Statement 21: Establish the training environment

Required

- **Criterion 76**: Establish compute resources and infrastructure for the training environment.
- **Criterion 77**: Secure the infrastructure.

Recommended

- **Criterion 78**: Reuse available approved AI modelling frameworks, libraries, and tools.

Statement 22: Implement model creation, tuning, and grounding

Required

- **Criterion 79**: Set assessment criteria for the AI models, with respect to pre-defined metrics for the AI system.
- **Criterion 80**: Identify and address situations when AI outputs should not be provided.
- **Criterion 81**: Apply considerations for reusing existing agency models, off-the-shelf, and pre-trained models.
- **Criterion 82**: Create or fine-tune models optimised for target domain environment.

Recommended

- **Criterion 83**: Create and train using multiple model architectures and learning strategies.

Statement 23: Validate, assess, and update model

Required

- **Criterion 84**: Set techniques to validate AI trained models.
- **Criterion 85**: Evaluate the model against training boundaries.
- **Criterion 86**: Evaluate the model for bias, implement and test bias mitigations.

Recommended

- **Criterion 87**: Identify relevant model refinement methods.

Statement 24: Select trained models

Recommended

- **Criterion 88**: Assess a pool of trained models against acceptance metrics to select a model for the AI system.

Statement 25: Implement continuous improvement frameworks

Required

- **Criterion 89**: Establish interface tools and feedback channels for machines and humans.
- **Criterion 90**: Perform model version control.

Evaluate

Statement 26: Adapt strategies and practices for AI systems

Required

- **Criterion 91**: Mitigate bias in the testing process.
- **Criterion 92**: Define test criteria approaches.

Recommended

- **Criterion 93**: Define how test coverage will be measured.
- **Criterion 94**: Define a strategy to ensure test adequacy.

Statement 27: Test for specified behaviour

Required

- **Criterion 95**: Undertake human verification of test design and implementation for correctness, consistency, and completeness.
- **Criterion 96**: Conduct functional performance testing to verify the correctness of the AI system Under Test (SUT) as per the pre-defined metrics.
- **Criterion 97**: Perform controllability testing to verify human oversight and control, and system control requirements.
- **Criterion 98**: Perform explainability and transparency testing as per the requirements.

- **Criterion 99**: Perform calibration testing as per the requirements.
- **Criterion 100**: Perform logging tests as per the requirements.

Statement 28: Test for safety, robustness, and reliability

Required

- **Criterion 101**: Test the computational performance of the system.
- **Criterion 102**: Test safety measures through negative testing methods, failure testing, and fault injection.
- **Criterion 103**: Test reliability of the AI output, through stress testing over an extended period, simulating edge cases, and operating under extreme conditions.

Recommended

- **Criterion 104**: Undertake adversarial testing (red team testing), attempting to break security and privacy measures to identify weaknesses.

Statement 29: Test for conformance and compliance

Required

- **Criterion 105**: Verify compliance with relevant policies, frameworks, and legislation.
- **Criterion 106**: Verify conformance against organisation and industry-specific coding standards.
- **Criterion 107**: Perform vulnerability testing to identify any well-known vulnerabilities.

Statement 30: Test for intended and unintended consequences

Required

- **Criterion 108**: Perform user acceptance testing (UAT) and scenario testing, validating the system with a diversity of end-users in their operating contexts and real-world scenarios.

Recommended

- **Criterion 109**: Perform robust regression testing to mitigate the heightened risk of escaped defects resulting from changes, such as a step change in parameters.

Integrate

Statement 31: Undertake integration planning

Recommended

- **Criterion 110**: Ensure the AI system meets architecture and operational requirements with the Australian Government Security Authority to Operate (SATO).
- **Criterion 111**: Identify suitable tests for integration with the operational environment, systems, and data.

Statement 32: Manage integration as a continuous practice

Recommended

- **Criterion 112**: Apply secure and auditable continuous integration practices for AI systems.

Deploy

Statement 33: Create business continuity plans

Required

- **Criterion 113**: Develop plans to ensure critical systems remain operational during disruptions.

Statement 34: Configure a staging environment

Recommended

- **Criterion 114**: Ensure the staging environment mirrors the production environment in configurations, libraries, and dependencies for consistency and predictability.
- **Criterion 115**: Measure the performance of the AI system in the staging environment against predefined metrics.
- **Criterion 116**: Ensure deployment strategies include monitoring for AI specific metrics, such as inference latency and AI output accuracy.

Statement 35: Deploy to a production environment

Required

- **Criterion 117**: Apply strategies for phased roll-out.

- **Criterion 118**: Apply readiness verification, assurance checks and change management practices for the AI system.

Recommended

- **Criterion 119**: Apply strategies for limiting service interruptions.

Statement 36: Implement rollout and safe rollback mechanisms

Recommended

- **Criterion 120**: Define a comprehensive rollout and rollback strategy.
- **Criterion 121**: Implement load balancing and traffic shifting methods for system rollout.
- **Criterion 122**: Conduct regular health checks, readiness, and startup probes to verify stability and performance on the deployment environment.
- **Criterion 123**: Implement rollback mechanisms to revert to the last stable version in case of failure.

Monitor

Statement 37: Establish monitoring framework

Recommended

- **Criterion 124**: Define reporting requirements.
- **Criterion 125**: Define alerting requirements.
- **Criterion 126**: Implement monitoring tools.
- **Criterion 127**: Implement feedback loop to ensure that insights from monitoring are fed back into the development and improvement of the AI system.

Statement 38: Undertake ongoing testing and monitoring

Required

- **Criterion 128**: Test periodically after deployment and have a clear framework to manage any issues.
- **Criterion 129**: Monitor the system as agreed and specified in its operating procedures.
- **Criterion 130**: Monitor performance and AI drift as per pre-defined metrics.
- **Criterion 131**: Monitor health of the system and infrastructure.
- **Criterion 132**: Monitor safety.

- **Criterion 133**: Monitor reliability metrics and mechanisms.
- **Criterion 134**: Monitor human-machine collaboration.
- **Criterion 135**: Monitor for unintended consequences.
- **Criterion 136**: Monitor transparency and explainability.
- **Criterion 137**: Monitor costs.
- **Criterion 138**: Monitor security.
- **Criterion 139**: Monitor compliance of the AI system.

Statement 39: Establish incident resolution processes

Required

- **Criterion 140**: Define incident handling processes.
- **Criterion 141**: Implement corrective and preventive actions for incidents.

Decommission

Statement 40: Create a decommissioning plan

Required

- **Criterion 142**: Define the scope of decommissioning activities.
- **Criterion 143**: Conduct an impact analysis of decommissioning the target AI system.
- **Criterion 144**: Proactively communicate system retirement.

Statement 41: Shut down the AI system

Required

- **Criterion 144**: Proactively communicate system retirement.
- **Criterion 146**: Disable computing resources or components specifically dedicated to the AI system.
- **Criterion 147**: Securely decommission or repurpose all computing resources specifically dedicated to the AI system, including individual and shared components.

Statement 42: Finalise documentation and reporting

Required

- **Criterion 148**: Securely decommission or repurpose all computing resources specifically dedicated to the AI system, including individual and shared components.

Use case assessment

The standard was assessed against a selection of use cases across government agencies. Outcomes were collated to identify how the standard can be used across each lifecycle stage.

The assessment considered:

- proofs of concept to those in operation
- the nature of the applications, whether used by internal staff or public facing
- the type of data involved, whether private, public, or a combination of both
- the risk level of the applications, ranging from low to high.

The applicability of the standard varied, based on who built each part of the AI system:

1. Fully built and managed in-house: Involves building AI systems from scratch.
2. Partially built and fully managed in-house: This includes using pre-trained or off-the-shelf models with or without grounding, RAG, and prompt engineering, such as large language models (LLMs) or reusing existing pre-trained machine learning or computer vision models. Note that fine-tuning a model would transfer the responsibility of applying the standard from the vendor to the agency.
3. Largely built and managed externally: Sourcing or procuring an AI system or SaaS product that is managed by a third party or an external provider, such as Copilot.
4. Incidental usage of AI: Using off-the-shelf software with AI as incidental feature. Examples include:
    - AI features built into desktop software such as grammar checks
    - internet search with AI functionality.

Applicability of the statements in the standard was tested against each AI use case. The process determined whether the standard could be applied to the use case or not. In some cases, such as when a pre-trained model is used, the applicability may be conditional. This means that the applicability depends on the use case, vendor responsibility, and how AI is integrated into the environment.

Applicability of the standard has been categorised as:

- **Applicable:** The statements in the standard fully apply to the use case.

- **Conditional:** The statements in the standard are applicable, but their implementation may require agreement with third-party providers or rigorous testing and monitoring. For example, when using GenAI without fine-tuning or grounding, parts of the standard will be implemented by the provider.
- **N/A (not applicable):** The use case falls outside the scope of the standard, and therefore the statements do not apply.

The following table shows the applicability of the standard against each lifecycle phase:

Phase	Built and managed in-house	Partially built and fully managed in-house	Largely built and managed externally	Incidental usage of AI
Whole of AI Lifecycle	Applicable	Applicable	Applicable	N/A
Design	Applicable	Applicable	Applicable	N/A
Data	Applicable	Conditional	Conditional	N/A
Train	Applicable	Conditional	Conditional	N/A
Evaluate	Applicable	Applicable	Conditional	N/A
Integrate	Applicable	Applicable	Conditional	N/A
Deploy	Applicable	Applicable	Conditional	N/A
Monitor	Applicable	Applicable	Applicable	N/A
Decommission	Applicable	Applicable	Applicable	N/A

Whole of AI lifecycle

Whole of AI lifecycle includes statements that apply across multiple AI product lifecycle stages, for ease of use and to minimise content duplication.

Across the lifecycle stages, agencies should consider:

- operational model – to ensure compliance, efficiency, and ethical standards
- reference architecture – to provide structured frameworks that guide the design, development, and management of AI systems
- people capabilities – having the specialised skills required for successful implementation
- auditability – enabling external scrutiny, supporting transparency, and accountability
- explainability – identifying what needs to be explained and when, making complex AI processes transparent and trustworthy
- system bias – maintaining the role of positive bias in delivering meaningful outcomes, while mitigating the source and impacts of problematic bias
- version control – tracking and managing changes to information to inform stakeholder decision-making
- watermarking – to embed visual or hidden markers into generated content so that its creation details can be identified.

NOTE: Agencies must consider intellectual property rights and ownership derived from procured services or datasets used (including general AI outputs) to comply with copyright law | ag.gov.au.

NOTE: Management of bias in an AI system is critical to ensuring compliance with Australia's anti-discrimination law | ag.gov.au.

NOTE: All documents relating to the establishment, design, and governance of an AI implemented solution must be retained to comply with information management legislation | naa.gov.au.

NOTE: Agencies must comply with data privacy and protection practices as per the Australian Privacy Principles | oaic.gov.au.

NOTE: Agencies must consider data and lineage | naa.gov.au compliance with Australian Government regulations.

NOTE: Agencies should refer to the Policy of responsible use of AI in government to implement AI fundamentals training | digital.gov.au for all staff, regardless of their role. To support agencies with their implementation of the Policy, the DTA provides Guidance for staff training on AI | digital.gov.au.

NOTE: Australian Government API guidelines | api.gov.au mandate the use of semantic versioning.

NOTE: Agencies should refer to the Australian parliamentary recommendations | aph.gov.au on AI including risk management, people capabilities, and implement measures for algorithmic bias | humanrights.gov.au.

NOTE: Any infrastructure, both software and hardware, for AI services and solutions must adhere to Australian Government regulations and should consider security as priority as recommended by the Australian Government guidance on AI system Development | cyber.gov.au, Deploying AI systems Securely | cyber.gov.au and Engaging with AI | cyber.gov.au. The recommendations include secure well-architected environments, whether on-premises, cloud-based, or hybrid, to maintain the confidentiality, integrity, and availability of AI services.

NOTE: Agencies using cloud-based systems should refer to Cloud Financial Optimisation (Cloud FinOps) | architecture.digital.gov.au.

NOTE: Agencies must consider security frameworks | cyber.gov.au, controls and practices with respect to the Information security manual (ISM) | cyber.gov.au, Essential Eight maturity model | cyber.gov.au, Protective Security Policy Framework | protectivesecurity.gov.au and Strategies to mitigate cyber security incidents | cyber.gov.au.

NOTE: Reuse digital, ICT, data and AI systems in line with the Australian Government Reuse standard | architecture.digital.gov.au. This includes pre-existing AI assets and components from organisational repositories or open-source platforms.

NOTE: The Budget Process Operational Rules (BPORs) | finance.gov.au mandate that entities must consult with the DTA before seeking authority to come forward for Expenditure Review Committee | directory.gov.au agreement to digital and ICT-enabled New Policy Proposals, to meet the requirements of the Digital and ICT Investment Oversight Framework | digital.gov.au. Digital proposals likely to have financial implications of $30 million or more, may be subject to the ICT Investment Approval Process (IIAP) | digital.gov.au.

NOTE: Management of human, society and environmental impact should ensure alignment with National Agreement on Closing the Gap | closingthegap.gov.au, Working for Women – A Strategy for Gender Equality | genderequality.gov.au, Australia's Disability Strategy 2021-2031 | ndis.gov.au, National Plan to End Gender Based Violence | dss.gov.au, APS Net Zero Emissions by 2030 Strategy | finance.gov.au, Environmentally Sustainable Procurement Policy | finance.gov.au.

NOTE: The DTA oversees sourcing of digital and ICT for the whole of government and provides a suite of policies and guidelines to support responsible procurement practices of agencies, such as the Procurement and Sourcing | architecture.digital.gov.au and Lifecycle | buyict.gov.au guidance. AI model clauses | buyict.gov.au provide guidance for purchasing AI systems.

Statement 1: Define an operational model

Agencies should:

Criterion 1: Identify a suitable operational model to design, develop, and deliver the system securely and efficiently.

Implementing effective operational models for AI systems needs careful consideration to ensure compliance, efficiency, and ethical standards. They also provide tools for traceability, reproducibility, and modularity.

Existing operational models can be used or extended for AI systems. Operational models can streamline the iterative nature of design, and develop and deliver AI systems more securely, efficiently, and reliably. Some examples include:

- Model operations (ModelOps) – set of practices and technologies to streamline lifecycle management for decision models, using interdisciplinary approaches and automation tools
- Machine learning operations (MLOps) – like ModelOps but for machine learning
- Large language model operations (LLMOps) – is a specialised form of MLOps that is distinct from MLOps due to a focus on prompt engineering, managing massive model weights, and specialised fine-tuning techniques
- Data operations (DataOps) – practices to streamline lifecycle management for data using interdisciplinary approaches and automated pipelines

- Development operations (DevOps) – a software development methodology combining software development and ICT operations for streamlined workflows.

The above list contains examples that are at varying levels of abstraction. For example, LLMOps is a type of MLOps as it inherits many of the same properties.

Ensure governance and security are integrated into the operational model.

Criterion 2: Consider the technology impacts of the operating model.

These include:

- the resources required for system development and maintenance, including computational power and data storage
- AI requirements including potential harm and bias, human oversight and intervention, AI model configuration, and pre and post processing options for fine-tuning models
- the data requirements of the system including sourcing and usage, provenance, training, data diversity, data used in pre-trained models, and intellectual property rights.

NOTE: The source of the impacts listed will be tied to selection decisions of the data and model, and additional training applied to the model.

Criterion 3: Consider suitable technology hosting strategies.

The hosting strategy can involve one of the following models:

- Infrastructure as a service (IaaS) – for maximum control and flexibility; generally suitable for complex AI
- Platform as a service (PaaS) – generally for AI experimentation and development; no in-house infrastructure management required
- Software as a service (SaaS) – generally for ready-made AI systems; no in-house infrastructure management and no in-house AI system development required.

The strategy to adopt should consider:

- use case and enterprise needs
- flexibility, scalability, control, computational performance
- AI development and support costs
- customisation

- vendor lock-in
- security and privacy considerations.

Statement 2: Define the reference architecture

The use of a reference architecture provides a structured framework that guides the design, development, and management of an AI system.

Agencies must:

Criterion 4: Evaluate existing reference architectures.

Make use of the Australian Government Architecture | dta.gov.au to:

- consider reusing pretrained models when applicable
- consider whether to build in-house or use off-the-shelf software or services
- embed flexible architecture practices to avoid vendor lock-in
- ensure strategic alignment with government's digital direction
- ensure consistency and interoperability across agencies.

Agencies should:

Criterion 5: Monitor emerging reference architectures to evaluate and update the AI system.

New architectural paradigms are emerging that address complex AI applications, including:

- Large Language Model (LLM) architectures: These architectures focus on deploying and managing large-scale language models. They encompass systems, tools, and design patterns that facilitate the integration of LLMs into applications, ensuring scalability and efficiency.
- AI infrastructure architectures: Conceptualised to streamline the production of AI models, AI factories provide comprehensive guidelines for building high-performance, scalable, and secure data centres dedicated to AI development. These architectures support the end-to-end lifecycle of AI system creation, from development to deployment.

- Generative AI (GenAI) reference architecture: This architecture outlines interfaces and components for GenAI applications, enabling users to interact with AI systems effectively. It emphasises modularity and flexibility, allowing for the integration of various AI functionalities to meet diverse user needs.

Statement 3: Identify and build people capabilities

Agencies must:

Criterion 6: Identify and assign AI roles to ensure a diverse team of business and technology professionals with specialised skills.

Specialist roles may include, noting that an individual may perform one or more of these roles:

- AI accountable official | digital.gov.au: A senior executive accountable for their agency's implementation of the Policy for the responsible use of AI in government | digital.gov.au
- Data scientists and analysts: Professionals who collect, process, and analyse datasets to inform AI models. They will have expertise in statistical analysis which supports the development of reliable AI systems
- AI integration engineers: Professionals responsible for planning, designing and implementing all components requiring integration in an AI system. The role includes reviewing client needs, developing and testing specifications and documenting outputs
- AI and machine language engineers: Specialists who design, build, and maintain AI models and algorithms. They work closely with data scientists to implement scalable AI systems
- AI test engineers: Specialists who verify and validate AI systems against business and technical requirements
- Ethics and compliance officers: Specialists who ensure that AI systems adhere to legal standards and ethical guidelines, mitigating risks associated with AI systems
- Domain experts: Individuals with specialised knowledge in specific fields, such as healthcare or finance, who provide context and insights to ensure that AI systems are relevant and effective within their respective domain.

Criterion 7: Build and maintain AI capabilities by undertaking regular training and education of end users, staff, and stakeholders.

This may involve:

- agencies should provide regular training programs keep staff updated on the latest tools, methodologies, ethical guidelines and regulatory requirements
- consider how to tailor training to the knowledge requirements of each role and provide staff involved in procurement, design, development, testing, and deployment of AI systems with specialised training. For example, individuals responsible for managing and operating AI decision-making systems should undergo specific AI ethics training
- consider tailoring training for people with disability
- consider interactive workshops, simulations, case study walk-throughs and computing sandpit environments to provide more immersive and real-world-like experiences especially for more complex aspects of AI.

Agencies should:

Criterion 8: Mitigate staff over-reliance, under-reliance, and aversion of AI.

This may involve:

- perform periodic technology-specific training, performance assessments, peer reviews, or random audits
- implement a regular feedback loop for incorrect AI outcomes.

Statement 4: Enable AI auditing

Agencies must:

Criterion 9: Provide end-to-end auditability.

End-to-end AI auditability refers to the ability to trace and inspect the decisions and processes involved in the AI system lifecycle. This enables internal and external scrutiny. Publishing audit results enables public accountability, transparency, and trust.

This may include:

- establishing documentation across the AI system lifecycle as agreed with the accountable official. This should demonstrate conformance with the AI technical standard, and compliance with relevant legislation and regulations.

- establishing traceability of decisions and changes from requirements through to operational impacts
- ensuring accessibility, availability, and explainability of technical and non-technical information to assist audits
- ensuring audit logging of the AI tools and systems are configured appropriately

  This may include:

  - enabling or disabling the capture of system inputs and outputs
  - detect and record modifications to the system's operation or performance
  - record who made the modification, under what authority, and the rationale for the modification
  - record system version and any other critical system information.
- reviewing of audit logs
- ensuring independence and avoiding conflict of interest when undertaking AI audits.

Criterion 10: Perform ongoing data-specific checks across the AI lifecycle.

This should address:

- data quality for AI training, capabilities, and limitations
- how data was evaluated for bias
- controls to detect and manage data poisoning
- legislative compliance.

Criterion 11: Perform ongoing model-specific checks across the AI lifecycle.

This should address:

- track and maintain experiments with new models and algorithms to ensure reproducibility, achieving similar model performance with the same dataset
- output flaws such as factually incorrect, nonsensical, or misleading information, which may be referred to as AI hallucinations
- bias and potential harms, such as ensuring fair treatment of all demographic groups
- model explainability
- controls to detect and manage model poisoning
- legislative compliance.

Statement 5: Provide explainability based on the use case

Agencies must:

Criterion 12: Explain the AI system and technology used, including the limitations and capabilities of the system.

AI algorithms and technologies such as deep learning models, are often seen as 'black boxes'. This can make it difficult to understand how they work and the factors that generate outcomes. Providing clear and understandable explanations of AI outputs helps maintain trust and transparency with AI systems.

Explainability on the specific context of the use case ensures clear understanding and reasoning behind AI system output. This supports accountability, trust, and ethical considerations.

This may include:

- explaining the AI system such as:
  - consideration of trade-offs such as cost and performance
  - what changes are made with AI system updates
  - how feedback is used in improving AI system performance
  - whether the AI system is static or learns from user behaviour
  - whether AI techniques would provide clearer explanations and validate AI actions and decisions.
- use cases that are impacted by legislation, regulation, rules, or third-party involvement
- explain how the system operates including situations that require human intervention
- explain technical and governance mechanisms that ensure ethical outcomes from the use of an AI system
- inform stakeholders when changes are made to the system
- persona level explainability adhering to need-to-know principles.
- Agencies should:

Criterion 13: Explain outputs made by the AI system to end users.

This typically includes:

- explaining:
  - AI outputs that have serious consequences
  - how outputs are based on the data used
  - consequences of system actions and user interactions
  - errors
  - high-risk situations
  - avoid explanations that are confusing or misleading.
- using a variety of methods to explain outputs.

Criterion 14: Explain how data is used and shared by the AI system.

This includes:

- how personal and organisational data is used and shared between the AI system and other applications
- who can access the data
- where identified data has been used, or will be used, for AI system training.

Statement 6: Manage system bias

Management of bias and its potential harms of an AI system is critical to ensuring compliance with federal anti-discrimination legislation. Australia's anti-discrimination law | ag.gov.au states:

> *…it is unlawful to discriminate on the basis of a number of protected attributes including age, disability, race, sex, intersex status, gender identity and sexual orientation in certain areas of public life, including education and employment.*

Certain forms of bias, such as affirmative measures for disadvantaged or vulnerable groups, play a constructive role in aligning AI systems to human values, intentions, and ethical principles. At the same time, it's important to identify and address biases that may lead to unintended or harmful consequences. A balanced approach to bias management ensures that beneficial biases are preserved while minimising the impact of problematic ones.

When integrating off-the-shelf AI products, it's essential to ensure they deliver fair and equitable outcomes in the targe operating environment. Conducting thorough bias

evaluations becomes especially important when documentation or supporting evidence is limited.

Agencies must:

Criterion 15: Identify how bias could affect people, processes, data, and technologies involved in the AI system lifecycle.

This includes:

- establishing a bias management plan, outlining how bias will be identified, assessed, and managed across the AI system lifecycle
- checking for systemic bias, which are rooted in societal and organisational culture, procedures, and practices that disadvantage or benefit specific cohorts. These biases manifest in datasets and in the processes throughout the AI lifecycle
- checking for algorithmic bias in decision-making systems, where an output from an AI system might produce incorrect, unfair or unjustified results
- checking for human bias, which can be conscious or unconscious biases in design decisions, data collection, labelling, test selection, or any process that requires judgment throughout the AI lifecycle
- checking for statistical and computational bias, which can occur when data used to train an AI system is not representative of the population
- checking for bias based on the application of AI, such identifying cognitive bias in a computer vision system
- considering intended bias, such as identifying specific circumstances for a person or a group
- considering inherent bias when reusing pre-trained AI models.

  Examples of sources of bias includes:

  - Cognitive bias – systematic human inclinations or reasoning, such as subconscious judgements based on the current norms of individuals. Based on how people interpret and understand information in their surroundings, such as only using data that reinforces an individual's belief
  - Authority bias – tendency to provide greater weighting or consideration of information from an authority source
  - Availability bias – providing undue weighting to information or processes that they are, or have been, actively involved with

- Confirmation bias – tendency to interpret, favour, or seek out information that reinforces a personal belief, value or understanding, such as a political alignment
- Contextual bias – reliance upon unnecessary or irrelevant information which may unduly influence a decision
- In-group or labelling bias – preferential treatment is provided to those who belong in the same group. Adversely, out-group bias is where unfavourable treatment is provided to those who belong in other groups
- Stereotype bias – generalisations about an individual or group of people based on shared characteristics, such as age, gender, or ethnicity
- Anchoring bias – tendency to rely heavily on the first piece of information they receive
- Group think – tendency for people to strive for consensus within a group
- Automation bias – tendency to rely on automated systems and ignore contradictory information made without automation.

Criterion 16: Assess the impact of bias on your use case.

This typically involves:

- identifying stakeholders and the potential harms to them
- identifying existing countermeasures and assessing their effectiveness
- engaging with diverse and multi-disciplinary stakeholders in assessing the potential impacts of bias
- using bias assessment tools relevant to your use case.

Criterion 17: Manage identified bias across the AI system lifecycle.

For off-the-shelf products, AI deployers should ensure that the AI system provides fair outcomes. Evaluating for bias will be critical where insufficient documentation from the off-the-shelf AI model supplier is provided.

This involves:

- engaging multi-disciplinary skillsets and diverse perspectives, including:
  - policy owners, legal, architecture, data, IT experts, program managers, service delivery professionals, subject matter experts
  - people with lived experience, for example people with disability, gender or sexual diversity and people who are culturally and linguistically diverse.

- implementing multiple approaches to reduce automation bias and monitor to detect unwanted bias that might emerge
- identifying bias-specific documentation requirements such as data and model provenance records:
  - document selection criteria for selecting stakeholders, metrics, and other design-related decisions
  - document any discarded requirement, design, data, model, or tests with corresponding rationale
  - document biases that resulted in decommissioning the data, the model, the application, or the system.
- performing periodic context-based bias awareness training for teams
- consideration of lifecycle stage-specific mitigations, including:
  - identify and validate root causes of bias before addressing them
  - identify corrective and preventive actions corresponding to the root causes of bias
  - identify fairness metrics at design. Performance metrics, such as accuracy and precision, aggregated over the entire dataset could hide bias. For example, a cancer detecting device with 90 per cent accuracy averaged across the entire dataset could hide underperformance on a minority population. Disaggregating performance metrics into suitable attributes can detect whether a system performs fairly across demographics, environmental conditions, and other risk factors
  - analyse data for bias and fix issues in the data. See Model and Context dataset section for more information
  - test independence strategy, functional performance testing, fairness testing, and user acceptance testing
  - configure, calibrate, and monitor bias-related metrics during phased roll-out
  - monitor bias-related metrics and unintended consequences during operations. Provide mechanisms for end users to report and escalate experiences of bias
  - audit for how risks of bias are identified, assessed, and mitigated throughout the lifecycle
  - find and use suitable tools that discover and test for unwarranted associations between the AI system outputs and protected input features

- implement bias mitigation techniques after harmful bias has been identified
- implement bias mitigation thresholds that can be configured post-deployment to ensure equity for cohorts, such as people with lived experience.

Statement 7: Apply version control practices

Version control is a process that tracks and manages changes to information such as data, models, and system code. This allows business and technical stakeholders to identify the state of an AI system when decisions are made, restore previous versions, and restore deleted or overwritten files.

AI system versioning can extend beyond traditional coding practices, which manages a package of identifiable code or configuration information. Version control for information such as training data, models, and hyperparameters will need to be considered.

Information across the AI lifecycle, that was used to generate a decision or outcome, must be captured. This applies to all AI products, including low code or no code third-party tools.

Agencies must:

Criterion 18: Apply version management practices to the end-to-end development lifecycle.

Australian Government API | api.gov.au guidelines mandate the use of semantic versioning. They should be enhanced to cater for AI related information and processes.

Version standards should clearly document the difference between production and non-production data, models and code.

This involves applying version management practices to:

- the model, training and operation dataset, data in the AI system, training algorithm, and hyperparameters
- maintaining design documentation outlining the end-to-end AI system state in line with existing organisational control mechanisms
- include point-in-time date and timestamps to data and any changes in data
- authorship, relevant licencing details, and changes since last version
- capturing approvals from accountable officials for workflow and model reviews, datasets used for training, and relevant hyperparameters

- managing any data poisoning and AI poisoning
- data versioning supporting AI interoperability should include the following:
  - consistency: data structures, exchanges, and formats across different sources are well-defined
  - integration: enables data sourced from different sources to be integrated in a seamless manner
  - all documents relating to the establishment, design, and governance of an AI implemented solution must be retained as per the *Archives Act 1983*.

This does not apply to:

- third-party software products, which are subject to existing controls.

Agencies should:

Criterion 19: Use metadata in version control to distinguish between production and non-production data, models, and code.

This includes:

- a simple and transparent way for all users of the system to understand the version of each component at the time a decision was made
- the use of tags in the version number to provide a visual representation of non-production versions without needing direct access to data or source control toolsets
- the use of metadata can also distinguish between different control states where outputs can vary, and core system functionality of the system has not changed.

Criterion 20: Use a version control toolset to improve useability for users.

Version toolsets improve the usability for service delivery and business users, addressing activities such as appeals, Ministerial correspondence, executive briefs, court cases, audit, assurance, privacy, and legislative reviews.

This includes:

- using purpose built in-house or commercial version management products
- storing sufficient information to allow rollback to a previous system state
- considering archival requirements of training data used in a test environment.

Criterion 21: Record version control information in audit logs.

This includes:

- use of a commit hash to identify the control state of all elements, to reduce the volume and complexity of audit log data. A commit hash, is a unique identifier for every single commit in a repository
- recording AI predictions and actions taken
- pro-active data analytics to be processed against the audit logs, to monitor and assess ongoing AI system performance
- where low code or no code third-party tools are used.

Statement 8: Apply watermarking techniques

AI watermarking can be used to embed visual or hidden markers into generated content, so that its creation details can be identified. It provides transparency, authenticity, and trust to content consumers.

Visual watermarks or disclosures provide a simple way for someone to know they are viewing content created by, or interacting with, an AI system. This may include generated media content or GenAI systems.

The Coalition for Content Provenance and Authenticity (C2PA) | c2pa.org is developing an open technical standard for publishers, creators, and consumers to establish the origin and edits of digital content. Advice on the use of C2PA is out of scope for the standard.

Agencies must:

Criterion 22: Apply visual watermarks and metadata to generated media content to provide transparency and provenance, including authorship.

This will only apply where AI generated content may directly impact a user. For instance, using AI to generate a team logo would not need to be watermarked.

Criterion 23: Apply watermarks and metadata that are WCAG compatible where relevant

Criterion 24: Apply visual and accessible content to indicate when a user is interacting with an AI system.

For example, this may include adding text to a GenAI interface so that users are aware they are interacting with an AI system rather than a human.

Agencies should:

Criterion 25: For hidden watermarks, use watermarking tools based on the use case and content risk.

This includes:

- including provenance and authorship information
- encrypting watermarks for high-risk content
- using an existing tool or technique when practicable
- embedding watermarks at the AI training stage to improve their effectiveness and allows additional information such as content modification to be included
- verifying that the watermark does not impact the quality or efficiency of content generation, such as image degradation or text readability
- including data sources, such as publicly available content used for AI training to manage copyright risks, and product details such as versioning information.

Criterion 26: Assess watermarking risks and limitations.

This includes:

- ensuring users understand there is a risk of third parties replicating a visual watermark and to not over-rely on watermarks, such as sourcing content from external sources
- preventing third-party use of watermarking algorithms to create their own content and act as the original content creator
- consider situations where watermarking is not beneficial. For example, watermarking can be visually distracting for decision makers, or when it's overused in low-risk applications
- consider situations where malicious actors might remove or replicate the watermark to reproduce content generated by AI
- managing copyright or trademark risks related to externally sourced data.

Design

The design stage includes concept development, requirements engineering, and solution design.

Designing AI systems that are effective, efficient, and ethical involves being clear on the problem, understanding the impacts of technical decisions, taking a design approach with humans at the centre and having a clear definition of success.

In the design stage agencies consider how the AI system will operate with and impact existing processes, people, data, and technology. This includes considering potential malfunctions and harms.

Without appropriate design an AI system could:

- cause harm due to incorrect information, caused by AI hallucinations, false positives, or false negatives
- be used beyond their purpose
- perpetuate existing injustices
- be misused, misunderstood, or abused
- be susceptible to malfunctions of another interacting system
- experience behaviour and performance issues caused by other external factors.

At the design stage agencies also determine the performance and reliability measures relevant to their AI system's tasks. Considerations when selecting metrics include business, performance, safety, reliability, explainability, and transparency.

NOTE: Under the Digital Experience Policy | digital.gov.au agencies must meet design standards for digital services.

NOTE: The Voluntary AI Safety Standard | industry.gov.au outlines the need to establish and implement a risk management process to identify and mitigate risks.

Statement 9: Conduct pre-work

Agencies must:

Criterion 27: Define the problem to be solved, its context, intended use, and impacted stakeholders.

This includes:

- analysing the problem through problem-solving frameworks such as root cause analysis, design thinking, and DMAIC (define, measure, analyse, improve, control)
- define user needs, system goals and the scope of AI in the system
- identifying and documenting stakeholders, including:
  - internal or external end users, such as APS staff or members of the public
  - indigenous Australians, refer to Framework for Governance of Indigenous Data | niaa.gov.au
  - people with lived experiences, including those defined by religion, ethnicity, or migration status
  - data experts, such as owners of the data being used to train and validate the AI system
  - subject matter experts, such as internal staff
  - the development team, including SROs, architects, and engineers.
- understanding the context of the problem such as interacting processes, data, systems, and the internal and external operating environment
- phrasing the problem in a way that is technology agnostic.

Criterion 28: Assess AI and non-AI alternatives.

This includes:

- starting with the simplest design, experimenting, and iterating
- validate and justify the need for AI by conducting an objective quality evidence assessment
- differentiating parts that could be solved by traditional software from parts that could benefit from AI
- determine why using AI would be more beneficial over non-AI alternatives by comparing KPI's
- considering the interaction of any AI and non-AI components

- considering existing agency solutions, commercial, or open-source off-the-shelf products
- examining capabilities, performance, cost, and limitations of each option
- conducting proofs of concept and pilots to assess and validate the feasibility of each option
- for transformative use cases, consider foundation and frontier models. Foundation models are quite versatile, trained on large data sets, and can be fine-tuned for specific contexts. Frontier models are at the forefront of AI research and development, trained on extensive datasets, and may demonstrate creativity or reasoning.

Criterion 29: Assess environmental impact and sustainability.

Developing and using AI systems may have corresponding trade-offs with electricity usage, water consumption, and carbon emissions.

Criterion 30: Perform cost analysis across all aspects of the AI system.

This includes:

- infrastructure, software, and tooling costs for:
  - acquiring and processing data for training, validation, and testing
  - tuning the AI system to your particular use case and environment
  - internally or externally hosting the AI system
  - operating, monitoring, and maintaining the AI system.
- cost of human resources with the necessary AI skills and expertise.

Criterion 31: Analyse how the use of AI will impact the solution and its delivery.

This includes:

- identifying the type of AI and classification of data required
- identifying the implications of integrating the AI system with existing departmental systems and data, or as a standalone system
- identifying legislation, regulations, and policies.

Statement 10: Adopt a human-centred approach

Agencies must:

Criterion 32: Identify human values requirements.

Human values represent what people deem important in life such as autonomy, simplicity, tradition, achievement, and social recognition.

Refer to human value requirements for AI systems paper. This includes:

- using traditional requirement elicitation techniques such as surveys, interviews, group discussions and workshops to capture relevant human values for the AI use case
- translating human values into technical requirements, which may vary depending on the risk level and AI use case
- reviewing feedback to identify ignored human-values in the AI system
- understanding the hierarchy of human values and emphasising those with higher relevance
- considering social, economic, political, ethical, and legal values when designing AI systems
- considering human values that are domain specific and based on the context of the AI system.

Criterion 33: Establish a mechanism to inform users of AI interactions and output, as part of transparency.

Depending on use case this may include:

- incorporating visual cues on the AI product when applicable
- informing users when text, audio or visual messages addressed to them are generated by AI
- including visual watermarks to identify content generated by AI
- providing transparency on whether a user is interacting with a person, or system
- including a disclaimer on the limitations of the system
- displaying the relevance and currency of the information being provided
- persona level transparency adhering to need-to-know principles

- providing alternate channels where a user chooses not to use the AI system. This may include channels such as a non-AI digital interface, telephony, or paper.

Criterion 34: Design AI systems to be inclusive, ethical, and meets accessibility standards using appropriate mechanisms.

This includes:

- identifying affirmative actions or preferential treatment that apply for any person or specific stakeholder groups
- ensuring diversity and inclusion requirements, and guidelines, are met throughout the entire AI lifecycle
- providing justification to situations such as pro-social policy outcomes
- reviewing and revisiting ethical considerations throughout the AI system lifecycle.

Criterion 35: Define feedback mechanisms.

This includes:

- providing options to users on the type of feedback method they prefer
- providing users with the choice to dismiss feedback
- provide the user with the option to opt-out of the AI system
- ensuring measures to protect personal information and user privacy
- capturing implicit feedback to reflect user's preferences and interactions, such as accepting or rejecting recommendations, usage time, or login frequency
- capturing explicit feedback via surveys, comments, ratings, or written feedback.

Criterion 36: Define human oversight and control mechanisms.

This includes:

- identifying conditions and situations that need to be supervised and monitored by a human, conditions that need to be escalated by the system to a supervisor or operator for further review and approval, and conditions that should trigger transfer of control from the AI system to a supervisor or operator
- defining the system states, errors, and other relevant information that should be observable and comprehensible to an informed human
- defining the pathway for the timely intervention, decision override, or auditable system takeover by authorised internal users

- subsets of inputs and outputs that may result in harm should be recorded for monitoring, auditing, contesting, or validation. This will facilitate reviewing of false positives against inputs that triggered them, and of false negatives that result in harms
- identifying situations where a supervising human might become disengaged and designing the system to attract the operator's attention
- map human oversight and control requirements to corresponding risks they mitigate
- identifying required personas and defining their roles
- adherence to privacy and security need-to-know principles.

Agencies should:

Criterion 37: Involve users in the design process.

The intention is to promote better outcomes for managing inclusion and accessibility by setting expectations at the beginning of the AI system lifecycle.

This includes:

- considering security guidance and the need-to-know principle
- involving users in defining requirements, evaluating, and trialling systems or products.

Statement 11: Design safety systemically

Agencies must:

Criterion 38: Analyse and assess harms.

This includes:

- Utilising functional safety standards that provide frameworks for a systematic and robust harms analysis.

Criterion 39: Mitigate harms by embedding mechanisms for prevention, detection, and intervention.

This includes:

- designing the system to avoid the sources of harm
- designing the system to detect the sources of harm
- designing the system to check and filter its inputs and outputs for harm
- designing the system to check for sensitive information disclosure

- designing the system to monitor faults in its operation
- designing the system with redundancy
- designing intervention mechanisms such as warnings to users and operators, automatic recovery to a safe state, transfer of control, and manual override
- designing the system to log the harms and faults it detects
- designing the system to disengage safely as per requirements
- for physical systems, designing proper protective equipment and procedures for safe handling
- ensuring the system meets privacy security requirements and adheres to the need-to-know principle for information security.

Agencies should:

Criterion 40: Design the system to allow calibration at deployment.

This includes:

- where initial setup parameters are critical to the performance, reliability, and safety of the AI system.

Statement 12: Define success criteria

Agencies must:

Criterion 41: Identify, assess, and select metrics appropriate to the AI system.

Relying on a single metric could lead to false confidence, while tracking irrelevant metrics could lead to false incidents. To mitigate these risks, analyse the capabilities and limitations of each metric, select multiple complementary metrics, and implement methods to test assumptions and to find missing information.

Considerations for metrics includes:

- value-proposition metrics – benefits realisation, social outcomes, financial measures, or productivity measures
- performance metrics – precision and recall for classification models, mean absolute error for regression models, or bilingual evaluation understudy (BLEU) for text generation. This can include summarisation tasks, inception score for image generation models, or mean opinion score for audio generation
- training data metrics – data diversity and data quality related measures

- bias-related metrics – demographic parity to measure group fairness, fairness through awareness to measure individual fairness, counterfactual fairness to measure causality-based fairness
- safety metrics – likelihood of harmful outputs, adversarial robustness, or potential data leakage measures
- reliability metrics – availability, latency, mean time between failures (MTBF), mean time to failure (MTTF), or response time
- citation metrics – measures related to proper acknowledgement and references to direct content and specialised ideas
- adoption-related metrics – adoption rate, frequency of use, daily active users, session length, abandonment rate, or sentiment analysis
- human-machine teaming metrics – total time or effort taken to complete a task, reaction time when human control is needed, or number of times human intervention is needed
- qualitative measures – checking the well-being of the humans operating or using the AI system, or interviewing participants and observing them while using the AI system to identify usability issues
- drift in AI system inputs and outputs - changes in input distribution, outputs, and performance over time.

After metrics have been identified, understand and assess the trade-offs between the metrics.

This includes:

- assessing trade-offs between different success criteria
- determining the possible harms with incorrect output, such as a false positive or false negative
- analysing how the output of the AI system could be used. For example, determine which instance would have greater consequences: a false negative that would fail to detect a cyberattack; or a false positive that incorrectly flags a legitimate user as a threat
- assessing the trade-offs among the performance metrics
- understanding the trade-offs with costs, explainability, reliability, and safety

- understanding the limitations of the selected metric and ensure measures are considered when building the AI system, such as selecting data and training methods
- ensuring trade-offs are documented, understood by stakeholders, and accounted for in selecting AI models and systems
- optimising the metrics appropriate to the use case.

Agencies should:

Criterion 42: Re-evaluate the selection of appropriate success metrics as the AI system moves through the AI lifecycle.

Criterion 43: Continuously verify correctness of the metrics.

Before relying on the metrics, verify the following:

- metrics are accurately reflected when the AI system does not have enough information
- metrics correctly reflect errors, failures, and successful task performance.

Data

The data stage involves establishing the processes and responsibilities for managing data across the AI lifecycle. This stage includes data used in experimenting, training, testing, and operating AI systems.

Data used by an AI system can be classified into development and deployment data.

Development data includes all inputs and outputs (and reference data for GenAI) used to develop the AI system. The dataset is made up of smaller datasets – train dataset, validation dataset, and test dataset.

- Train dataset – this dataset is used to train the AI system. The AI system learns patterns in the train dataset. The train dataset is the largest subset of the modelling dataset. For GenAI, the train dataset may also include reference or contextual datasets such as retrieval-augmented generation (RAG) datasets and prompt datasets
- Validation dataset – this dataset is used to evaluate the model's performance during model training. It is used to fine-tune and select the best-performing model, such as through cross validation
- Test dataset – this dataset is used to evaluate the final model's performance on previously unseen data. This dataset helps provide unbiased evaluation of model performance.

Deployment data includes AI system inputs such as live production data, user input data, configuration data, and AI system outputs such as predictions, recommendations, classifications, logs, and system health data. Deployment stage inputs are new and previously unseen by the AI system.

The performance of an AI system is dependent on robust management of data quality and the availability of data.

Key workstreams within this stage include:

- data orchestration – establishing central oversight of and planning the flow of data to an AI system from across datasets
- data transformation – converting and optimising data for use by the AI system
- feature engineering – methods to improve AI model training to better identify and learn patterns in the data

- data quality – measuring dimensions of a dataset associated with greater performance and reliability
- data validation – testing the consistency, accuracy, and reliability of the data to ensure it meets the requirements of the AI system
- data integration and fusion – combining data from multiple sources to synchronise the flow of data to the AI system
- data sharing – promoting reuse, reducing resources required for collection and analysis, and helping to build interoperability between systems and datasets
- model dataset establishment – using real-world production data to build, refine, and contextualise a high-quality AI model.

NOTE: Requirements for handling personal and sensitive data within AI systems are included in the Privacy Act | oaic.gov.au, the Australian Privacy Principles | oaic.gov.au, Privacy and Other Legislation Amendment Act 2024 | aph.gov.au and the Handling personal information | oaic.gov.au guidance.

NOTE: Data archival and destruction must comply with the Information management legislation | naa.gov.au.

NOTE: The Framework for the Governance of Indigenous Data | niaa.gov.au provides guidelines on Indigenous data sovereignty.

NOTE: The Office of the Australian Information Centre (OAIC) provides Guidelines on data matching in Australian Government administration | oaic.gov.au, which agencies must consider prior to data integration and fusion activities.

NOTE: The Information management for records created using AI technologies | naa.gov.au provides guidelines to manage data for AI.

NOTE: The Data Availability and Transparency Act 2022 | datacommissioner.gov.au (DATA Scheme) requires agencies to identify data as open, shared, or closed.

NOTE: The Guidelines for data transfers | cyber.gov.au provide guidance on the processes and procedures for data transfers and transmissions.

NOTE: The APS Data Ethics Use Cases | finance.gov.au provide guidance for agencies to manage and mitigate data bias.

NOTE: The report on Responding to societal challenges with data | oecd.org provides guidance on data access, sharing, and reuse of data.

Statement 13: Establish data supply chain management processes

Agencies must:

Criterion 44: Create and collect data for the AI system and identify the purpose for its use.

It is important to identify:

- what data will be used and is fit-for-purpose for the AI system
- the sensitivity of the data, such as personal, protected, or otherwise sensitive
- consent provided on usage including when to retain or destroy data, ensuring the proposed uses in the AI system align with the original limits of the consent
- speed and mode of the data supply
- how the data will be used at each stage of the AI system
- where the data will be stored at each stage of the AI system
- changes to the data at different points of the AI system
- methods to manage and monitor data access
- methods to manage any real-time data changes
- data retention policies
- cross-agency or cross-border data governance, if relevant
- any risks and challenges associated with data elements of off-the-shelf AI models, products, or services in the AI system
- cyber supply chain management
- data quality monitoring and remediation
- comprehensive documentation at each stage of the AI system to facilitate traceability and accountability
- adherence to relevant legislation.

The consent framework for use of data across the AI system should satisfy the following:

- clear framework
- kept up to date
- individuals are provided with informed consent for how their data will be used
- a dedicated team to own and maintain a register on how data is being used and to show compliance with the terms of the consent

The data should be thought of in groupings or packages, including:

- the data within the organisation
- the data surrounding the algorithm, APIs, and user interface
- the data used to train the AI system
- the data used for testing and integration
- data inputted at regular intervals in monitoring the data
- the data used at deployment, including input and output data from and to users.

Criterion 45: Plan for data archival and destruction.

Consider the following:

- will data be made available for future use, and what data
- restrictions and access controls in place
- will data be restricted until a specific date
- file formats to ensure data remains available during the archival period
- alignment with data sharing arrangements
- arrangements for data used to train and test AI models, and associated model management arrangements
- clear criteria for data archival and destruction for the data used at each stage of the AI lifecycle
- guidelines in the Information management for records created using AI technologies | naa.gov.au.

Agencies should:

Criterion 46: Analyse data for use by mapping the data supply chain and ensuring traceability.

Mapping the data supply chain to the AI system involves capturing how data will be stored, shared, and processed, particularly at the training and testing stages, which involve regular injections of data. When mapping the data account for:

- how data was sourced
- what data is required by the system, ensuring that excess data or data irrelevant to the functioning of system is not consumed by the system
- the amount and type of data the system will use
- what could affect the reliable accessibility of data
- how data will be fused and transformed
- how will the data be secured at rest and in transit
- how will the data be used by the system.

Ensuring traceability entails maintaining awareness of the flow of data across the AI system.

This includes:

- data sovereignty controls and considerations including legal implications for geographic locations for data (including its metadata and logs) when at rest, in transit, or in use. For classified data processing on cloud platforms, it is recommended to use cloud service providers and cloud services located in Australia, as per Cloud assessment and authorisation | cyber.gov.au
- providing the level of detail for debugging data errors and troubleshooting
- enforcing organisational policies on information management
- enhancing visibility over changes to the data occurring during migrations, system updates, or other errors
- supporting users to identify and fix data issues with a clear information audit trail
- supporting diagnosis for bias
- managing the quality of data to maintain availability and consistency.

Criterion 47: Implement practices to maintain and reuse data.

This involves determining ongoing mechanisms for ensuring data is protected, accessible, and available for use in line with the original consent parameters.

Any changes in data scope, including expansion in scope and usage patterns, would need to be monitored and addressed.

Statement 14: Implement data orchestration processes

Agencies must:

Criterion 48: Implement processes to enable data access and retrieval, encompassing the sharing, archiving, and deletion of data.

Considerations include:

- security classifications and permissions of the data
- speed or mode of the data, such as streaming or batch data
- alignment to Guidelines for data transfers | cyber.gov.au.

Agencies should:

Criterion 49: Establish standard operating procedures for data orchestration.

This includes:

- defining responsibilities between business areas and identifying mutual outcomes to be managed across teams. This is particularly important for business areas that are owners of datasets
- considering inclusion of infrastructure arrangements and use of cloud arrangements for data storage or processing.

Practices to be defined include:

- data governance
- data testing
- security and access controls.

Criterion 50: Configure integration processes to integrate data in increments.

This includes:

- enabling agencies to better manage incident identification and intervention during data integration

- ensuring risks of creating personal identifiable information from data integration are managed appropriately.

Criterion 51: Implement automation processes to orchestrate the reliable flow of data between systems and platforms.

Criterion 52: Perform oversight and regular testing of task dependencies.

This should involve having comprehensive backup plans in place to handle potential outages or incidents.

The following should be considered:

- regular backups of critical data
- failover mechanisms
- detailed recovery procedures to minimise downtime and data loss.

Criterion 53: Establish and maintain data exchange processes.

This includes:

- how often will data need to be accessed by the system
- at what points will the frequency, magnitude, or speed of access change
- how will security processes adapt when data is exposed to new risks across the AI system
- how will data be monitored for changes to accessibility or completeness
- will the sensitivity of the data change once processed or analysed
- how to validate data trust and authenticity.

Statement 15: Implement data transformation and feature engineering practices

Agencies should:

Criterion 54: Establish data cleaning procedures to manage any data issues.

Data cleaning involves appropriately treating data errors, inconsistencies, or missing values to improve performance of the AI system. Data cleaning should be documented,

and possibly included in the metadata, each time it is conducted to manage issues such as:

- blanks, nulls, or trailing spaces
- structural errors or unwanted formatting
- missing data
- spelling mistakes
- repetition of words
- irrelevant characters
- content or observations irrelevant to the purpose of the AI system.
- For open-source data, or data that has not yet been validated or can be trusted, consider using a sandbox environment.

Criterion 55: Define data transformation processes to convert and optimise data for the AI system.

This could leverage existing Extract, Transform and Load (ETL) or Extract, Load and Transform (ELT) processes.

Consider the following data transformation techniques:

- data standardisation – convert data from various sources into a consistent format
- data reorganisation – organise data to make it easier to query and analyse
- data integration – combine data from different sources for a single unified view
- discretisation – convert continuous data into discrete intervals
- missing value imputation – analyse what values need to be imputed and the method
- convert data from one source to another, such as log transformation
- smoothing – to even out fluctuations
- convert unstructured data to structured data
- Optical Character Recognition (OCR) – convert images of text into machine readable format
- object labelling and tracking – in images, audio, and video
- signal processing and transformation
- point in time of data – a snapshot of data at a specific point in time.

Criterion 56: Map the points where transformation occurs between datasets and across the AI system.

Consider:

- security checks.

Criterion 57: Identify fit-for-purpose feature engineering techniques.

Feature engineering techniques include:

- feature creation and extraction – deriving features from existing data to help the AI system produce better quality outputs
- feature selection – selecting attributes or fields that provide relevant context to the AI model
- encoding – converting data into a format that can be better used in AI algorithms
- binning – grouping data into categories
- specific conversion – changing data from one format to another for AI compatibility
- scaling – mapping all data to a specific range to help improve AI outputs.

Criterion 58: Apply consistent data transformation and feature engineering methods to support data reuse and extensibility.

Consider:

- metadata and tagging of the data
- data transformation not limited to AI models and processes.

Statement 16: Ensure data quality is acceptable

Agencies must:

Criterion 59: Define quality assessment criteria for the data used in the AI system.

Data quality can be measured across a variety of dimensions (in line with the ABS Data Quality Framework | abs.gov.au) by identifying institutional environment, relevance, timeliness, accuracy, coherence, interpretability, and accessibility.

A report on data quality can include:

- data quality statement (see ABS Data Quality Statement Checklist | abs.gov.au)

- metrics for measuring data quality, including its correctness and credibility
- frequency of reporting on data quality
- delegating ownership to a business area to be responsible for managing data quality
- monitoring for changes in quality across the supply chain
- intervening and addressing data quality issues as they arise.
- Consider:
- any existing data standard frameworks that are used by the agency.

Agencies should:

Criterion 60: Implement data profiling activities and remediate any data quality issues.

This involves analysing the structure, content, and quality of the data to determine its fitness for purpose for an AI system.

Data profiling can investigate the following characteristics:

- frequency
- volume, range, and distribution
- invalid entry identification
- error detection
- duplicates identification
- noise identification
- specific pattern identification.

Methods that can be used to explore and analyse the data include:

- descriptive statistics, such as mean, median, mode, or frequencies
- business rules – apply business knowledge
- clustering or dendrogram – group similar observations together
- visualisation – to get a visual representation of the data from various types of graphs and charts, such as histograms, bar-plot, boxplots, density plots, or heatmaps
- correlation analysis – measure relationships between variables, usually between numerical variables
- scatter plots – visualise relationships between two numerical variables
- cross-tabulations – analyse relationships between multiple categorical variables

- principal component analysis – analyse variables with the most variance
- factor analysis – helps reveal hidden patterns.

Criterion 61: Define processes for labelling data and managing the quality of data labels.

Data labelling can be done for the purposes of managing and storing data, audit purposes and AI model training purposes. Humans with appropriate skills and knowledge can perform the data labelling or it could be supported by automated labelling tools.

Setting data labelling practices can help optimise performance across the AI system by describing the context, categories, and relationships between data types, creating lineage of data through the AI system via versioning, distinguishing between pre-deployment and live data, and identifying what data will be reused, archived, or destroyed.

These include:

- establishing naming schemes, taxonomy, tagging, and data labelling practices
- considering different techniques such as manual or automated labelling, crowdsourcing, and quality checks
- defining quality control methods to improve consistency of labelling and assist in reducing bias
- considering changes to the raw data and data imputations, and associated impact
- providing data labels for AI training approaches or testing the AI models. Labels can provide the ground truth data for AI models and can influence AI validation. Different types of data labelling include:
  - classification
  - regression
  - visual object labels
  - audio labels
  - entity tagging.
- applying quality assurance measures to data labels, labelling personnel, and automated data-labelling support tools
- implementing bias mitigation practices in labelling:

- establishing a review process. Diverse people could independently label the same data so correlation could be analysed. Final labels could go through spot check review by subject matter experts
- establishing feedback loops. Labellers should be able to report issues and suggest improvements, and automated systems should be updated to be consistent with corrections made by human labellers
- establishing performance management for staff. Data labellers should undergo periodic training, performance reviews, and random audits for quality control
- implementing metadata labelling techniques that capture the type of data categories within the system and the relationship between these categories. Metadata labels can be prepared for model bias evaluation by annotating metadata with suitable dimensions. Ensure the metadata labelling aligns to the Guide on Metadata Attributes | datacommissioner.gov.au and Australian Government Recordkeeping Metadata Standard | naa.gov.au.

- assessing and monitoring quality of all automated data labelling support tools. Determine the regularity and criteria for these quality checks and report on findings
- updating and maintaining the labelling tools and processes to adapt to new data types, and labelling requirements
- considering potential harm to data labellers who may need to access sensitive or distressing content. This can occur when training an AI model to prevent responses including violence, hate speech, or sexual abuse.

Statement 17: Validate and select data

Agencies must:

Criterion 62: Perform data validation activities to ensure data meets the requirements for the AI system's purpose.

This involves including AI-specific validations in schema migrations to ensure data pipelines and feature stores remain functional. Suitable data validation techniques include:

- type validation – ensuring data is in the correct data type
- format validation – ensuring data aligns to a predefined pattern
- range validation – checking whether data falls within a specific range

- outlier detection – checking for data points that significantly deviate from the general data pattern
- completeness – verifying that all required fields are filled
- diversity – ensuring the data represents a variety of data points

Considerations include:

- a quality framework
- online near real-time and offline batch data validation mechanisms to support the purpose and operations of the AI system.

Criterion 63: Select data for use that is aligned with the purpose of the AI system.

This includes:

- alignment with the agency's business intent and the goals of the AI system, as well as ensuring data meets the data quality criteria previously established
- maintaining a live test dataset to test the AI system in production, to help monitor and maintain the operational integrity of the AI system.

Statement 18: Enable data fusion, integration and sharing

Agencies should:

Criterion 64: Analyse data fusion and integration requirements.

This includes:

- datasets, their sources and their owners
- purpose of the datasets for the AI system and intended outcomes
- data interdependencies
- risks associated with the datasets and mitigation plans
- data fusion and integration methodology for the AI system
- metrics to assess the quality of the fusion and data integration process and its outputs
- security, storage, and access requirements
- scalability intentions
- documentation and traceability

- regular audits and reviews
- data sharing principles and the risk management framework data as per the Data Availability and Transparency Act 2022 | datacommissioner.gov.au (DATA Scheme)
- compliance with the Guidelines on data matching in Australian Government administration guidelines | oaic.gov.au
- ethical considerations and guidance on data use as per the Data Ethics Framework | finance.gov.au.

Data fusion is a method to integrate or combine data from multiple sources and this can help an AI system create a more comprehensive, reliable, and accurate output. Meaningful data sharing practices across the agency can build interoperability between systems and datasets. Data sharing also promotes reuse, reducing resources for collection and analysis.

Criterion 65: Establish an approach to data fusion and integration.

This approach should involve one or more of the following processes:

- ETL (Extract, Transform and Load) – batch movements of data
- ELT (Extract, Load and Transform) – batch movements of data
- Application programming interface (API) – allowing the movement and syncing of data across multiple applications
- data streaming – moving data in or near real-time from source to target
- data virtualisation – combining streaming data virtually from different sources on demand
- chaining of AI models – linking multiple AI models in a sequence where the output from one model becomes the input for another.

Consider:

- data migration guidelines and any agency data management agreements, if relevant.

Agencies can optimise data fusion and integration processes by automating scheduling and data integration tasks and by deploying intuitive interfaces to diagnose and resolve errors.

Criterion 66: Identify data sharing arrangements and processes to maintain consistency.

Data sharing considerations include:

- whether other systems could leverage the data analysed by the AI system
- which areas within the agency would benefit from analysed data being shared with them
- what data containers could improve with access to the system's data sources
- whether data on how the system was trained could be used to train other systems
- documentation such as a memorandum of understanding, or similar, for data sharing arrangements intra-agency, inter-agency, or with external parties
- addressing risks of creating personal identifiable information
- what can be published for public, government, or internal benefit
- any legislative implications.

Statement 19: Establish the model and context dataset

Agencies must:

Criterion 67: Measure how representative the model dataset is.

Key considerations for measuring and selecting a model dataset include:

- whether it is representative of the true population relevant to the purpose of the AI system – this will improve model generalisation and minimise overfitting
- ensuring the dataset has the required features, volumes, distribution, representation and demographics, including people with lived experience and intersectional dimensions. For example, someone with cultural or linguistic diversity, may also be a person with disability, the dataset must consider how multiple dimensions of a person intersect and create unique experiences or challenges
- for GenAI, assess data quality thresholds and mechanisms in the data setup for modelling to help avoid unwanted bias and hallucinations.

Criterion 68: Separate the model training dataset from the validation and testing datasets.

Agencies must maintain the separation between these datasets to avoid any misleading evaluation for trained models.

Agencies can refresh these datasets to account for timeframes, degradation in AI performance during operation, and compute resource constraints.

Criterion 69: Manage bias in the data.

Techniques for agencies to manage and mitigate problematic bias in their model dataset includes:

- data collection analysis – examining how data was generated and verified, and checking the methodologies used to ensure the data is diverse and represents the real population
- data source analysis – investigating limitations and assumptions around the origin of the data
- data diversity – determining various demographics, sources and types of data, inclusion and exclusion considerations
- statistical testing – determining the likelihood of the population being accurately represented in the data
- class imbalance – analysing data for class imbalance before using it to train classification models, and applying relevant data and algorithm techniques and metrics, such as precision or F1-score, to address this
- outlier detection – identifying outliers or unusual data points in the data and ensuring they are handled appropriately
- exploratory data analysis – using descriptive statistics and data visualisation tools to identify patterns and discrepancies
- removing any irrelevant data from the training data that does not improve the performance of the model
- ensuring that any sensitive and protected data are retained in the test datasets for the purpose of evaluating for bias
- data augmentation – deploying measures to address the completeness of the model dataset, through supplementary data collection or synthetic data generation
- transparency – identifying bias and where it originated from through transparency on data sourcing and processing
- domain knowledge – ensuring practitioners have relevant domain knowledge on the datasets the AI system uses to serve the scope of the AI, including an understanding of the data characteristics and what it represents for the organisation
- documentation of data use – documenting the use of data by the AI system and any potential change of use, providing an audit trail of any incidence and causation of bias.

Agencies should:

Criterion 70: For generative AI, build reference or contextual datasets to improve the quality of AI outputs.

A reference or a contextual dataset for GenAI, can be in the form of (and not limited to) a retrieval-augmented generation (RAG) dataset or a prompt dataset.

Key considerations include:

- building high-quality reference or contextual datasets to support more accurate and context aware AI outputs, and reduce hallucinations
- implementing pre-defined prompts tailored to ensure consistent and reliable responses from GenAI models
- establish workflows for prompt engineering and data preparation to streamline development and deployment of GenAI systems.

Train

The train stage covers the creation and selection of models and algorithms. The key activities in this stage include modelling, pre- and post-processing, model refinements, and fine-tuning. It also considers the use of pre-trained models and associated fine-tuning for the operational context.

AI training involves processing large amounts of data to enable AI models to recognise patterns, make predictions, draw inferences, and generate content. This process creates a mathematical model with parameters that can range from a few to trillions. Training an AI model might require adjustment of these parameters, entailing increased processing power and storage.

Training a model can be compute-heavy, relying on infrastructure that may be significantly expensive. The model architecture, including choice of the AI algorithm and learning strategy, together with the size of the model dataset, will influence the infrastructure requirements for the training environment.

The AI Model encapsulates a complex mathematical relationship between input and output data that it derives from patterns in a modelling dataset. AI models can be chained together to provide more complex capabilities.

Pre-processing and post-processing augment the capabilities of the AI model. Application, platform, and infrastructure components are shown here as well as they all contribute to the overall behaviour and performance of the whole AI system.

Due to the number of mathematical computations involved and time taken to execute them, training can be a highly intensive stage of the AI lifecycle. This will depend on the infrastructure resources available, the algorithms used to train the AI model and the size of the training datasets.

Key considerations during this stage include:

- the model architecture, including the AI model and how components within the model interact, as well as the use of off-the-shelf or pre-trained models
- selection and development of the algorithms and learning strategies used to train the AI model
- an iterative process of implementing model architecture, setting hyperparameters, and training on model datasets

- model validation tests, supplemented by human evaluation, which evaluate whether the model is fit-for-purpose and reliable
- trained model selection assessments, which streamline development and enhance capabilities by comparing various models for the AI system
- continuous improvement frameworks which set processes for measuring model outputs, business, and user feedback to manage model performance.

If after multiple attempts of refinement, the model does not meet requirements or success criteria, a new model may need to be created, business requirements updated, or the model is retired.

See the Design lifecycle stage for details on measuring model outputs, as well as business and user feedback, to manage AI model performance.

See the Apply version control practices statement in the Whole of AI lifecycle section for detail on tracking changes to training models, trained models, algorithms, learning types, and hyperparameters.

Statement 20: Plan the model architecture

Agencies must:

Criterion 71: Establish success criteria that covers any AI training and operational limitations for infrastructure and costs.

Ensure alignment with AI system metrics selected at the design stage.

Consider:

- AI system purpose and requirements including explainability
- pre-defined AI system metrics including AI performance metrics
- impact and treatment for false positives and false negatives
- AI operational environment including scalability intentions
- frequency of change in context
- limitations on compute infrastructure
- cost constraints
- operational models such as ModelOps, MLOps, LLMOps, DataOps, and DevOps (see Statement 1).

AI training can occur in offline mode, or in online or real-time mode. This is dependent on the business case and the maturity of the data and infrastructure architecture. The risk of the model becoming stale is higher in offline mode, while the risk of the model exhibiting unverified behaviour is higher in online mode.

The training process is interdependent to the infrastructure in the training environment. Complex model architectures with highly specialised learning strategies and large model datasets generally require tailored infrastructure to manage costs.

Criterion 72: Define a model architecture for the use case suitable to the data and AI system operation.

The following will influence the choice of the model architecture and algorithms:

- business requirements – risk thresholds or performance criteria
- purpose of the system – identified stakeholders and the intended outcomes, safety, reproducibility level of AI model outputs, or explainability level for AI outputs
- data – bias, quality, and managing the supply of data to the system
- supporting infrastructure – computational demands, costs, and speed with respect to business needs
- resourcing – the capabilities involved with documentation, oversight, and intervention in training the AI model or reusable assets
- design – the training process will include necessary human oversight and intervention, to ensure responsible AI practices are in place. Consider embedding flexible architecture practices to avoid vendor lock-in.

The model architecture will highlight the variables that will impact the intended outcomes for the system. These variables will include the model dataset, use case application and scalability intentions. These variables will influence which algorithms and learning strategies are chosen to train the AI model.

An AI scientist can test and analyse the model architecture and dataset to identify what is needed to effectively train the system. Additionally, they can outline requirements for the model architecture to comply with data, privacy, and ethical expectations.

Consider starting off with simple and small architectures, and add complexity progressively depending on the purpose of the system to simplify debugging and reduce errors. Note that an AI system can contain a combination of multiple models which can add to the complexity.

Generally, a single type of algorithm and training process may not be sufficient to determine optimal models for the AI system. It is usually good practice to train multiple models with various algorithms and training methodologies.

There are options to develop a chain of AI models, or add more complexity, if that better meets the intent of the AI system. Each model could use a different type of algorithm and training process.

Analysis of support and maintenance of the AI system in operation can influence the model architecture. For some use cases, a complete model refresh may be required, noting cost considerations. Alternatives such as updates to pre or post processing could be considered, including updates to the configuration or knowledge repository for RAG for GenAI.

It may not be necessary to retrain models every time new information becomes available, and this should be considered when defining the model architecture. For example, for GenAI, adding new information in RAG can help the AI system remain up to date without the need to retrain the AI model, saving on costs without impacting AI accuracy.

Criterion 73: Select algorithms aligned with the purpose of the AI system and the available data.

There are various forms of algorithms to train an AI model, and it is important to select them based on the AI system requirements, model success criteria, and the available model dataset. A learning strategy is a method to train an AI model and dictates the mathematical computations that will be required during the training process.

Depending on use case, some examples of the types of training processes may include:

- supervised learning – training an AI model with a dataset, made up of observations, that has desired outputs or labels, such as support vector machines or tree-based models
- unsupervised learning – training a model to learn patterns in the dataset itself, where the training dataset does not have desired outputs or labels, such as anomaly detection or transformer LLMs
- reinforcement learning – training a model to maximise pre-defined goals, such as Monte Carlo tree search or fine-tuning models

- transfer learning – a model trained on one task, such as a pre-trained model, is reused as a starting point to enhance model performance on a related, yet different, task
- parameter tuning – optimising a model's performance by adjusting parameters or hyperparameters of a model, usually adjusted automatically
- model retraining – updating a model with new data
- online or real-time mode – continuously train the model using live data (note that this can significantly increase vulnerability of the AI system, such as data poisoning attacks).

Like traditional software, there are options to reuse, reconfigure, buy, or build models. An agency could reuse off-the-shelf models as-is, fine-tune pre-trained models, use pre-built algorithms, or create new models. The approach taken to training will vary across model types.

Criterion 74: Set training boundaries in relation to any infrastructure, performance, and cost limitations.

Agencies should:

Criterion 75: Start small, scale gradually.

Consider:

- starting off with simple and small architectures and add complexity progressively, depending on the purpose of the system to simplify debugging and reduce errors
- that an AI system can contain a combination of multiple models which can add to the complexity.

Statement 21: Establish the training environment

Agencies must:

Criterion 76: Establish compute resources and infrastructure for the training environment.

This allows for infrastructure and computational constraints to be considered in relation to business needs and supports configuration of learning strategies best optimised for the infrastructure environment.

Criterion 77: Secure the infrastructure.

Implement required security and access controls for infrastructure used for training, validating, and testing the AI model which are dependent on the security classification of the data. For details, see the Information security manual (ISM) | cyber.gov.au, Essential Eight maturity model | cyber.gov.au, Protective Security Policy Framework | protectivesecurity.gov.au and Strategies to mitigate cyber security incidents | cyber.gov.au.

Agencies should:

Criterion 78: Reuse approved AI modelling frameworks, libraries, and tools.

Statement 22: Implement model creation, tuning, and grounding

Agencies must:

Criterion 79: Set assessment criteria for the AI model, with respect to pre-defined metrics for the AI system.

These criteria should address:

- success factors specific to user stories
- model quality thresholds and performance of the AI system
- explainability and interpretability requirements
- security and privacy requirements
- ethics requirements
- tolerance for error for model outputs
- tolerance for negative impacts
- error rates by scale and similar processing by humans.

Considerations for modelling include:

- model training, maintenance, and support costs
- data and compute infrastructure constraints
- likelihood of the AI models becoming outdated
- whether the model can be legally used for the intended use case

- whether methods can be implemented to mitigate risk of new harms being introduced into the AI system
- bias, security, and ethical concerns
- whether the model meets the explainability and interpretability requirements
- use of model interpretability tools to analyse important features and decision logic.

Criterion 80: Identify and address situations when AI outputs should not be provided.

These situations include:

- low confidence scores
- when user input and context are ambiguous or lack reliable sources
- complex questions as input
- limited knowledge base
- privacy concerns and potential breach of safety
- harmful content
- unlawful content
- misleading content.

For GenAI, implementing techniques such as threshold settings or content filtering could address these situations.

Criterion 81: Apply considerations for reusing existing agency models, off-the-shelf, and pre-trained models.

These include:

- whether the model can be adapted to meet the KPIs for the AI system
- suitability of pre-defined AI architecture
- availability of AI specialist skills or skills required for configuration and integration
- whether the model is relevant to the target operating domain or can be adapted to it, such as fine-tuning, retrieval-augmented generation (RAG), and pre-processing and post-processing techniques
- cybersecurity assessment in line with Australian Government policies and guidance (see Whole of AI Lifecycle for more details).

Criterion 82: Create or fine-tune models optimised for target domain environment.

This includes:

- model testing on target operating environment and infrastructure
- using pre-processing and post-processing techniques
- addressing input and output filtering requirements for safety and reliability
- grounding such as RAG, which can augment a large language model (LLM) with trusted data from a database or knowledge base internal to an agency
- for GenAI, prompt engineering or establishing a prompt library, which can streamline and improve interactions with an AI model
- consider cost and performance implications associated with the adaptation techniques
- perform unit testing for the training algorithm, pre-processing, and post-processing algorithms
- track model training implementations systematically to speed up the discovery and development of models.

Agencies should:

Criterion 83: Create and train using multiple model architectures and learning strategies.

Systematically track model training implementations to speed up the discovery and development of models. This will help select a more optimal trained model.

Statement 23: Validate, assess, and update model

Agencies must:

Criterion 84: Set techniques to validate AI trained models.

There are multiple qualitative and quantitative techniques and tools for model validation, informed by the AI system success criteria (see Design section), including:

- correct classifications, predictions or forecasts, and factual correctness and relevance
- identify between positive and negative instances, and distinguish between classes

- benchmarking
- consistency in responses, clarity and coherence
- source attribution
- data-centric validation approaches for GenAI models.

Criterion 85: Evaluate the model against training boundaries.

Evaluation considerations include:

- poor or degraded performance of the model
- change of AI context or operational setting
- data retention policies
- model retention policies.

Criterion 86: Evaluate the model for bias, implement and test bias mitigations.

This includes:

- using suitable tools that test and discover unwarranted associations between an algorithm's protected input features and its output
- evaluating performance across suitable and intersectional dimensions
- checking if bias could be managed through updating the training data (see Statement 18)
- implementing bias mitigation thresholds that can be configured post-deployment
- implementing pre-processing or post-processing techniques such as disparate impact remover, equalised odds post-processing, content filtering, and RAG.

Agencies should:

Criterion 87: Identify relevant model refinement methods.

These considerations may trigger model refinement or retirement and can include:

- model parameter or weight adjustments – further training or re-training the model on a new set of observations, or additional training data
- adjusting data pre-processing or post-processing components
- model pruning – to reduce redundant mathematical calculations and speed up operations.

Statement 24: Select trained models

Agencies should:

Criterion 88: Assess a pool of trained models against acceptance metrics to select a model for the AI system.

This involves:

- defining clear needs and expectations
- comparing multiple trained models, usually generated based on different configurations
- prioritising based on metrics such as ‘simplest’ or ‘most effective’
- documenting the rationale for selection based on results from training models with various model architectures, learning strategies, and configurations
- any risk and mitigation plans
- a model refresh and re-training plan and register
- implementing mechanisms for explainability of model outputs to system users
- feedback channels and mechanisms implemented for monitoring and managing model performance
- an audit plan
- documenting a method for retiring the model.

Statement 25: Implement continuous improvement frameworks

Agencies must:

Criterion 89: Establish interface tools and feedback channels for machines and humans.

This also involves providing appropriate human-machine interface tools for human interrogation and oversight.

Criterion 90: Perform model version control.

AI model versioning and tracking is key to comparing performance over time, identifying factors affecting performance and updating the model when needed.

AI model tracking and versioning can involve:

- dataset tracking and versioning
- model tracking and versioning – each trained model can have the following details:
- algorithm, learning type and hyperparameter settings
- compile-time parameters
- the tools or version of tools used to compile the model.

Set-up rollback options to historical models to help develop safety nets in the AI system, reduce risk of deploying new models, and provide deployment flexibility.

Evaluate

This stage includes testing, verification, and validation of the whole AI system. It is assumed that agencies have existing capability on test management and on testing traditional software and systems.

Testing should be done continuously at each component of the AI system lifecycle. The level of testing at each stage could be at unit, integration, system, or system-of-systems, depending on the scope of development at that stage and the test strategy.

Testing can be divided into formal and informal phases. Formal testing is the phase when the system under test (SUT) is formally versioned, and the outcomes are evaluated for deciding whether to deploy to production or not. Statements within the standard apply to the formal testing phase.

Where an AI system or components have been procured, deployment and integration into the deployer's environment may need to be done before starting formal testing.

Testing of AI systems differs from testing traditional software as they can be probabilistic and non-deterministic in nature. While probabilistic systems do not give you an exact value, non-deterministic systems may have different outputs using the same inputs.

Other key differences include your approach to regression testing. While making small changes to a non-AI system may have limited consequences, a step change in test data or in parameters can be significant for an AI system. This means you need to conduct more robust regression testing to mitigate the heightened risk of escaped defects.

Note that an AI system which learns dynamically will change its behaviour without being formally updated. This means changes may occur on the same deployment version and without formal testing. This will require a more rigorous continuous monitoring post deployment. The development team or supplier should confirm whether your AI system has been designed to learn dynamically or statically.

Statement 26: Adapt test strategies and practices for AI systems

Agencies must:

Criterion 91: Mitigate bias in the testing process.

This includes:

- differentiating test data for formal testing from the data used during model development
- ensuring test subjects are not involved in the development of the SUT
- providing testers and developers a degree of independence from each other
- using test-driven development, such as designing test cases based on the requirements, and prior to implementing data, models, and non-AI components
- conducting peer reviews, bias awareness training, and documenting test-related decisions and processes.

Criterion 92: Define test criteria approaches.

The test criteria determines whether a test case has passed or failed. It involves comparing an actual output with an expected output.

This includes:

- considering statistical techniques – in probabilistic and non-deterministic systems, a single test case run may not be sufficient to determine success or failure. It may involve repeating a test case multiple times, defining thresholds, minimum values, and average efficiency. An example of statistical approach in regulatory setting is EU's General Safety Regulation on driver drowsiness | eur-lex.europa.eu
- performing baseline testing against a reference system or operation – this is useful in cases where there are no comprehensive specifications to define expected output. This entails comparing the AI system's behaviour and performance against a reference system. The reference system may be non-AI, manual processing, robotic process automation, or earlier system versions
- considering metamorphic testing – for GenAI or systems where an exact output cannot be specified for a given input. Metamorphic testing involves verifying relations, which are rules about how the output should change relative to how the input is changed. Examples include semantic consistency, intended style and tone,

and preservation of facts. This can be done computationally through vectors or user acceptance testing.

Agencies should:

Criterion 93: Define how test coverage will be measured.

This includes:

- identifying the limitations of existing static and dynamic coverage measures
- ensuring there is a method for measuring code coverage. This may include when code has been added, such as pre-processing, post-processing, prompts, components being integrated, and model-specific coverage. Code coverage tools can surface untested code and reduce defects or harms
- tracing test cases against requirements, design, and risks to check for gaps and demonstrate test coverage.

Criterion 94: Define a strategy to ensure test adequacy.

Achieving full test coverage for AI systems may be challenging and not viable.
To maximise test coverage, consider:

- performing combinatorial testing where comprehensive testing of inputs, preconditions, and corresponding outputs is not feasible due to the complexity of real-world combinations. Combinatorial testing involves deriving test data by sampling from an extremely large set of possible inputs, preconditions, and outputs
- automating as much as practicable. With testing needs required to validate an AI system during development and maintenance phases, performing manual testing with sufficient coverage will not be feasible
- performing appropriate virtual and real-world testing. This includes:
  - using a mix of data sources such as real-world data and synthetic data
  - understanding limitations when using synthetic data to test AI systems
  - considering requirements for embedded systems, such as model-in-the-loop, software-in-the-loop, processor-in-the-loop, and hardware-in-the-loop
  - understanding limitations of the scope and data sampling size for real-world testing
  - documenting virtual environment assumptions and correlations with the real world
  - documenting the quantity of real-world testing and the rationale for selecting the real-world test cases
  - determine the statistical significance of the test results while considering limitations.

Statement 27: Test for specified behaviour

Agencies must:

Criterion 95: Undertake human verification of test design and implementation for correctness, consistency, and completeness.

Criterion 96: Conduct functional performance testing to verify the correctness of the SUT as per the pre-defined metrics.

This includes testing for fairness and bias to inform affirmative actions.

For off-the-shelf systems, consider benchmark testing using industry-standard benchmark suites and comparisons with competing AI systems.

Criterion 97: Perform controllability testing to verify human oversight and control, and system control requirements.

Criterion 98: Perform explainability and transparency testing as per the requirements.

This involves:

- testing that AI outputs are understandable for the target audience, ensuring diversity of test subjects and representativeness of the target population
- testing that the right information is available for the right user.

Criterion 99: Perform calibration testing as per the requirements.

This involves:

- measuring functional performance across various operating or installation conditions
- testing that changes in calibration parameters are detected
- testing that any out-of-range calibration parameters are rejected by the AI system in a transparent and explainable way.

Criterion 100: Perform logging tests as per the requirements.

This involves verifying that the system records:

- system warnings and errors
- relevant system changes with corresponding details of who made the change, timestamp, and system version.

Statement 28: Test for safety, robustness, and reliability

Agencies must:

Criterion 101: Test the computational performance of the system.

This includes:

- testing for response times, latency, and resource usage under various loads
- network and hardware load testing.

Criterion 102: Test safety measures through negative testing methods, failure testing, and fault injection.

- This includes:
- testing for incorrect or harmful inputs.

Criterion 103: Test reliability of the AI output, through stress testing over an extended period, simulating edge cases, and operating under extreme conditions.

Agencies should:

Criterion 104: Undertake adversarial testing (red team testing), attempting to break security and privacy measures to identify weaknesses.

AI-specific attacks can be executed before, during, and after training.

Examples of attacks that can be made before and during training includes:

- dataset poisoning
- algorithm poisoning
- model poisoning
- backdoor attacks.

Examples of attacks that can be made after training includes:

- input attack and evasion
- reverse engineering the model and data.

Statement 29: Test for conformance and compliance

Agencies must:

Criterion 105: Verify compliance with relevant policies, frameworks, and legislation.

Criterion 106: Verify conformance against organisation and industry-specific coding standards.

This includes static and dynamic source code analysis. While agencies may use traditional analysis tools for the whole system, it is important to note their limitations with respect to AI models and consider finding tools built specifically for AI models.

Criterion 107: Perform vulnerability testing to identify any well-known vulnerabilities.

This includes:

- testing for entire AI system.

Statement 30: Test for intended and unintended consequences

Agencies must:

Criterion 108: Perform user acceptance testing (UAT) and scenario testing, validating the system with a diversity of end users in their operating contexts and real-world scenarios.

Agencies should:

Criterion 109: Perform robust regression testing to mitigate the heightened risk of escaped defects resulting from changes, such as a step change in parameters.

Traditional software regression testing is insufficient.

This may include:

- back-to-back testing to compare two versions of system or software using historical data

- A/B software testing to simultaneously compare multiple versions in a real-world setting. This allows agencies to assess the impact of a specific model or software package on the overall system in its intended operating environment
- performance regression, checking for any degradation in model accuracy, fairness, or other key metrics.

Integrate

The integrate stage of the AI lifecycle focuses on implementing and testing an AI system within an agency's internal organisational environment, including with its systems and data.

Deploying a standalone or integrated AI system into existing IT infrastructure involves assessing compatibility and confirming data interoperability. Comprehensive integration may require the reconfiguration of current systems.

Agencies can achieve the best outcomes at this stage by adopting practices that closely align with those implemented at the test stage. This ensures the AI system has been thoroughly tested against its intended purpose prior to integration – a key measure before the AI system potentially contaminates the business environment. See Test section for a detailed discussion on testing methods.

Following recommended practices for managing code integration workflows for AI systems will help agencies to maintain quality, security, and consistency.

Statement 31: Undertake integration planning

Agencies should:

Criterion 110: Ensure the AI system meets architecture and operational requirements with the Australian Government Security Authority to Operate (SATO).

This aspect of integration planning includes:

- assessing the AI system and its third-party dependencies against the agency's requirements to identify risks
- assessing the AI system against the agency's architecture principles
- identifying any gaps between the agency's current and target infrastructure to support the AI system
- ensuring the AI system meets security and privacy requirements for handling classified data.

Criterion 111: Identify suitable tests for integration with the operational environment, systems, and data.

This includes:

- ensuring robust test methods are selected
- incorporating auto testing processes
- ensuring that environment controls satisfy security and privacy requirements for the data in the AI system.

Statement 32: Manage integration as a continuous practice

Agencies should:

Criterion 112: Apply secure and auditable continuous integration practices for AI systems.

Continuous integration (CI) pipelines enable agencies to build, test, and validate changes upon every commit or merge, while accounting for computational requirements resulting from re-testing expensive model training processes. The CI pipeline should include any automated tests defined in the test stage, automating model training, as well as static and dynamic source code analysis.

These pipelines typically involve:

- ensuring end-to-end integration to include data pipeline and data encryption practices
- verifying and managing dependency checks for outdated or vulnerable libraries
- validating infrastructure-as-code (IaC) scripts to ensure environments are deployed consistently
- steps to build and validate container images for AI applications
- continuous training and delivery of AI models and systems
- employing fail-fast mechanisms to halt builds upon detection of silent failures and critical errors, such as test failures or vulnerabilities
- avoiding the propagation of unverified changes from failed workflows to production environments
- establishing a centralised artifact and model registry, and include steps to package and store artifacts, such as models, APIs, and datasets.

Deploy

The deploy stage involves introducing all the AI technical components, datasets, and related code into a production environment where it can start processing live data.

Deployment involves rigorous testing, governance, and security practices to ensure systems perform as intended in live environments.

Defining structures for secure deployment of an AI system is particularly crucial in a government setting. Deployment strategies should be incremental, consistent, and non-disruptive.

NOTE: The Australian Signals Directorate's Australian Cyber Security Centre report on Deploying AI systems Securely | cyber.gov.au provides further advice on deploying AI systems securely.

Statement 33: Create business continuity plans

Agencies must:

Criterion 113: Develop plans to ensure critical systems remain operational during disruptions.

This includes:

- identifying and managing potential risks to AI operations
- defining disaster recovery, backup and restore, monitoring plans
- testing business continuity plans for relevance
- regularly reviewing and updating objectives, success criteria, failure indicators, plans, processes and procedures to ensure they remain appropriate to the use case and its operating environment.

Statement 34: Configure a staging environment

Agencies should:

Criterion 114: Ensure the staging environment mirrors the production environment in configurations, libraries, and dependencies for consistency and predictability suited to the use case.

Criterion 115: Measure the performance of the AI system in the staging environment against predefined metrics.

Criterion 116: Ensure deployment strategies include monitoring for AI-specific metrics, such as inference latency and output accuracy.

Statement 35: Deploy to a production environment

Agencies must:

Criterion 117: Apply strategies for phased roll-out.

Consider splitting traffic between the current and new version being rolled out, or rolling out to a subset of users to gradually introduce changes and detect issues before full deployment.

Criterion 118: Apply readiness verification, assurance checks, and change management practices for the AI system.

This typically involves:

- the readiness verification, which includes all tests and covers the entire system – code, model, data, and related components
- consent for data governance, data use, and auditing frameworks
- ensuring all production deployments follow change management protocols, including impact assessment, notifying stakeholders, updating training, assurance, approvals, testing, and documentation
- including the rationale for deploying or updating AI systems in the change records to ensure accountability and transparency

- understanding the implications of AI model auto-updates in production, including options to disable
- understanding the implications of AI system online and dynamic learning in production, including options to disable.

Agencies should:

Criterion 119: Apply strategies for limiting service interruptions.

This typically involves:

- implementing strategies to avoid service interruptions and reduce risk during updates where zero downtime is required
- configuring instance draining to ensure active requests are not interrupted while allowing completion of long-running AI inference tasks
- include cost tracking on deployment workflows for additional resources used during deployment
- include real-time monitoring and alerting to detect and respond to issues during deployment processes and transitions.

Statement 36: Implement rollout and safe rollback mechanisms

Agencies should:

Criterion 120: Define a comprehensive rollout and rollback strategy.

This should safeguard data and limit data corruption.

Criterion 121: Implement load balancing and traffic shifting methods for system rollout.

This includes:

- using load balancers to distribute traffic dynamically between old and new deployments during updates
- creating traffic shifting policies to safeguard against overwhelming newly deployed AI systems with high inference demands.

Criterion 122: Conduct regular testing, health checks, readiness, and startup probes to verify stability before routing traffic for all deployed AI services.

- Consider using probes to continuously monitor during deployment, to detect issues early and rollback upon failure.

Criterion 123: Implement rollback mechanisms to revert to the last stable version in case of failure.

This includes:

- implementing automated rollback mechanisms to revert to the last stable version in case of pre-defined critical failure for AI deployments
- failures that do not satisfy the trigger for automated rollback require human intervention to analyse and decide the next steps.

Monitor

The monitor stage of the AI lifecycle includes operating and maintaining the AI system. Monitoring is critical to ensuring the reliability, availability, performance, security, safety, and compliance of an AI system after it is deployed.

Monitoring AI systems is critical because changes in the operating environment and inputs could result to degradation and potential harms. Effective monitoring includes continuous performance evaluation, anomaly detection, intervention, and proactive incident response.

The measures implemented at this stage helps identify if a system is generating outputs that misalign with its intended purpose and promptly remedy issues.

Statement 37: Establish monitoring framework

Agencies should:

Criterion 124: Define reporting requirements.

This includes:

- establishing a plan for providing different stakeholders with reports
- for each group of stakeholders (persona), define what needs to be reported, why, when, and how.

Criterion 125: Define alerting requirements.

This includes:

- defining what information needs alerting
- defining what information is critical to be alerted in real-time
- defining severity levels, such as major, minor, warning
- defining thresholds, out-of-pattern behaviour, and other triggers for each alert level
- defining who needs to be alerted and the method of alert such as SMS or e-mail.

Criterion 126: Implement monitoring tools.

This includes:

- monitoring the information needed to satisfy alerting and reporting requirements

- automating monitoring, alerting, and reporting
- implementing management information and dashboards
- implementing role-based access to protect sensitive information and meet security requirements
- implementing real-time alerting requirements.

Criterion 127: Implement feedback loop to ensure that insights from monitoring are fed back into the development and improvement of the AI system.

This includes:

- a decision matrix outlining guidance on what components in the AI system would need an update or refresh, such as pre or post processing components, AI model, or a RAG knowledge base in a GenAI system
- a framework to provide and track recommended actions from the insights
- a guideline for identifying actions to address insights, with considerations to costs, delays, AI trust, and effectiveness.

Statement 38: Undertake ongoing testing and monitoring

Agencies must:

Criterion 128: Test periodically after deployment and have a clear framework to manage any issues.

This assures that the system still operates as intended. See Test section for applicable tests.

Criterion 129: Monitor the system as agreed and specified in its operating procedures.

Ensure the operators understand when, why, and how to intervene.

Criterion 130: Monitor performance and AI drift as per pre-defined metrics.

Criterion 131: Monitor health of the system and infrastructure.

This includes:

- monitoring logs for errors

- services or processes
- resources such as compute, memory, storage, and network.

Criterion 132: Monitor safety.

This includes:

- monitoring inputs and outputs for abuse, misuse, sensitive information disclosure, and other forms of harm.

Criterion 133: Monitor reliability metrics and mechanisms.

This includes:

- error rates
- fault detection
- recovery
- redundancy
- failover mechanisms.

Criterion 134: Monitor human-machine collaboration.

This includes:

- reviewing the human experience
- assessing the effectiveness of the human oversight and control measures
- analysing the usage metrics for friction points and finding opportunities for improving the overall outcome from human-machine collaboration
- considering different monitoring methods. While surveys could be cost effective, face-to-face interviews and observing users live while interacting with the AI system could provide better insights.

Criterion 135: Monitor for unintended consequences.

This typically includes:

- implementing various channels for people to provide feedback, issues, or contest outcomes
- consider if anonymous channels are needed
- tracing how the outputs of the AI system are used
- analysing quantitative and qualitative data for recurring harms

- look for missing data, such as checking if certain demographics are not using the system.

Criterion 136: Monitor transparency and explainability.

Periodically check that transparency and explainability requirements are met post deployment.

Criterion 137: Monitor costs.

The cost model for using AI systems may be different and much more costly than traditional software and systems.

Criterion 138: Monitor security.

This may include logging AI services in use to satisfy security requirements and ensuring appropriate data loss prevention (DLP).

Identify the scope of deployment data for the AI system.

These include:

- data submitted by the user (prompts)
- agency data augmented into the prompts
- content generated by the service via completions, images, and embedding operations
- training and validation data from the department that will be used for fine-tuning a model.

DLP includes:

- ensuring that the supplier does NOT use agency data for improving the supplier's AI systems or other products
- ensuring the system only accesses data that the end user is authorised to access
- ensuring that human review is performed by authorised users only
- monitoring for sensitive data disclosure
- monitoring data access and usage
- automating data classification
- monitoring for anomalies and suspicious activities
- ensuring data encryption is enabled for data-at-rest and data-in-transit

- ensuring that any data provided to the model, or generated by the model, can be deleted completely by the authorised user.

Criterion 139: Monitor compliance of the AI system.

Statement 39: Establish incident resolution processes

Agencies must:

Criterion 140: Define incident handling processes.

This involves establishing a structured process for incident management that ensures identified incidents are allocated a severity level and addressed promptly and effectively. This includes security incident, reporting, and monitoring.

This must comply with the Australian Government Protective Security Policy Framework (PSPF) | protectivesecurity.gov.au and the Information security manual (ISM) | cyber.gov.au.

Criterion 141: Implement corrective and preventive actions for incidents.

This includes:

- defining clear protocols for root cause analysis, implementing corrective actions, and preventive actions
- maintaining detailed logs and documentation to facilitate troubleshooting, provide input into longer term problem management, and assist continuous improvement of AI systems.

Decommission

The Decommissioning stage of the AI lifecycle focuses on the planning, delivery, and documentation of decommissioning activities.

Decommissioning an AI-enabled system encompasses the entire AI production system. It includes retiring, shutting down, or repurposing system components. It is part of the full AI lifecycle process and is distinct from other activities that may occur during the lifecycle like retiring data, AI models, infrastructure, or captured data.

Taking a structured approach will allow you to safely decommission an AI system – mitigating the risk of unauthorised access or a data breach while ensuring ongoing compliance.

NOTE: The Pilot AI assurance framework | digital.gov.au recommends understanding the implications of decommissioning an AI system to ensure agencies can address all potential consequences.

NOTE: The Voluntary AI Safety Standard | industry.gov.au promotes proactive stakeholder engagement during the retirement stage, as well as the importance of maintaining detailed records.

NOTE: The Policy for the responsible use of AI in government | digital.gov.au encourages business and technology process enhancement to assist APS AI capability uplift over time.

Statement 40: Create a decommissioning plan

Agencies must:

Criterion 142: Define the scope of decommissioning activities.

Decommissioning plans clearly identify the system components being shut down, disabled, reused, or repurposed and the reason for decommissioning.

Ensure compliance with Information management for records created using AI technologies | naa.gov.au.

Criterion 143: Conduct an impact analysis of decommissioning the target AI system.

Assessing the potential impacts on an agency's business operations, stakeholders and compliance obligations allows for the identification of dependencies, risks and any alternative solutions required to maintain service continuity.

Criterion 144: Proactively communicate system retirement.

This involves:

- informing all affected parties – employees, partners, and users about the decommissioning schedule, reasons for decommissioning, and any expected impacts
- addressing any issues or concerns
- providing support and information about alternative systems, and any forecast implementation schedules
- considering a retrospective review of the AI system's performance and lifecycle to provide valuable insights for future projects. The review should involve analysing the system's successes, challenges, and areas for improvement.

Statement 41: Shut down the AI system

Agencies must:

Criterion 145: Retain AI system compliance records.

Any records related to an AI system, including those generated during retirement, must be preserved for agencies to demonstrate compliance and effectively respond to future audits and inquiries.

Criterion 146: Disable computing resources or components specifically dedicated to the AI system.

Criterion 147: Securely decommission or repurpose all computing resources dedicated to the AI system, including individual and shared components.

This involves:

- systematically shutting down and wiping servers, storage devices, and network components

- identifying which system or network components will be repurposed, and disconnecting or reconfiguring accordingly
- terminating instances and services associated with cloud resources, ensuring no data remains.

Statement 42: Finalise documentation and reporting

Agencies must:

Criterion 148: Finalise decommissioning information and update organisational documentation.

This involves:

- recording all decommissioning information in the final system documentation, compiling records of decommissioning activities, decisions, and lessons learnt
- delivering a final report to relevant stakeholders
- providing all documentation related to the decommissioning of the AI system or components, including detailed accounts of:
- the decommissioning process
- compliance adherence
- implications for ongoing operations.