Australian Government
Digital Transformation Agency

dta

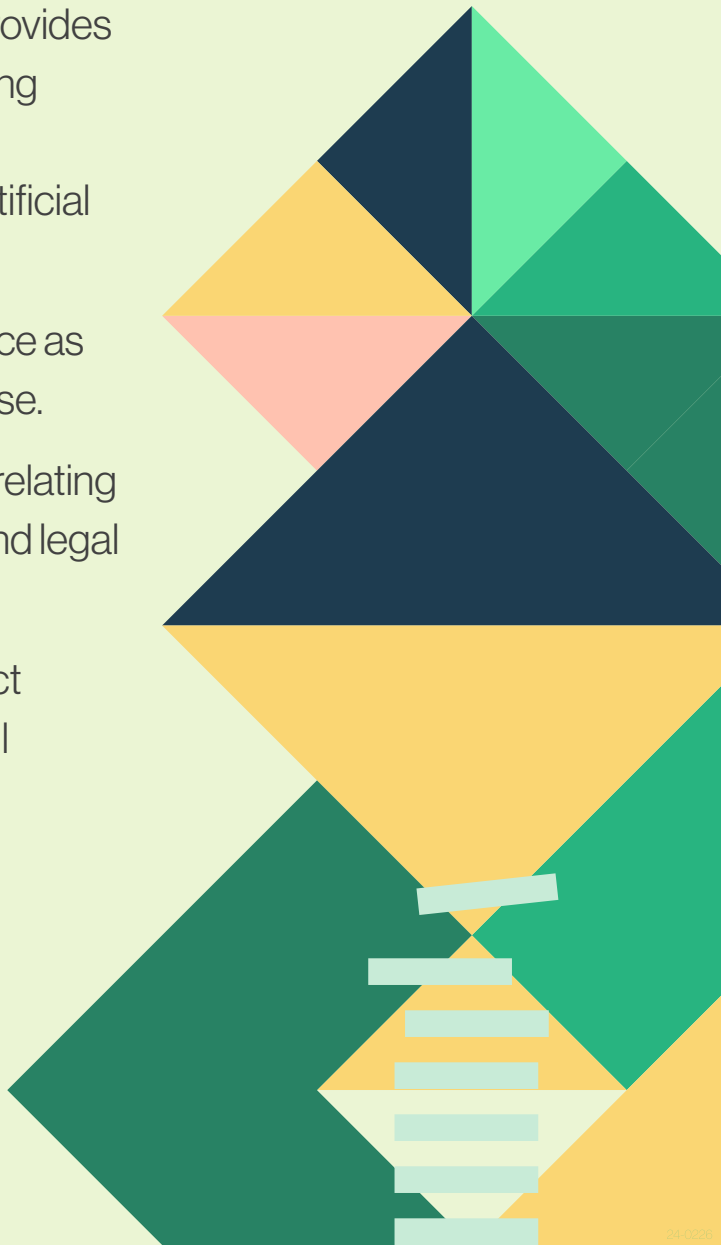
# Guidance for the artificial intelligence impact assessment tool

The Digital Transformation Agency (DTA) provides the AI impact assessment tool and supporting guidance to assist Australian Government agencies to assess their proposed use of artificial intelligence (AI).

Agencies should not treat the tool or guidance as legal advice or as authorising proposed AI use.

Agencies are responsible for any decisions relating to their use of AI and for seeking technical and legal advice as appropriate.

For questions and suggestions on the impact assessment tool and guidance, please email ai@dta.gov.au



dta.gov.au

**Digital Transformation Agency**



# Document control

This document is maintained by the DTA and provides practical advice for Australian Government staff using the AI impact assessment tool. The tool supports implementation of the Policy for the responsible use of AI in government (the AI policy).

It will be updated periodically as the policy and technology evolve. For enquiries, please email ai@dta.gov.au.

To view the AI policy, and to check you have the most up-to-date versions of the AI impact assessment tool and guidance, please visit https://digital.gov.au/ai/ai-in-government-policy.

| Version | Description | Date |
|---|---|---|
| v1.0 | Published | 01/12/2025 |

# Contents

# Introduction

This document provides guidance for completing the Australian Government artificial intelligence (AI) impact assessment tool (the tool). Use this supporting guidance to understand, interpret and complete the tool.

The impact assessment tool is for Australian Government teams working on an AI use case. It helps teams identify, assess and manage AI use case impacts and risks against Australia's AI Ethics Principles.

Agencies can use the tool to fulfill AI use case impact assessment requirements under the Policy for the responsible use of AI in government (the AI policy). Refer to the impact assessment tool document for assessment instructions and to the AI policy for key definitions and implementation requirements.

This guidance mirrors the AI impact assessment tool's 12-section structure. For advice on completing a section in the tool, find the corresponding section number in this guidance.

Assessing officers should familiarise themselves with the AI policy and the AI Ethics Principles. Also consider other Digital Transformation Agency (DTA) resources designed to support government AI adoption, including:

- the AI technical standard
- AI procurement resources, including the Guidance on AI procurement in government, AI contract template and AI model clauses
- guidance on the use of public generative AI tools for agencies and for staff.

The DTA welcomes user feedback on the tool and supporting guidance. Please send questions or suggestions to ai@dta.gov.au.

# 1. Basic information

## 1.1 AI use case profile

This section is intended to record basic information about the AI use case.

### Name of AI use case

Choose a clear, simple name that accurately conveys the nature of the use case.

### Internal reference number or identifier

Assign a unique reference number or other identifier for your assessment. This is intended to assist with internal record keeping and engagement with the DTA.

### Lead agency

The agency with primary responsibility for the AI use case. Where 2 or more agencies are jointly leading, nominate one as the contact point for assessment.

## 1.2 Establishing impact assessment responsibilities

### Assessing officer

An officer assigned to complete the assessment, coordinate the end-to-end process and serve as the contact point for any assessment queries. Depending on the use case and agency context, they may be a technical, data, governance or risk specialist, or a policy or project officer from the business area implementing the AI use case in its operations.

### Accountable use case owner(s)

This role is described in the AI policy and the Standard for accountability.

### Approving officer

This should be an officer with appropriate authority to approve the AI use case assessment, including the inherent risk ratings. Similar to the assessing officer role above, the approving officer's specific role in the AI use case will depend on the agency and use case context.

## 1.3 Additional roles and responsibilities

Clear roles and responsibilities are essential for ensuring accountability in the development and use of AI systems. In this section, you are asked to identify any additional individual officers that may have responsibilities related to your AI use case and the underlying AI system(s). Consider the roles already outlined – such as assessing officers or approving officers – and consider other positions that contribute to the AI system's lifecycle or oversight.

This is not intended to create new requirements for specific roles under the AI policy or this impact assessment. It is intended to help agencies record relevant roles and responsibilities, maintain transparency and facilitate accountability during AI use case implementation. For example, you could identify the person(s) responsible for:

- the decision to use the AI system and the scope of the AI system
- designing, developing and maintaining the system, such as key personnel of third-party suppliers
- applying interpreting the AI system's outputs, including decisions or actions based on those outputs
- controlling the AI system, with authority to start, stop, or deactivate the system under normal operating conditions
- monitoring and maintaining performance and safety, meeting quality standards and detecting errors, biases, and unintended consequences
- disengaging or stopping the system, if immediate intervention is required to prevent or stop harm
- the governance of the data used for operating, training or validating the AI system

You should consider distributing these roles among multiple officers where feasible, to avoid excessive concentration of responsibilities in a single individual, while ensuring responsible officers are appropriately skilled and senior.

## 1.4 AI use case description

Briefly explain how you are using or intending to use AI. This should an 'elevator pitch' that gives the reader a clear idea of the kind of AI use intended, without going into unnecessary technical detail, which is captured in your other project documentation. Use simple, clear language, avoiding technical jargon where possible. You may wish to include:

- a high level description of the problem the AI use case is trying to solve
- the way AI will be used
- the outcome it is intended to achieve.

## 1.5 In-scope use case

Record whether your AI use case is in scope of the Policy for the responsible use of AI in government (the AI policy). Appendix C of the AI policy specifies the criteria to determine if an AI use case is in scope.

At a minimum, an AI use case is in scope of the AI policy if any of the following apply:

- The use, misuse or failure of AI could lead to more than insignificant harm to individuals, communities, organisations, the environment or the collective rights of cultural groups including First Nations peoples.
- The use of AI will materially influence administrative decisions that affect individuals, communities, organisations, the environment or the collective rights of cultural groups including First Nations peoples.
- It is possible the public will directly interact with, or be significantly impacted by, the AI or its outputs without human review.
- The AI is designed to use personal or sensitive data[1] or security classified information[2].
- It is deemed an elevated risk AI use case as directed by the DTA.

Agencies may wish to apply the AI policy to AI use cases that do not meet the above criteria. This includes use cases with specific characteristics or factors unique to an agency's operating environment that may benefit from applying an impact assessment and governance actions.

The AI policy has been designed to exclude incidental and lower risk uses of AI that do not meet the criteria. Incidental uses of AI may include off-the-shelf software with AI features such as grammar checks and internet searches with AI functionality. The AI policy recognises that incidental usage of AI will grow over time and focuses on uses that require additional oversight and governance.

In assessing whether a use case is in scope, agencies should also carefully consider AI use in the following areas:

- recruitment and other employment-related decision making
- automated decision making of discretionary decisions
- administration of justice and democratic processes
- law enforcement, profiling individuals, and border control
- health
- education
- critical infrastructure.

While use cases in these areas are not automatically high-risk, they are more likely to involve risks that require careful attention through an impact assessment.

For information on how the policy applies when doing early-stage experimentation, refer to Appendix C of the AI policy.

1 As defined by the Privacy Act 1988 (Cth).
2 As defined by the Australian Government Protective Security Policy Framework

If your use case is within scope, record the AI policy criteria that apply using the checklist. If your use case meets multiple criteria, tick each one. If you are unsure, it is best practice to select the criteria that most closely reflect your use case.

The criteria are designed to help identify uses of AI that require additional oversight and governance. This provides a clearer picture of the types of uses across government.

Refer to the AI policy for further detail on mandatory use case governance actions. Consult your accountable use case owner and your agency's AI accountable official for agency-specific guidance on fulfilling the mandatory AI policy actions and any internal agency requirements in addition to the mandatory actions.

Note you can also apply the AI policy and impact assessment tool to use cases that do not meet the criteria. In this case, you can select 'not applicable' for this question.

## 1.6 Type of AI technology

Briefly explain what type of AI technology you are using or intend to use. For example, supervised or unsupervised learning, computer vision, natural language processing, generative AI.

This may require a more technical answer than the use case description. Aim to be clear and concise with your answer and use terms that a reasonably informed person with experience in the AI field would understand.

## 1.7 Usage pattern

Select the AI system usage pattern or patterns that apply to your use case. For usage pattern definitions, refer to the Classification systems for AI use.

## 1.8 Administrative decisions

Only complete this section if you selected 'Decision-making and administrative action' in assessment section 1.7 and if AI automated decision-making is used for an administrative decision under an Act.

Express legislative authority is generally required to automate decision-making of an administrative decision under an Act. Legal advice should be obtained for any proposed use of AI in this context.

Agencies using automated decision-making should review the Commonwealth Ombudsman's Better Practice Guide on Automated Decision Making.

Agencies should generally consider:

- any legislation or framework which requires a particular decision-maker such as a minister to make a decision – for example, the minister may not have discharged their duty if they rely on the AI outputs without proper validation

- an official's duty of care and diligence under the *Public Governance, Performance and Accountability Act 2013* (Cth) – for example, an official fails to validate AI outputs
- administrative law requirements of legality and procedural fairness – for example, how an automated decision can be challenged.

## 1.9 Domain

Select the AI system domain or domains that apply to your use case. For domain definitions, refer to the Classification systems for AI use.

## 1.10 Expert contributions

List any expert consultation undertaken during the assessment process, including the nature of their expertise, the specific contributions they made, and how their input informed the assessment process. While such consultation is not mandatory, agencies should consider engaging relevant internal or external expertise based on the complexity, novelty or potential impacts of the AI system.

Agencies should:

- Identify the areas where expert input is most needed, such as ethical considerations, legal risk and compliance, and technological challenges.
- Engage with a diverse range of experts to ensure a comprehensive assessment.
- Document the consultation process thoroughly, including the date of consultation, the experts' names and affiliations, and their key recommendations.
- Summarise how the expert input was integrated into the assessment and any resulting changes to the AI system or its deployment.
- Review and update the record of expert consultations periodically to ensure it remains relevant and accurate.

## 1.11 Impact assessment review log

As new information becomes available or design choices are refined, you should reassess all identified risks and consider whether previous responses still reflect the current state of the project. When data sources, functionality, user groups or other project elements change, revise previous answers to maintain clear and accurate records of the risk profile.

The AI policy specifies requirements for monitoring, evaluating and re-validating use cases following deployment. For details, refer to the AI policy.

# 2. Purpose and expected benefits

## 2.1 Problem definition

Describe the problem that you are trying to solve.

For example, the problem might be that your agency receives a high volume of public submissions, and that this volume makes it difficult to engage with the detail of issues raised in submissions in a timely manner.

Do not describe how you plan to fix the problem or how AI will be used.

Though 'problem' implies a negative framing, the problem may be that your agency is not able to take full advantage of an opportunity to do things in a better or more efficient way.

## 2.2 AI use case purpose

Clearly and concisely describe the purpose of your use of AI, focusing on how it will address the problem you described at section 2.1.

Your answer may read as a positive restatement of the problem and how it will be addressed.

For example, the purpose may be to enable you to process public submissions more efficiently and effectively and engage with the issues that they raise in more depth.

## 2.3 Non-AI alternatives

Briefly outline non-AI alternatives that could address the problem you described at section 2.1.

Non-AI alternatives may have advantages over solutions involving AI. For example, they may be cheaper, safer or more reliable.

Considering these alternatives will help clarify the benefits and drawbacks of using AI and help your agency make a more informed decision about whether to proceed with an AI based solution.

## 2.4 Identifying stakeholders

Conduct a mapping exercise to identify the individuals or groups who may be affected by the AI use case. Consider holding a workshop or brainstorm with a diverse team to identify the different direct and indirect stakeholders of your AI use case.

The stakeholder mapping aid attached to the impact assessment tool may help generate discussion on the types of stakeholder groups to consider. Please note the table has been provided as a prompt to aid discussion and is not intended as a prescriptive or comprehensive list.

## 2.5 Expected benefits

This section requires you to explain the expected benefits of the AI use case, considering the stakeholders identified in the previous question. The AI Ethics Principles specify that throughout their lifecycle, AI systems should benefit individuals, society and the environment.

This analysis should be supported by specific metrics or qualitative analysis. Metrics should be quantifiable measures of positive outcomes that can be measured after the AI is deployed to assess the value of using AI. Any qualitative analysis should consider whether there is an expected positive outcome and whether AI is a good fit to accomplish the relevant task, particularly when compared to the non-AI alternatives you identified previously. Benefits may include gaining new insights or data.

Consider consulting the following resources for further advice:

- DTA Benefits management policy
- Australian Government Guide to Policy Impact Analysis.

# 3. Inherent risk assessment

To complete the inherent risk assessment, follow these steps.

## Sections 3.1 to 3.8

### Definitions

Inherent risk: reflects the level of risk that exists before any additional or new controls are applied. This is the risk level under standard operating conditions, assuming only existing baseline or standard controls are in place.

Residual risk: reflects the level of risk that remains after new or additional treatments, controls or safeguards have been implemented.

### Determine risk likelihood and consequence

For each risk category listed in the assessment table, determine the likelihood and consequence of the risk occurring for your AI use case. The likelihood descriptors are provided in **Table 1** of the impact assessment tool, and consequence descriptors are in the **appendix** to this guidance.

The inherent risk assessment should reflect the intended scope and function of the AI use case. In conducting your assessment, you should be clear on:

- key factors contributing to the likelihood and consequence of the risk
- any assumptions or uncertainties affecting your risk assessment.

### Determine inherent risk rating

Use the risk matrix provided in **Table 2** of the impact assessment tool to determine the risk rating for each category.

### Provide explanations

Provide clear and concise explanations for each risk rating.

## Key considerations

When completing the inherent risk assessment, keep the following in mind:

- Try to be objective and honest in your assessment of risks. Underestimating risks at this stage could lead to inadequate risk management.
- Determining risk ratings can be challenging. Seek guidance from others to assist you, especially subject matter experts and those experienced in safe and responsible AI risk assessments.
- Consider the perspectives of stakeholders, including those identified at section 2.4, in assessing the likelihood and consequence of risks.
- Consider the perspectives of marginalised groups, including First Nations people, especially in relation to risks relating to discrimination and stereotyping. You may not have the background or life experience to fully appreciate these risks.
- Consider both intended and unintended consequences and outcomes. This includes evaluating the impact of system failure, the impact of misuse or malicious use and other deviations from expected use.
- Where there is uncertainty or disagreement about the appropriate inherent risk rating, choose the higher rating.
- Document key assumptions or evidence used in determining the inherent risk rating, as this will help explain the rationale for your assessment to reviewers.
- Consider the expected benefits of the AI use case before deciding whether to proceed based on significant but mitigable risks.

# 3.9 Overall inherent risk rating

In this section, you are required to determine the threshold risk rating for the AI use case based on the ratings selected in the sections above. The highest risk rating identified in any earlier sections must be used as the overall risk rating.

# Resources

CSIRO Data61 Responsible AI Pattern Catalogue

Massachusetts Institute of Technology (MIT) AI Risk Repository

United States National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) and AI RMF Playbook

# 4. Threshold assessment outcome

## 4.1 Assessing officer recommendation

Once completed, if the assessing officer is satisfied all risks are low, they may recommend that a full assessment is not required and that the approving officer accept the low risks and endorse the use case.

If one or more risks are medium or higher, the assessing officer must either:

- complete a full assessment
- amend the scope, or function to a point where the threshold assessment results in a low risk rating
- decide to not accept the risk and not proceed with the AI use case.

## 4.2 Approving officer review

Once the assessing officer has made their recommendation, the approving officer must:

- review the recommendation
- confirm whether they are satisfied by the supporting analysis
- agree that a full assessment is or is not necessary for the use case.

# 5. Fairness

## 5.1 Defining fairness

Fairness is a core principle in the design and use of AI systems, but it is a complex and contextual concept. Australia's AI Ethics Principles state that AI systems should be inclusive and accessible and should not involve or result in unfair discrimination. However, there are different and sometimes conflicting definitions of fairness and people may disagree on what is fair.

For example, there is a distinction between:

- individual fairness – treating individuals similarly
- group fairness – similar outcomes across different demographic groups.

Different approaches to fairness involve different trade-offs and value judgments. The most appropriate fairness approach will depend on the specific context and objectives of your AI use case.

When defining fairness for your AI use case, you should be aware that AI models are typically trained on broad sets of data that may contain bias. Bias can arise in data where it is incomplete, unrepresentative or reflects societal prejudices. AI models may reproduce biases present in the training data, which can lead to misleading or unfair outputs, insights or recommendations. This may disproportionally impact some groups, such as First Nations people, people with disability, LGBTIQ+ communities and culturally and linguistically diverse communities. For example, an AI tool used to screen job applicants might systematically disadvantage people from certain backgrounds if trained on hiring data that reflects past discrimination.

When defining fairness for your AI use case, consider the inclusivity and accessibility of the AI. AI can lead to unfairness if it creates barriers for individuals or groups who wish to access government services. For example, an AI chatbot designed to provide social security information may produce unfair outcomes because it is more difficult for vulnerable or underrepresented groups to access the digital technologies required to access the chatbot.

When defining fairness for your AI use case, it is recommended that you:

- consult relevant domain experts, affected parties and stakeholders (such as those you have identified at assessment section 2.4) to help you understand the trade-offs and value judgements that may be involved
- document your definition of fairness in your response to assessment section 5.1, including how you have balanced competing priorities and why you believe it to be appropriate to your use case
- be transparent about your fairness definition and be open to revisiting it based on stakeholder feedback and real-world outcomes.

You should also ensure that your definition of fairness complies with anti-discrimination laws. In Australia, 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. Australia's federal anti-discrimination laws are contained in the following legislation:

- *Age Discrimination Act 2004*
- *Disability Discrimination Act 1992*
- *Racial Discrimination Act 1975*
- *Sex Discrimination Act 1984.*

Where the AI will produce information or be involved in decision-making, you should also ensure that your definition of fairness reflects the administrative law principle of procedural fairness, which requires that decision-making is transparent and challengeable.

## 5.2 Measuring fairness

You may be able to use a combination of quantitative and qualitative approaches to measuring fairness. Quantitative fairness metrics can allow you to compare outcomes across different groups and assess this against fairness criteria. Qualitative assessments, such as stakeholder engagement and expert review, can provide additional context and surface issues that metrics alone might miss.

### Quantifying fairness

The specific quantitative metrics you use to measure fairness will depend on the definition of fairness you have adopted for your use case. When selecting fairness metrics, you should:

- choose metrics that align with your fairness definition, recognising the trade-offs between different fairness criteria and other objectives like accuracy
- confirm if you have appropriate data to assess those metrics, including compliance with the Australian Privacy Principles where personal or sensitive information is being collected and used
- set clear and measurable acceptance criteria (see guidance for section 6.4)
- establish a plan for monitoring these metrics (see section 6.6) and processes for remediation, intervention or safely disengaging the AI system if those thresholds are not met

For examples of commonly used fairness metrics, see the Fairness Assessor Metrics Pattern from the CSIRO's Data61 unit.

## Qualitatively assessing fairness

Consider some of these qualitative approaches, which may be useful to overcome data limitations and to surface issues that metrics may overlook.

### Stakeholder engagement

Consult affected communities, stakeholders and domain experts to understand their perspectives and identify potential issues

### User testing and feedback

Test your AI system with diverse users and solicit their feedback on the fairness and appropriateness of the system's outputs. Seek out the perspectives of marginalised groups and groups that may be impacted by the AI system.

### Expert review

Engage experts, such as AI ethicists or accessibility and inclusivity specialists, to review the fairness of your system's outputs and the overall approach to fairness. Identify potential gaps or unintended consequences.

# Resources

For advice on bias measurement and minimisation techniques, see the National AI Centre's report on Implementing Australia's AI Ethics Principles.

CSIRO Data61's Responsible AI Pattern Catalogue includes a Fairness Assessor Metrics Pattern.

Consider resources on fairness in AI in the OECD Catalogue of Tools & Metrics for Trustworthy AI.

# 6. Reliability and safety

## 6.1 Data suitability

The data used to operate, train and validate your AI system has a significant impact on its performance, fairness and safety. In your answer, explain why the chosen data is suitable for your use case. Some relevant considerations are outlined below.

When choosing between datasets, consider whether the data can be separated by marginalised groups, particularly by Indigenous status identifiers. If the data is Indigenous data, see section 6.2 below, which references the Framework for Governance of Indigenous Data.

Agencies should also refer to the Australian Public Service (APS) Data Ethics Framework for guidance on managing and using data and analytics ethically in government, including where AI is used in analytics. The framework is underpinned by 3 key principles: trust, respect and integrity. It provides advice on implementation across different major use cases and agency operations and encourages agencies to assess potential risks and benefits, consider fairness and inclusivity, and engage with stakeholders where appropriate. Visit the Department of Finance website to access the APS Data Ethics Framework.

Data quality should be assessed prior to use in AI systems. Agencies should select applicable metrics to determine a data set's quality and identify any remediation required before using it for training or validation in AI systems. Relevant metrics to consider include diversity, relevance, accuracy, completeness, timeliness, validity and lack of duplication. One method to ensure good quality data is to set minimum thresholds appropriate to specific use cases, such as through acceptance criteria discussed below at section 6.4. An example of a specific framework for determining data quality in statistical uses is the ABS Data Quality Framework.

Where third party material or data is being used to operate, train or validate an AI system, it is important to protect the rights of intellectual property holders. If the AI may use, modify or otherwise handle material in which intellectual property exists, agencies should confirm that both the following are true:

- the AI provider holds the necessary intellectual property rights in the AI output material
- the agency holds the necessary intellectual property rights in the input material.

The AI may otherwise infringe third party intellectual property rights.

Agencies should also confirm that the AI system has safeguards in place to prevent the unauthorised use or disclosure of confidential information.

Where data used to operate, train and validate the AI system includes personal information, agencies should confirm that collection, use and disclosure is in accordance with the Australian Privacy Principles (APPs) under the *Privacy Act 1988* (Cth) (see section 7 below).

The relevance of the data used in training the AI model may influence the output and may not be relevant to the use case and Australian context. Consider whether the model is likely to make accurate or reliable predictions concerning matters relating to Australian subject matter if it has been trained on, for example, US-centric data.

You should also consider data provenance, lineage and volume – as outlined below:

### Data provenance

Involves keeping records of the data collected, processed and stored by the AI system and creating an audit trail to assign custody and trace accountability for issues. It provides assurance of the chain of custody and its reliability, insofar as origins of the data are documented.

### Data lineage

Involves documenting data origins and flows to enable stakeholders to better understand how datasets are constructed and processed. This fosters transparency and trust in AI systems.

### Data volume

Consider the volume of data you need to support the operation, training and validation of your AI system.

## 6.2 Indigenous data

Describe how any components of your AI system have used or will use Indigenous data, or where any outputs relate to First Nations individuals, communities or groups.

All Australian Public Service (APS) agencies are required to implement the Framework for Governance of Indigenous Data. This framework adopts the definition of 'Indigenous data' as provided by Maiam nayri Wingara Indigenous Data Sovereignty Collective:

> Information or knowledge, in any format or medium, which is about and may affect Indigenous peoples both collectively and individually.

If the data used to operate, train or validate your AI system, or any outputs from your AI system, meet this definition of Indigenous data, refer to the Framework for Governance of Indigenous Data for guidance on applying the framework.

The framework is based on the principles of:

- respect for cultural heritage
- informed consent
- privacy, including collective or group privacy
- trust.

The Framework for Governance of Indigenous Data is also informed by 2 complementary data governance frameworks:

- FAIR Guiding Principles (Findable, Accessible, Interoperable, Reusable) – providing technical standards for scientific data management and stewardship
- CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics) – which focuses on Indigenous data governance, reflecting the crucial role of data in self-determination.

Relevant practices to consider in this context include:

- Checking if datasets used to train the AI included diverse and representative samples of cultural expression, artifacts, languages and practices. This supports the AI system being able to recognise and appropriately respond to a greater range of cultural contexts in a less biased manner.
- Describing any mechanisms in place for engaging with Indigenous individuals, communities or group representatives and collecting and incorporating their feedback on the AI system's performance, especially regarding cultural aspects.
- Describing processes to review documentation and protocols that ensure the project has incorporated the GID principles. Look for evidence of meaningful engagement with and input from suitably qualified and experienced Indigenous individuals, communities and groups. Assess if the system includes features or options that allow Indigenous stakeholders to control how their data is used and represented and describe how benefits of the project to First Nations Peoples, to which the data relate, have been considered.

Also consider the use of Indigenous data in the context of the United Nations Declaration on the Rights of Indigenous Peoples and apply the concept of 'free, prior and informed consent' in relation to the use of Indigenous data in AI systems.

## 6.3 Suitability of procured AI model

If you are procuring an AI model or system from a third-party provider, your procurement process should consider if the provider has appropriate data management including data quality and data provenance in relation to the model. This will help you to identify whether the AI model is fit for the context and purpose of your AI use case.

This may include:

- governance
- data sourcing
- privacy
- security
- intellectual property
- cybersecurity practices.

There are many other considerations you should take into account when selecting a procured AI model and contracting with a supplier. The following considerations may be relevant to your use case:

- Determine if data will be hosted overseas and if it could be subject to foreign laws. Consider the potential for foreign ownership, control, or influence (FOCI) and refer to the Department of Home Affairs FOCI Risk Assessment Guidance.
- Determine if processes and practices are in place to address risks along the supplier's supply chain, such as risks relating to FOCI, security, transparency and business practices). Agencies should refer to Australian Signals Directorate guidance on cyber supply chain risk management.
- Assess whether the AI model meets the functional requirements for your use case.
- Determine how the model was evaluated, including the test data and benchmarks used.
- Determine how versioning for the AI model is handled.
- Consider the support the supplier provides for users and procurers.
- Review provisions regarding potential liability issues and clarify accountability between your agency and the provider if the product fails.
- Establish security precautions, such as handover or destruction of agency data upon termination or expiry of the procurement contract, and identify any residual risks and mitigation measures.
- Confirm what controls the agency has if the AI system malfunctions, produces harmful outputs or behaves in an unintended way.
- Review any guarantees that data handling and management across the entire lifecycle of the data meet internal agency and legislative requirements.
- Review any warranties the supplier will provide, such as suitability of the AI system for the intended use, absence of defects, and development with reasonable care and skill.
- Ensure the supplier has a contractual obligation to comply with relevant legislation and frameworks, including privacy, discrimination, AI Ethics Principles, the AI policy and the Protective Security Policy Framework (PSPF).
- Clarify any supplier responsibilities for training, monitoring and validation of the AI system.
- Clarify ownership of intellectual property rights in relation to the AI model, inputs, outputs and other materials such as user manuals or technical documentation.
- Review measures taken to prevent or reduce hallucinations, unwanted bias and model drift. For example, evaluation of training data for harm or bias and adjustments made to compensate.
- Assess whether the level of human oversight, transparency, explainability and interpretability of the model is sufficient for your use case.
- Specify the kinds of records the supplier will provide to the agency, such as records of how agency data is used by the AI system.
- Determine the computing and storage capacity requirements for operating the model on premises.
- Assess the capability needed to maintain the AI model and whether this can be done in-house or require external sourcing.

- If using a platform as a service (PaaS) to run and support your AI system or AI model, consider risks associated with outsourcing.
- Evaluate whether the AI system could be designed or influenced to promote certain products or services, and how such behaviour could be detected and addressed. For example, if the supplier accepts advertising or sponsorship to give prominence to products or services.

Consider also how your agency will support transparency across the AI supply chain, for example, by notifying the developer of issues encountered in using the model or system.

Refer to the DTA's AI procurement resources including the:

- Guidance on AI procurement in government
- AI contract template
- Digital Sourcing ClauseBank AI model clauses.

## 6.4 Testing

Testing is a key element for assuring the responsible and safe use of AI models – for both models developed in-house and externally procured – and in turn, of AI systems. Rigorous testing helps validate that the system performs as intended across diverse scenarios. Thorough and effective testing helps identify problems before deployment.

Testing AI systems against test datasets can reveal biases or possible unintended consequences or issues before real-world deployment. Testing on data that is limited or skewed can fail to reveal shortcomings.

Consider establishing clear and measurable acceptance criteria for the AI system that, if met, would be expected to control harms that are relevant in the context of your AI use case. Acceptance criteria should be specific, objective and verifiable. They are meant to specify the conditions under which a potential harm is adequately controlled.

Consider developing a test plan for the acceptance criteria to outline the proposed testing methods, tools and metrics. Documenting results through a test report will assist with demonstrating accountability and transparency. A test report could include the following:

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

In your explanation, outline any areas of concern in results from testing. If the AI system has not yet undergone testing, outline elements to be considered in testing plans.

# Model accuracy

As an example. model accuracy is a key metric for evaluating the performance of an AI system. Accuracy should be considered in the specific context of the AI use case, as the consequences of errors or inaccuracies can vary significantly depending on the domain and application. This can include:

- unfairness – for example, where a decision has been made based on inaccurate data
- breach of individual rights – for example, where information produced by AI is defamatory
- non-compliance with legislation – for example, presenting false or misleading information in breach of Australian Consumer Law, or acting in a discriminatory manner in breach of anti-discrimination laws.

Some of the factors that can influence AI model output accuracy and reliability include:

- choice of AI model or model architecture
- quality, accuracy and representativeness of training data
- presence of bias in the training data or AI model
- robustness to noise, outliers and edge cases
- ability of the AI model to generalise to new data
- potential for errors or 'hallucinations' in outputs
- environmental factors (such as lighting conditions for computer vision systems)
- adversarial attacks (such as malicious actors manipulating input data to affect outputs)
- stability and consistency of performance over time
- whether AI model allows for sponsorship or advertising to give prominence to certain outputs.

Ways to assess and validate the accuracy of your model for your AI use case include:

- quantitative metrics
- qualitative analysis – such as manual review of output, error analysis, and user feedback
- domain-specific benchmarks or performance standards
- comparison to human performance or alternative models.

It is important to set accuracy targets that are appropriate for the risk and context of the use case. For high stakes decisions, you should aim for a very high level of accuracy and have clear processes for handling uncertain or borderline cases.

## 6.5 Pilot

Conducting a pilot study is a valuable way to assess the real-world performance and impact of your AI use before full deployment. A well-designed pilot can surface issues related to reliability, safety, fairness and usability that may not be apparent in a controlled development environment.

If you are planning a pilot, your explanation should provide a brief overview of the pilot's:

- scope and duration
- objectives and key results (OKRs)
- key performance indicators (KPIs)
- participant selection and consent process
- risk mitigation strategies.

If you have already completed a pilot, reflect on the key findings and lessons learned, including by:

- assessing how the pilot outcomes compared to your expectations.
- identifying any issues or surprises that emerged during the pilot.
- documenting how you adapted your AI use case based on the pilot results.

If you are not planning to conduct a pilot, explain why not. Consider whether the scale, risk or novelty of your use case warrants a pilot phase. Discuss alternative approaches you are taking to validate the performance of your AI use case and gather user feedback prior to full deployment.

## 6.6 Monitoring

Monitoring is key to maintaining the reliability and safety of AI systems over time. It enables active rather than passive oversight and governance, and ensures the agency has ongoing accountability for the AI-assisted performance and decision-making processes.

Your monitoring plan should be tailored to the specific risks and requirements of your use case. In your explanation, describe your approach to monitoring any measurable acceptance criteria (as discussed above at section 6.4) and other relevant metrics such as performance metrics or anomaly detection. In your plan, include your proposed monitoring intervals for your use case. The AI policy requires agencies to establish a clear process to address AI incidents aligned to their ICT management approach. Incident remediation must be overseen by an appropriate governance body or senior executive and should be undertaken in line with any other legal obligations.

Periodically evaluate your monitoring and evaluation mechanisms to ensure they remain effective and aligned with evolving conditions throughout the lifecycle of your AI use case. Examples of events that could influence your monitoring plan are system upgrades, error reports, changes in input data, performance deviation or feedback from stakeholders.

Monitoring can help identify issues that can impact the safety and reliability of your AI system, such as:

- concept drift – changes in the relationship between input data and the feature being predicted
- data drift – changes in input data patterns compared to the data used to train the model.

Vendors offer monitoring tools that may be worth considering for your use case. For more information on continuous monitoring, refer to the NAIC's Implementing Australia's AI Ethics Principles report.

## 6.7 Preparedness to intervene or disengage

Relevant stakeholders, including those who operate, use or interact with the AI system, those who monitor AI system performance, and affected stakeholders identified at section 2.4, should have the ability to raise concerns about insights or decisions assisted by the AI system.

Agencies must develop clear pathways for staff or other relevant stakeholders to report AI safety concerns, including AI incidents. Agencies should also document and take appropriate steps in relation to any interventions that occur to ensure consistency and fairness.

In addition, agencies should be prepared to quickly and safely disengage an AI system when an unresolvable issue is identified. This could include a data breach, unauthorised access or system compromise. Consider such scenarios in business continuity, data breach and security response plans.

### Techniques to avoid overreliance on AI system outputs

Agencies should consider the following techniques to avoid overreliance on AI system outputs.

Three techniques to consider at the **system design stage**:

- Build in transparency about system limitations, by incorporating prompts to remind users to critically analyse outputs. These could include explanations of outputs, hallucination reminders, reference source checking and accuracy scores.
- Build in 2-way feedback pathways by prompting users to assess the quality of the AI system's outputs and provide feedback. Similarly, provide feedback to users on their interactions with the systems, such as feedback on ineffective prompts or alerts when the user has accepted a risky decision.
- Build in steps that require human decision-making, for example by designing the AI system to provide options to choose from rather accept a single outcome, prompting users to engage with and evaluate AI outputs.

At the **evaluation stage**, focus on validating whether the system supports human judgement as intended. Engage directly with users to understand their experience, encourage them to assess outputs critically and suggest improvements. Review user behaviour, feedback loops and decision-making patterns and prompts to confirm that safeguards against overreliance are effective. Use these insights to refine system design, guidance and training materials.

# 6.8 Training of AI system operators

AI system operators play a crucial role in ensuring the responsible and effective use of AI. They must have the necessary skills, knowledge and judgment to understand the system's capabilities and limitations, how to appropriately use the system, interpret its outputs and make informed decisions based on those outputs.

In your answer, describe the process for ensuring AI system operators are adequately trained and skilled. This may include:

## Initial training

Consider what training operators receive before being allowed to use the AI system. Does this training cover technical aspects of the system, as well as ethical and legal considerations?

As a baseline, you may expect that operators:

- understand the limitations of the AI system
- are able to monitor the AI system, so that anomalies, errors and unexpected performance can be detected and addressed
- are aware of the possible tendency of relying, or over-relying, on AI outputs
- are able to correctly interpret AI outputs, taking into account the particular characteristics of the system
- are able to decide when to disregard, override or reverse the AI outputs.

## Ongoing training

This includes processes for continuous learning and skill development, and for keeping officers up to date with changes or updates to the AI system.

## Evaluation

This can include skills and knowledge assessment, certification or qualification requirements for operators.

## Support

Ensure resources and support are available to operators if they have questions or encounter issue. Consider whether this needs to be tailored to the specific needs and risks of your AI system or proposed use case or whether general AI training requirements are sufficient.

# 7. Privacy protection and security

## 7.1 Minimise and protect personal information

### Compliance with the Australian Privacy Principles

Agencies should consider how the AI use case will comply with the Australian Privacy Principles (APPs) in Schedule 1 to the *Privacy Act 1988* (Cth). The APPs apply to personal information inputted into an AI system, as well as the output generated or inferred by an AI system that contains personal information. Under the APPs:

- **APP 1:** Agencies must implement practices, procedures and systems to ensure compliance with APPs. Agencies must also have a clearly expressed and up-to-date privacy policy. This can include establishing clear processes for verification of AI outputs containing personal information, and adding transparent information about its use of AI in its privacy policy.
- **APP 3:** AI inputs or outputs generated or inferred by AI, which contain personal information, must be reasonably necessary for, or directly related to, the agency's functions or activities. Additionally, if the AI input or output comprises sensitive personal information, the individual must consent unless another exception applies. Collection must occur by fair and lawful means.
- **APP 5:** Agencies should notify individuals of AI-related purposes for which their personal information is being collected and any proposed use of AI to generate outputs which contain personal information.
- **APP 6:** Agencies may only input an individual's personal information into an AI system, or use or disclose AI outputs which contain personal information, for the primary purpose for which the agency collected the information, unless they have consent or another exception applies – for example, if the agency can establish a related secondary use would be reasonably expected by the individual.
- **APP 10:** Agencies must take reasonable steps to ensure personal information collected, used and disclosed by the AI system is accurate, up-to-date, complete and relevant.
- **APP 11:** Agencies must take reasonable steps to protect personal information from misuse, interference and loss, as well as unauthorised access, modification or disclosure.

For more information, refer to the APP guidelines and the Office of the Australian Information Commissioner (OAIC) Guidance on privacy and the use of commercially available AI products.
Also consider your agency's internal privacy policy and resources and consult your agency's privacy officer.

## Privacy enhancing technologies

Your agency may want or need to use privacy enhancing technologies to assist in de-identifying personal information under the APPs or as a risk mitigation/trust building approach. Where the risk of re-identification is very low, de-identified information will no longer comprise personal information and agencies can use the information in ways that the Privacy Act would normally restrict.

Consider the Office of the Australian Information Commissioner's (OAIC) detailed guidance on De-identification and the Privacy Act. The OAIC has also jointly developed a resource with CSIRO Data61 on De-identification Decision-Making Framework.

# 7.2 Privacy Threshold and/or Impact Assessment

The Australian Government Agencies Privacy Code (the Privacy Code) requires Australian Government agencies subject to the *Privacy Act 1988* to conduct a privacy impact assessment (PIA) for all 'high privacy risk projects'. A project may be a high privacy risk if the agency reasonably considers that the project involves new or changed ways of handling personal information that are likely to have a significant impact on the privacy of individuals.

To determine whether a PIA is required, you should complete a privacy threshold assessment (PTA). A PTA will help you identify your use case's potential privacy impacts and screen for factors that point to a 'high privacy risk project' requiring a PIA under the Code.

Agencies should conduct a PTA and, if required, a PIA at an early stage of AI use case development or procurement– for example, after identifying the minimum viable product. This will enable the agency to fully consider whether to proceed with the AI use case or to change the approach if the PIA identifies significant negative privacy impacts. It may be appropriate to conduct a PTA and, if required, a PIA earlier than your AI impact assessment using this tool.

If you have not completed a PTA or PIA, explain how you considered potential privacy impacts – for example, if you have determined the AI use case will not involve personal information. Privacy assessments should consider if relevant individuals have provided informed consent, where required, to the collection, use and disclosure of their personal information in the AI system's training or operation, or as an output for making inferences. Also consider any consent obtained has been recorded, including a description of processes used to obtain the consent.

For more information, refer to the Office of the Australian Information Commissioner (OAIC) advice for Australian Government agencies on when to conduct a privacy impact assessment. You can also consult your agency's privacy officer and internal privacy policy and resources.

If your AI system has used or will use Indigenous data, you should also consider whether principles of collective or group privacy of First Nations people are relevant and refer to the Framework for Governance of Indigenous Data (see section 6.2).

## 7.3 Security risks

Agencies should consider the digital and cyber security risks associated with operation of the AI. Agencies may wish to refer to the frameworks and guidance noted below in considering what measures the AI will have in place to address security risks.

The Protective Security Policy Framework (PSPF) applies to non-corporate Commonwealth entities subject to the *Public Governance, Performance and Accountability Act 2013* (PGPA Act). Agencies should refer to the PSPF to understand security requirements relevant to AI technologies. These include managing procurement risks, incorporating and enforcing security terms in contracts, addressing FOCI risks, protecting classified information, and ensuring systems are authorised in accordance with the Information Security Manual (ISM).

You should engage with your agency's ITSA early in the AI use case development and assessment process to ensure it meets all PSPF and ISM requirements.

Agencies should implement security measures to align with Australian Signals Directorate (ASD) guidance on AI data security. This outlines data security risks in the development, testing and deployment of AI, and sets out best practices for securing AI data across stages of the AI lifecycle to address these risks.

Agencies should ensure appropriate procedures are in place to address a data breach or security incident. This may include processes to mitigate the immediate consequences of a data breach or security incident and to ensure any actual or potential ongoing loss to the agency is minimised.

For further mitigation considerations for organisations to consider refer to ASD's guidance on Engaging with AI. It is highly recommended that your agency engages with and implements the mitigation considerations in the guidance. This includes:

- enforcing multi-factor authentication or privileged access for AI systems
- managing backups of the AI system and training data
- ensuring the AI system is secure-by-design, including across its supply chain
- conducting periodic health checks on the AI system.

Agencies should also consider the requirements outlined in the Department of Home Affairs PSPF Policy Advisory on OFFICIAL Information Use with Generative AI. These include only providing access to certain generative AI products that meet hosting and other security criteria and ensuring staff have relevant training.

# 8. Transparency and explainability

## 8.1 Consultation

You should consult with a diverse range of internal and external stakeholders at every stage of your AI use case development and deployment to help identify potential biases, privacy concerns, and other ethical and legal issues present in your AI use case. This process can also help foster transparency, accountability, and trust with your stakeholders and can help improve their understanding of the technology's benefits and limitations. Refer to the stakeholders you identified in section 2.4.

If your project has the potential to significantly impact First Nations individuals, communities or groups, it is critical that you meaningfully consult with relevant community representatives.

### Consultation resources

APS Framework for Engagement and Participation

Sets principles and standards that underpin effective APS engagement with citizens, community and business and includes practical guidance on engagement methods.

Best practice consultation guidance note

This resource from the Office of Impact Analysis details the Australian Government consultation principles outlined in the Guide to Policy Impact Analysis.

Principles for engagement in projects concerning Aboriginal and Torres Strait Islander peoples

This resource from the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) provides non-Indigenous policy makers and service designers with the foundational principles for meaningfully engaging with Aboriginal and Torres Strait Islander peoples on projects that impact their communities.

## 8.2 Public visibility

Where appropriate, you should consider options to make the scope and goals of your AI use case publicly available. For instance, consider including this information on the relevant program page on your agency website or through other official communications. This information could include:

- use case purpose
- overview of model and application, including how the AI will use data to provide relevant outputs
- benefits

- risks and mitigations
- training data sources
- contact information for public enquiries.

All agencies in scope of the AI policy are required to publish an AI transparency statement. Your agency's AI accountable official is responsible for ensuring your agency's transparency statement complies with the AI policy. More information on this requirement is contained in the AI policy and associated Standard for transparency statements. Consult your agency's AI accountable official for specific advice on your use case.

Furthermore, to comply with APP 1 and APP 5, agencies should consider updating their privacy policies with information about their use of AI. For example, to advise that personal information may be disclosed to AI system developers or owners.

## Considerations for publishing

In some circumstances it may not be appropriate to publish detailed information about your AI use case. When deciding whether to publish this information you should balance the public benefits of AI transparency with the potential risks as well as compatibility with any legal requirements around publication.

For example, you may choose to limit the information you publish, or not publish any information at all, if the use case is still in the experimentation phase, or if publishing may:

- have negative implications for national security
- have negative implications for law enforcement or criminal intelligence activities
- significantly increase the risk of fraud or non-compliance
- significantly increase the risk of cybersecurity threats
- jeopardise commercial competitiveness – for example, revealing trade secrets or commercially valuable information
- breach confidentiality obligations held by the agency under a contract
- breach statutory secrecy provisions.

# 8.3 Maintain appropriate documentation and records

Agencies should comply with legislation, policies and standards for maintaining reliable and auditable records of decisions, testing, and the information and data assets used in an AI system. This will enable internal and external scrutiny, continuity of knowledge and accountability. For example, when responding to information requests under the *Freedom of Information Act 1982* (Cth). This will also support transparency across the AI supply chain. For example, this documentation may be useful to any downstream users of AI models or systems developed by your agency.

Agencies should document AI technologies they are using to perform government functions as well as essential information about AI models, their versions, creators and owners. In addition, artefacts used and produced by AI – such as prompts, inputs and raw outputs – may constitute Commonwealth records under the *Archives Act 1983* and may need to be kept for certain periods of time identified in records authorities issued by the National Archives of Australia (NAA). Such Commonwealth records must not be destroyed, disposed of, transferred, damaged or altered except in limited circumstances listed in the Archives Act.

To identify their legal obligations, business areas implementing AI in agencies may want to consult with their information and records management teams. The NAA can also provide advice on how to manage data and records produced by different AI use cases.

Refer to NAA advice on:

- principles and expectations for the creation and management of government business information, contained in the Information Management Standard for Australian Government
- Information management for records created using AI technologies.

## AI documentation types

Where suitable, you should consider creating the following forms of documentation for any AI system you build. If you are procuring an AI system from an external provider, it may be appropriate to request these documents as part of your tender process.

### System factsheet/model card

A system factsheet (sometimes called a model card) is a short document designed to provide an overview of an AI system to non-technical audiences (such as users, members of the public, procurers, and auditors). These factsheets usually include information about the AI system's purpose, intended use, limitations, training data, and performance against key metrics.

### Datasheets

Datasheets are documents completed by dataset creators to provide an overview of the data used to train and evaluate an AI system. Datasheets provide key information about the dataset including its contents, data owners, composition, intended uses, sensitivities, provenance, labelling and representativeness.

### System decision registries

System decision registries record key decisions made during the development and deployment of an AI system. These registries contain information about what decisions were made, when they were made, who made them and why they were made (the decision rationale).

### Reliability and safety documentation

It is also best practice to maintain documentation on testing, piloting and monitoring and evaluation of your AI system and use case, in line with the practices outlined in section 6.

For more on AI documentation, see Implementing Australia's AI Ethics Principles.

## 8.4 Disclosing AI interactions and outputs

You should design your use case to inform people that they are interacting with an AI system or are being exposed to content that has been generated by AI. This includes disclosing AI interactions and outputs to internal agency staff and decision-makers, as well as external parties such as members of the public engaging with government.

### When to disclose use of AI

You should ensure that you disclose when a user is **directly interacting with an AI system**, especially:

- when AI plays a significant role in critical decision-making processes
- when AI has potential to influence opinions, beliefs or perceptions
- where there is a legal requirement regarding AI disclosure (for example, updated privacy policies under APP 1 and APP 5)
- where AI is used to generate recommendations for content, products or services.

You should ensure that you disclose when someone is being exposed to **AI-generated content** including where:

- any of the content has not been through a contextually appropriate degree of fact checking and editorial review by a human with the appropriate skills, knowledge or experience in the relevant subject matter
- the content purports to portray real people, places or events or could be misinterpreted that way

- the intended audience for the content would reasonably expect disclosure
- there is a legal requirement regarding AI disclosure (for example, updated privacy policies under APP 1 and APP 5).

Exercise judgment and consider the level of disclosure that the intended audience would expect, including where AI-generated content has been through rigorous fact-checking and editorial review. Err on the side of greater disclosure – norms around appropriate disclosure will continue to develop as AI-generated content becomes more ubiquitous.

## Mechanisms for disclosure of AI interactions:

When designing or procuring an AI system, you should consider the most appropriate mechanism(s) for disclosing AI interactions. Some examples are outlined below:

### Verbal or written disclosures

Verbal or written disclosures are statements that are heard by or shown to users to inform that they are interacting with (or will be interacting with) an AI system.

For example, disclaimers/warnings, specific clauses in privacy policy and/or terms of use, content labels, visible watermarks, by-lines, physical signage, communication campaigns.

### Behavioural disclosures

Behavioural disclosure refers to the use of stylistic indicators that help users to identify that they are engaging with AI-generated content. These indicators should generally be used in combination with other forms of disclosure.

For example, using clearly synthetic voices or formal, structured language, robotic avatars.

### Technical disclosures

Technical disclosures are machine-readable identifiers for AI-generated content.

For example, inclusion in metadata, technical watermarks, cryptographic signatures.

Agencies should consider using AI systems that use industry-standard provenance technologies, such as those aligned with the standard developed by the Coalition for Content Provenance (C2PA).

## Ability to request a non-AI alternative

In certain contexts, it may be best practice not to provide a non-AI alternative, particularly where the AI system is low-risk, improves service delivery without affecting rights or entitlements, and where alternate pathways would create unnecessary cost, complexity, or delay. However, in other situations, offering the ability to request a non-AI alternative can be important.

# 8.5 Offer appropriate explanations

Explainability refers to accurately and effectively conveying an AI system's decision process to a stakeholder, even if they don't fully understand the specifics of how the model works. Explainability facilitates procedural fairness, transparency, independent expert scrutiny and access to justice by ensuring that agencies have the material that is required to provide affected individuals with evidence that forms the basis of a decision when needed. To interpret the AI's output and offer an explanation to relevant stakeholders, you should consider whether the agency can access:

- the inputs from the agency
- the logic behind an individual output
- the model that the AI System uses and the sources of data for the model
- information on which features of the AI contributed to the output
- automatic records of events which allow for traceability of the AI's functioning
- any risk management measures in place which would allow the agency to understand risks and adjust use of the AI accordingly (for example, technical limitations such as error rates of an AI model).

You should be able to clearly explain how a government decision or outcome has been made or informed by AI to a range of technical and non-technical audiences. You should also be aware of any requirements in legislation to provide reasons for decisions, both generally and in relation to the particular class of decisions that you are seeking to make using AI.

Explanations may apply globally (how a model broadly works) or locally (why the model has come to a specific decision). You should determine which is more appropriate for your audience.

## Principles for providing effective explanations

### Contrastive

Outline why the AI system output one outcome instead of another outcome.

### Selective

Focus on the most-relevant factors contributing to the AI system's decision process.

### Consistent with the audience's understanding

Align with the audience's level of technical (or non-technical) background.

### Generalisation to similar cases

Generalise to similar cases to help the audience predict what the AI system will do.

## Tools for explaining non-interpretable models

Providing explanations is relatively straightforward for interpretable models with low complexity and clear parameters. However, in practice, most AI systems have low interpretability and require effective post-hoc explanations that balance accuracy and simplicity. Among other matters, you should also consider defining appropriate timeframes for providing explanations in the context of your use case.

When developing explanations, consider the range of available approaches based on your model type and use case.

- For traditional machine learning models, feature importance methods and visualisation techniques can help explain individual predictions or overall model behaviour.
- For neural networks and deep learning systems, specialised interpretation methods have been developed that analyse network activations, attention patterns, and gradients.
- Large language models and foundation models require distinct approaches, including prompt-based explanations and emergent interpretability techniques.
- Model-agnostic methods offer flexibility across different architectures, while example-based approaches use counterfactuals and contrastive examples to make predictions more understandable.

Advice on appropriate explanations is available in the National AI Centre's Implementing Australia's AI Ethics Principles report.

Other reputable resources for explainability tools include open-source libraries maintained by academic institutions and research communities and documentation from major cloud platform providers. When selecting tools, prioritise those with active maintenance, clear documentation, and validation through published research.

However, explainable AI algorithms are not the only way to improve system explainability. Human-centred design can also play an important part, including:

- developing effective explanation interfaces tailored to different stakeholder audiences
- determining appropriate levels of detail for various contexts
- ensuring explanations are actionable and meaningful for decision-makers

# 9. Contestability

## 9.1 Notification of AI affecting rights

You should notify individuals, groups, communities or businesses when an administrative action materially influenced by an AI system has a legal or significant effect on them. This promotes transparency and access to justice, by ensuring individuals can understand how government uses AI to perform actions that affect them and have the opportunity to seek review of that decision.

This notification should state that the action was materially influenced by an AI system and include information on available review rights and how the individual can challenge the action. The notification should be clear, up-to-date, concise and understandable, and should not be complex, lengthy, legalistic or vague. It may be appropriate to provide notification prior to the action being taken or at the same time that the action occurs (for example, an applicant may be asked to acknowledge that AI will be used to a stated extent to assess their application).

An action producing a **'legal effect'** is when an individual, group, community or business's legal status or rights are affected, and includes an effect on the:

- provision of rights or benefits granted by legislation or common law
- imposition of penalties or orders (civil or criminal), and
- contractual rights.

An action producing a **'significant effect'** is when an individual, group, community or business's circumstances, behaviours, interests or choices are affected, and includes an effect on the provision of:

- critical government services or support, such as housing, insurance, education enrolment, criminal justice, employment opportunities and health, disability or aged care services
- basic necessities, such as food and water.

An action may be considered to have been **'materially influenced'** by an AI system if:

- the action was automated by an AI system, with little to no human oversight
- a component of the action was automated by an AI system, with little to no human oversight – for example, a computer performs the first 2 limbs of an action, with the final limb made by a human
- the AI system is likely to influence actions that are performed – for example, the AI system output recommended a decision to a human for consideration or provided substantive analysis to inform a decision.

**'Administrative action'** is any of the following:

- making, or refusing or failing to make, a decision
- exercising, or refusing or failing to exercise, a power
- performing, or refusing or failing to perform, a function or duty.

### Advisory note

This guidance is designed to supplement, not replace, existing administrative law requirements pertaining to notification of administrative decisions. The Attorney-General's Department is leading work to develop a consistent legislative framework for automated decision-making (ADM), as part of the government's response to recommendation 17.1 of the Robodebt Royal Commission Report.

## 9.2 Challenging administrative actions influenced by AI

Individuals, groups, communities or businesses should be provided with a timely opportunity to challenge an administrative action that has a legal or significant effect on them when the action was materially influenced by an AI system. This is an important administrative law principle. It also promotes accountability and improves the quality and consistency of government decisions.

Administrative actions may be subject to both merits review and judicial review.

### Merits review

Considers whether a decision made was the correct or preferable one in the circumstances, and may include internal review conducted by the agency or external review by the Administrative Review Tribunal.

Where an action can be challenged via internal review (as permitted by relevant legislation), you should consider what processes are in place to allow for internal review of an action materially influenced by AI, for example, by another or more senior officer in the agency.

### Judicial review

Examines whether an action was lawful (for example, whether the decision maker had the power to make a decision or whether a legal error has occurred in making a decision), and is limited to actions which affect an individual's liberties, vested rights or legitimate expectations.

You should ensure review rights that ordinarily apply to human-made decisions or actions are not impacted or limited because an AI system has been used.

Notifications discussed at section 9.1 should include information about available review mechanisms so that people can make informed decisions about disputing administrative actions.

Ensure a person within your agency is able to answer questions in a court or tribunal about an administrative action taken by an AI system if that matter is ultimately challenged. Review mechanisms also impact on the obligation to provide reasons. For example, the *Administrative Decisions (Judicial Review) Act 1977* gives applicants a right to request reasons for administrative decisions.

# 10. Human-centred values

## 10.1 Incorporating diversity

Diversity of perspective promotes inclusivity, mitigates biases, supports critical thinking, mitigates the risk of non-compliance with anti-discrimination laws and should be incorporated in all AI system lifecycle stages.

AI systems require input from stakeholders from a variety of backgrounds, including different ethnicities, genders, ages, abilities and socio-economic statuses. This also includes people with diverse professional backgrounds, such as ethicists, social scientists and domain experts relevant to the AI application. Determining which stakeholders and user groups to consult, which data to use, and the optimal team composition will depend on your AI system.

Failing to adequately incorporate diversity into relevant AI lifecycle stages can have unintended negative consequences, as illustrated in a number of real-world examples:

- AI systems ineffective at predicting recidivism outcomes for defendants of colour and underestimating the health needs of patients from marginalised racial and ethnic backgrounds.
- AI job recruitment systems unfairly affecting employment outcomes.
- Algorithms used to prioritise patients for high-risk care management programs were less likely to refer black patients than white patients with the same level of health.
- An AI system designed to detect cancers had shown biases towards lighter skin tones stemming from an oversight in collecting a more diverse set of skin tone images, potentially delaying life-saving treatments.

Resources, including approaches, templates and methods to ensure sufficient diversity and inclusion of your AI system, are described in the NAIC's Implementing Australia's AI Ethics Principles report.

## 10.2 Human rights obligations

You should consult an appropriate source of advice or otherwise ensure that your AI use case and use of data align with human rights obligations. If you have not done so, explain your reasoning.

It is recommended that you complete this question after you have completed the previous sections of the assessment. This will provide more complete information to enable an assessment of the human rights implications of your AI use case.

In Australia, 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. Australia's federal anti-discrimination laws are contained in the following legislation.

- *Age Discrimination Act 2004*
- *Disability Discrimination Act 1992*
- *Racial Discrimination Act 1975*
- *Sex Discrimination Act 1984.*

Human rights are defined in the *Human Rights (Parliamentary Scrutiny) Act 2011* as the rights and freedoms contained in the 7 core international human rights treaties to which Australia is a party, namely the:

- International Covenant on Civil and Political Rights (ICCPR)
- International Covenant on Economic, Social and Cultural Rights (ICESCR)
- International Convention on the Elimination of All Forms of Racial Discrimination (CERD)
- Convention on the Elimination of All Forms of Discrimination against Women (CEDAW)
- Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)
- Convention on the Rights of the Child (CRC)
- Convention on the Rights of Persons with Disabilities (CRPD).

In addition to other rights referred to in this guidance, human rights you may consider as part of your assessment of the AI use case include:

- a right to privacy – for example, where AI is being used for tracking and surveillance)
- freedom of expression and information – for example, where AI is used to moderate a forum and therefore possibly suppress legitimate forms of expression
- human agency – for example, where AI makes an automated decision on an individual's behalf.

# 11. Accountability

## 11.1 Ensuring accountability during the life cycle of the AI system

Agencies should consider putting mechanisms in place during the life cycle of the AI system to ensure that the agency itself, or the relevant decision-maker, remains responsible and accountable for a government decision which involves the use of AI. Such mechanisms should clearly define how ultimate responsibility for the decision is retained, even when AI is used to analyse data or generate recommended outcomes.

Accountability should be considered at all stages of the AI system lifecycle. Some of the relevant considerations for different stages are outlined below., including:

### During the design and development phase

- How the AI will be constructed in a way that is consistent with the scope of a decision-maker's discretion and any legislative framework which confers authority on a decision-maker
- How the AI will be designed in a way that ensures that the decision-maker takes into account any matters which it is required to consider as part of decision-making
- Whether the decision-maker will have the ability to override or disregard decisions made by AI– for example, where its outputs are based on biased data.

### During the deployment and operation phase

- What information the decision-maker needs to have oversight of the AI (for example, information on the capacities and limitations of the AI, and the process that the AI will use to reach a conclusion)
- What processes are in place to ensure that, where appropriate, final discretion or judgement lies with the decision-maker (for example, the decision-maker analyses the information provided by the AI before making a decision)
- What records will be kept of the decision-maker's reasoning at any decision points which require discretion and judgment. This contributes to the contestability of decisions in accordance with section 9.

### Where the scope of the use case changes or is developed during the life of the AI system

- What assessment, review and acceptance-testing processes are to be applied to the changes to the AI system to ensure the above.

# 12 Use case review and next steps

## 12.1 Alignment with relevant legal frameworks

This question looks to confirm that you have identified and documented any agency specific legislation, regulations, or binding policy instruments that are relevant to your AI use case.

When completing this section:

- review your agency's legislative and regulatory frameworks. Identify any provisions that may be affected by, or place restrictions on, the design, operation, or outputs of the AI system
- if there is any uncertainty, engage your agency's legal area early, and maintain legal professional privilege where appropriate.

## 12.2 Legal advice

This section asks whether your agency has sought or obtained legal advice in relation to the AI use case. If you answer 'yes', you should summarise the nature of the legal issue without including the content of the advice. This information should not be disclosed to anyone other than those who need to know or access the information within the agency.

Note that including the actual content of legal advice in this tool may result in waiver of legal professional privilege, meaning the advice could be legally required to be disclosed to others. To avoid unintended waiver, only summarise the subject matter of the advice–for example, 'privacy compliance' or 'intellectual property risks rather than reproducing or paraphrasing the advice itself.

## 12.3 Risk summary table

To complete the risk summary table:

- list any risks assessed as **medium** or **high** at the inherent risk assessment stage in section 3
- summarise any mitigations or controls that have been or will be applied
- explain how these mitigations have influenced the residual risk rating

## 12.4 Overall residual risk rating

To complete this section, choose an overall residual risk rating for the AI use case. Refer to your response to section 12.3.

## 12.5 Internal governance body review

If your use case's inherent risk is rated as **high** at section 3, you are required under the AI policy to apply specific actions, including creating or reusing a governance body for the purpose of governing high-risk AI. You may document the outcome of the governance body review here, including any recommendations and agreed next steps.

# Appendix: Risk consequence guidance table

This table is designed to help you select the appropriate consequence level for the risk questions in sections 3.1 to 3.8. Examples are illustrative, not exhaustive.

| Risk | Insignificant | Minor | Moderate | Major | Severe |
|---|---|---|---|---|---|
| **3.1 Negatively affecting public accessibility or inclusivity of government services** | Insignificant compromises to accessibility or inclusivity of services.<br><br>Minor technical issues causing brief inconvenience but no actual barriers to access or inclusion.<br><br>Issues rapidly resolved with minimal impact on user experience. | Limited, reversable compromises to accessibility or inclusivity of services.<br><br>Some people experience difficulties accessing services due to technical issues or design oversights.<br><br>Barriers are short-term and addressed once identified, with additional support provided to people affected. | Many compromises are made to the accessibility or inclusivity of services.<br><br>Considerable access challenges for a modest number of users.<br><br>Resolving access issues requires substantial effort and resources.<br><br>Certain groups may be disproportionately impacted.<br><br>Affected users experience frustration and delays in receiving services. | Extensive compromises are made to the accessibility or inclusivity of services, may include some essential services.<br><br>Ongoing delays that require external technical assistance to resolve.<br><br>Widespread inconvenience, frustration, public distress and potential legal implications.<br><br>Vulnerable user groups disproportionately impacted. | Widespread irreversible ongoing compromises are made to the accessibility or inclusivity of services, including some essential services.<br><br>Majority of users, especially vulnerable groups affected.<br><br>Essential services inaccessible for extended periods, causing significant public distress, legal implications, and a loss of trust in government efficiency.<br><br>Comprehensive and immediate actions are urgently needed to rectify the situation. |
| **3.2 Unfair discrimination against individuals, communities or groups** | Negligible instances of discrimination, with virtually no discernible effect on individuals, communities, or groups.<br><br>Issues are proactively identified and rapidly addressed before causing harm. | Limited instances of unfair discrimination occur, affecting a small number of individuals.<br><br>Relatively isolated cases, and corrective measures minimise their impact. | Moderate levels of discrimination leading to noticeable harm to certain individuals, communities, or groups.<br><br>These incidents raise bias and fairness concerns and require targeted interventions. | Significant discrimination results in major, tangible harm to individuals and multiple communities or groups.<br><br>Rebuilding trust requires substantial reforms and remediation efforts. | Pervasive and systemic discrimination causes severe harm across a broad spectrum of the population, particularly marginalised and vulnerable groups.<br><br>Public outrage, potential legal action, and a profound loss of trust in government.<br><br>Immediate, sweeping reforms and accountability measures are required. |

| Risk | Insignificant | Minor | Moderate | Major | Severe |
|---|---|---|---|---|---|
| **3.3 Perpetuating stereotyping or demeaning representations of individuals, communities or groups** | Inadvertently reinforce mild stereotypes, but these instances are quickly identified and rectified with no lasting harm or public concern. | Isolated cases of stereotyping, affecting limited members of community with some noticing and raising concerns.<br><br>Prompt action mitigates the issue, preventing broader impact. | Moderate stereotyping by AI systems leads to noticeable public discomfort and criticism.<br><br>Disproportionally affecting certain communities or groups.<br><br>Requires targeted corrective measures to address and prevent recurrence. | Significant and widespread reinforcement of harmful stereotypes and demeaning representations.<br><br>Causes public outcry and damages the relationship between communities and government entities.<br><br>Urgent, comprehensive strategies are needed to rectify these representations and restore trust. | Pervasive and damaging stereotyping severely harms multiple communities, leading to widespread distress.<br><br>Potential legal consequences, and a profound breach of trust in government use of technology.<br><br>Requires immediate, sweeping actions to address the harm, including system overhauls and public apologies. |
| **3.4 Harm to individuals, communities, groups, organisations or the environment** | Inconsequential glitches with no real harm to the public, business operations or ecosystems.<br><br>Easily managed through routine measures. | Isolated incidents mildly affecting the public.<br><br>Slight inconveniences or disruptions to businesses, leading to manageable financial costs.<br><br>Limited manageable environmental disturbances affecting local ecosystems or resource consumption. | Noticeable negative effects on the public.<br><br>Businesses face operational challenges or financial losses, affecting their competitiveness.<br><br>Obvious environmental degradation, including pollution or habitat disruption, prompting public concern. | Significant public harm causing distress and potentially lasting damage.<br><br>Significant harm to a wide range of businesses, resulting in substantial financial losses, layoffs, and long-term reputational damage.<br><br>Compromises ecosystem wellbeing causing substantial pollution, loss of biodiversity, and resource depletion. | Widespread, profound harm and severe distress affecting broad segments of the public.<br><br>Profound damage across the business sector, leading to bankruptcies, major job losses, and a lasting negative impact on the economy.<br><br>Comprehensive environmental destruction, leading to critical loss of biodiversity, irreversible ecosystem damage, and severe resource scarcity. |

| Risk | Insignificant | Minor | Moderate | Major | Severe |
|---|---|---|---|---|---|
| **3.5** **Raising privacy concerns** | Insignificant data handling errors occur without compromising sensitive information.<br><br>Incidents are quickly rectified, maintaining public trust in data security. | Isolated exposure of limited sensitive data affects a small group of individuals.<br><br>Swift actions taken to secure the data and prevent further incidents. | Breach of moderate amounts of sensitive data, leading to privacy concerns among the affected populace.<br><br>Some individuals experience inconvenience and distress. | Serious misuse of sensitive private data affects a large segment of the population, leading to widespread privacy violations and a loss of public trust.<br><br>Comprehensive measures are urgently required to secure data and address the privacy breaches. | Significant potential to expose sensitive information of a vast number of individuals, causing severe harm, identity-theft risks; use of sensitive personal information in a way that is likely to draw public criticism with limited ability for individuals to choose how their information is used.<br><br>Significant potential to harm trust in government information handling with potential for lasting consequences. |
| **3.6** **Raising security concerns due to the sensitivity or security classification of the data being used by an AI system** | Inconsequential security lapses occur without actual misuse of sensitive data.<br><br>Quickly identified and corrected with no real harm done.<br><br>These types of incidents may serve as prompts for reviewing security protocols. | A limited security breach involves unauthorised access to protected data affecting a small number of records with minimal impact.<br><br>Immediate actions secure the breach, and affected individuals are notified and supported.<br><br>Incident is catalyst for review of security protocols. | Security incident leads to the compromise of a moderate volume of sensitive data, raising concerns over data protection and privacy.<br><br>The breach necessitates a thorough investigation, enhanced security measures. | A significant security breach results in extensive unauthorised access to sensitive or protected data, causing considerable concern and distress among the public.<br><br>Urgent security upgrades and support measures for impacted individuals are implemented. to restore security and trust. | A massive security breach exposes a vast amount of sensitive and protected data, leading to severe implications for national security, public safety, and individual privacy.<br><br>This incident triggers an emergency response, including legal actions, a major overhaul of security systems, and long-term support for those affected. |

| Risk | Insignificant | Minor | Moderate | Major | Severe |
|---|---|---|---|---|---|
| **3.7** **Raising security concerns due to implementation, sourcing or characteristics of the AI system** | Inconsequential security concerns arise due to characteristics of the AI system, such as software bugs, which are promptly identified and fixed with no adverse effects on overall security.<br><br>These issues may serve as lessons, leading to slight improvements in the system's security framework. | Certain characteristics of the AI system lead to vulnerabilities that are exploited in a limited manner, causing minor security breaches.<br><br>Immediate remediation measures are taken, and the system is updated to prevent similar issues. | A moderate security risk is realised when intrinsic features of the AI system allow for unintended access or data leaks.<br><br>Incident affects a noticeable but contained component of the AI system.<br><br>Prompts a comprehensive security review of the AI system and the implementation of more robust safeguards. | Significant security flaws in the AI system's design result in major breaches, compromising a large amount of data and severely affecting system integrity.<br><br>Incident leads to an urgent overhaul of security measures and protocols, alongside efforts to mitigate the damage. | Critical security vulnerabilities inherent to the AI system lead to widespread breaches, exposing vast quantities of sensitive data and jeopardising national security or public safety.<br><br>The incident results in severe consequences, necessitating emergency responses, extensive system redesigns, and long-term efforts to recover from the breach and prevent recurrence. |
| **3.8** **Posing a reputational risk or undermining public confidence in the government** | Isolated reputational issues arise, quickly addressed and explained.<br><br>Causes negligible damage to public trust in government capabilities. | Small-scale AI mishaps lead to brief public concern, slightly denting the government's reputation.<br><br>Prompt clarification and corrective measures minimize long-term impact on public confidence<br><br>Seen by the government as poor management. | Misapplications result in moderate public dissatisfaction and questioning of government oversight.<br><br>Requires remedial actions to mend trust and address concerns.<br><br>Seen by government and opposition as failed management. | Widespread public scepticism and criticism, majorly affecting the government's image.<br><br>Requires substantial efforts to rebuild public confidence through transparency, accountability, and improvement of AI governance.<br><br>High profile negative stories, seen by government and opposition as significant failed management. | Severe misuse or failure of AI systems leads to profound public distrust and criticism.<br><br>Significantly undermining confidence in government effectiveness and integrity.<br><br>Requires comprehensive, long-term strategies for rehabilitation of public trust, including systemic changes and ongoing engagement.<br><br>Seen by government and opposition as catastrophic failure of management.<br><br>Minister expresses loss of confidence or trust in agency. |