Australian Government
Department of Industry, Science and Resources

National Artificial Intelligence Centre

Being clear about AI-generated content

A guide for business

November 2025

Copyright

Disclaimer

The purpose of this publication is to provide general guidance related to the use of Artificial Intelligence.

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

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

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

Acknowledgements

The National AI Centre (NAIC) would like to acknowledge that the Being clear about AI-generated content guidance was developed by the NAIC working with Data 61.

Over 160 organisations participated in consultations in the development of the Guidance for AI Adoption which included consultation on the transparency of AI-generated content.

This guidance was produced with AI assistance. Full review and editorial control remains with the team at the National AI Centre.

Introduction

When you do business, it is good practice to tell people if you have used artificial intelligence (AI) to generate or modify your content. In some contexts, you should also be able to show where your AI-generated content came from, if it has been modified and other details.

This guidance provides Australian businesses with up-to-date best practice approaches to AI-generated content transparency based on the latest research and international governance trends.

It explains why and how to tell people that you've used content that has been generated or modified by AI in your business.

Following this guidance will help you to:

- ensure that your AI-generated content is clearly identifiable
- follow industry best practice
- contribute to emerging standards of transparency for AI-generated content.

This guidance is voluntary and supports the *Guidance for AI Adoption: Implementation practices*. It builds on practices outlined in Practice 4. Share essential information.

This guidance on how to inform people about AI-generated content **does not cover** how to:

- tell users when they are engaging with AI systems
- tell users when they may be impacted by AI-enabled decisions
- be transparent about non-AI-generated content
- manage copyright or other intellectual property implications of AI-generated content
- use AI-generated content detection mechanisms.

Contents

Who is this for?

This guidance is for:

- all businesses that use AI to generate or modify content
- all businesses that build, design, train, adapt or combine AI models or systems that can generate or modify content.

Everyone involved in the AI lifecycle is responsible for being transparent about their AI-generated or modified content.

Why be transparent about AI-generated content?

AI is a rapidly developing technology, and it's changing the way people do business. AI-generated content is already common in business and marketing contexts, and its realism and reach has increased as the technology has advanced.

Because of this, it can now be difficult to tell if content has been modified or generated by AI. This can make it more difficult for people to trust the content they encounter. It can also make it easier for people to commit fraud and other malicious acts.

Being transparent about when your content is AI-generated can help to:

Build trust in the digital content ecosystem

Being transparent about AI-generated content can contribute to greater accountability, reliability and trust in the digital content we engage with.

Reduce regulatory and reputational risks

The regulatory landscape around AI is evolving, both in Australia and internationally. Being transparent about AI-generated content may be necessary for your business to meet its regulatory requirements and adapt to new ones. It can also help to reduce the risk of harms of misleading or deceptive AI-generated content.

Improve collective digital literacy

Clearly identifying AI-generated content is an important way to help people recognise and understand that information that they encounter. This can support public education efforts and build skills in critically evaluating the authenticity and reliability of information.

Build competitive advantage

People are more likely to engage with AI-generated content when they understand where it comes from and how reliable it is. Being transparent about your use of AI-generated content may help to create a point of difference with your competitors. It can also support your business to build a foundation of trust with your consumers.

Transparency mechanisms

We use the term ‘transparency mechanisms’ to mean the ways you can show your users when and how you’ve used AI to create or modify content. Transparency mechanisms can also show where AI-generated content, including images, text, audio and video, come from.

In this guidance, we talk about 3 transparency mechanisms:

- labelling
- watermarking
- metadata recording.

You can apply these mechanisms individually or combine them. When combined, they provide a greater level of transparency.

Your context and regulatory obligations (such as responsibilities under the Online Safety Act) will inform whether you need to use one or more transparency mechanisms.

Content transparency mechanisms can help to build awareness and trust and improve accountability and safety. They can help users to distinguish AI-generated content from human-authored material. This can help users think critically about the accuracy of content they consume, which can reduce the risks of seriously harmful disinformation and misinformation.

Digital content transparency mechanisms are evolving. There isn’t yet a standardised approach to transparency for AI-generated content, but AI Safety Institutes are progressing research on technical approaches. Industry-led initiatives such as the Coalition for Content Provenance and Authenticity (C2PA) are being adopted.

Transparency mechanisms are not failsafe. They can be misused or tampered with and remain vulnerable to attack. As a business, you need to judge how to implement transparency mechanisms which best suit your context.

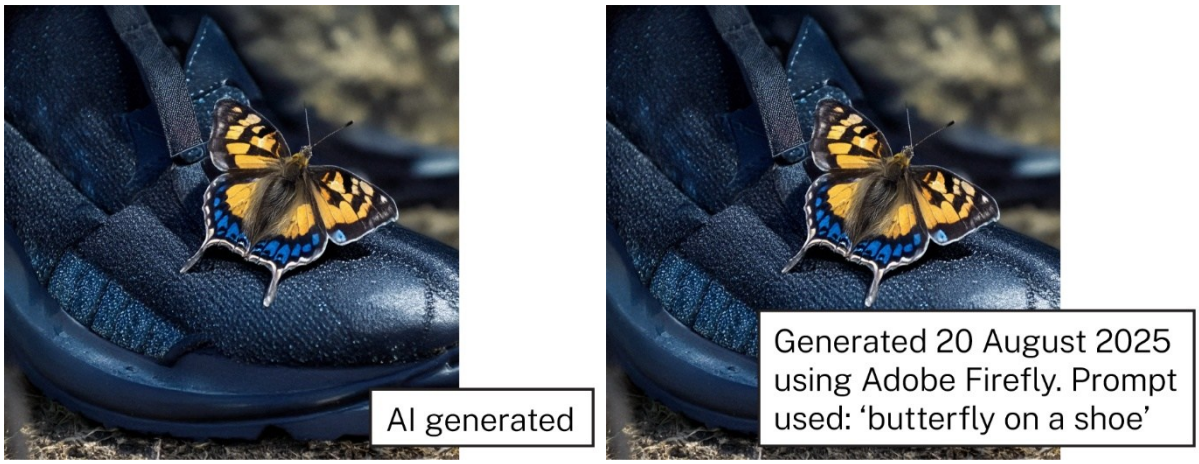
Learn more about the limitations of transparency mechanisms.

Labelling

Labelling means using visible text to tell users if something is AI-generated and where it came from.

Labelling is the easiest transparency mechanism to use, but it can need extra support from watermarking and accurate metadata for it to be credible.

Labelling can range from very simple to complex.



Simple AI generated label (left) and complex AI generated label (right).

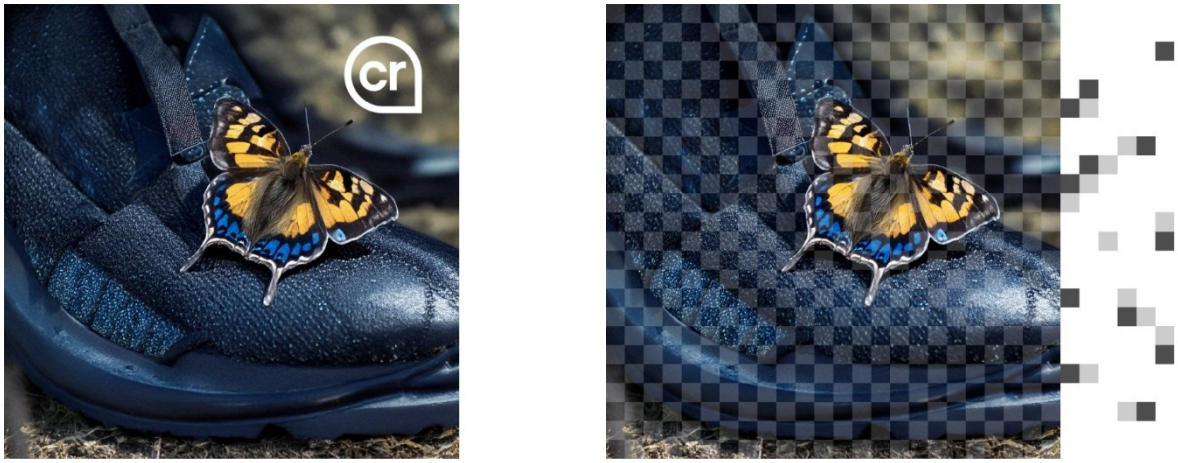
Watermarking

Watermarking is a way to embed information into digital content so you can trace its origin or verify its authenticity. This information can be visible or invisible to people. Watermarking is different to labelling, which is easier to remove.

Visible watermarking can appear in several ways. In images and videos, a watermark may appear as a semi-transparent text overlay. In audio, it could take the form of an audible disclosure stating, 'This audio was generated by AI'.

Invisible watermarking involves embedding hidden data into the content. To verify the watermark, a user would need to extract the watermark data by using a special watermark detection tool to examine the content.

Typically, AI model developers are responsible for building watermarks into their model. System developers and deployers are responsible for implementing, checking and using these watermarks, depending on their needs and use case. Find out more in the National institute of Standards and Technology's overview of technical approaches to digital content transparency (PDF).



Visible watermarking (left) and invisible watermarking (right).

Metadata recording

Metadata is descriptive information about a piece of content that's included with the content file. Often metadata appears automatically. An example of this is digital photography. Depending on the device you use, the metadata of a digital photo will record where and when it was taken.

Metadata recording is versatile. It can include many details about a piece of content, like who created it and whether it's been edited. Accurate metadata can support the credibility of both watermarking and labelling.

Metadata capabilities are usually the responsibility of AI system developers or model developers. If you're an AI system deployer, you should check if the system you're using has metadata recording capabilities.

Metadata recording can contain more than just the AI generation information.

Legal responsibilities when generating content with AI

Businesses that generate or modify content using AI must comply with existing legal and regulatory obligations, especially relating to privacy, competition, human rights, copyright and online safety. Under the Australian Consumer Law all businesses are required to ensure your conduct and the representations you make are truthful and not deceptive, including where AI was involved in creating content. Domestic and international legal obligations, technical capabilities and transparency mechanisms are evolving rapidly, and businesses should monitor them closely.

When developing and using AI systems and associated transparency mechanisms, all developers and deployers should be aware of their obligations under the *Privacy Act 1988*. This includes obligations relating to the collection, use and disclosure of personal and sensitive information when training or fine-tuning AI models, generating content or recording metadata in addition to the security and accuracy of personal information.

Metadata including personal information such as author name, creation date or IP address, should generally be obscured, though developers and deployers should be aware of their obligations under the *Copyright Act 1968* in relation to electronic rights management information. You should also respect individual and collective rights as well as cultural sensitivities. Read more about privacy and generative AI on the OAIC website or work underway on Envisioning Aboriginal and Torres Strait Islander AI Futures.

The spectrum of AI-generated content

We use the term 'AI-generated content' to mean content that has been created or modified, in whole or in part, by AI. We consider AI-generated content on a spectrum:

AI-*assisted* content

You use AI to assist you in minor ways. For example:

- spelling and grammar checks
- automatic photo touch-ups like removing red eyes.

AI-*enhanced* content

You use AI to modify or refine content through inputs or instructions you enter into the AI system. For example:

- fact-checking or editing a complex document with substantial rewrites
- removing specific details from the background of images, such as logos.

Fully AI-generated content

Through a simple action, like a prompt, or uploading a file, you create a complete output, with little to no human oversight. For example:

- creating an interview-style video from a still image and a text script
- creating a new poster artwork from a verbal prompt.

Figure 1: Spectrum of content creation

How businesses can improve transparency of AI-generated content

In this guidance, we talk about businesses as AI model developers, AI system developers and AI system deployers. Businesses may fall into more than one category. All 3 have a role to play in improving the transparency of AI-generated content.

Table 1: Roles and responsibilities across the AI lifecycle

AI system deployers	AI system developers	AI model developers
Apply user-friendly labels • fit-for-purpose information • consider accessibility **Educate and verify** • verify mechanisms • explain their use • offer support contacts	**Develop and apply system-level tools** • adapt to context and users • offer configurable options • verify persistence of watermarks	**Develop model-level tools** • use common standards • provide robust tools to deployers • ensure resilience against attacks **Establish feedback with deployers** • understand risks • foster continuous improvement

AI system deployers

You are an AI system deployer if you are a business or person that uses an AI system to operate or to provide a product or service. This might look like:

- using chat assistants such as ChatGPT, Microsoft Copilot, Claude or Gemini to create content for use internally or externally
- using AI-enabled image editors such as Adobe, Canva or Picsart
- giving your customers access to an AI chat assistant to answer frequently asked questions from your website
- using AI to monitor system performance
- AI-enabled handling of user feedback
- using AI to maintain systems.

Examples of AI system deployers include Australian businesses who are using AI to improve their operations. For example, call centres who deploy AI to improve their customer support experience and minimise follow-up calls.

Read more examples of AI system deployment in When to use transparency mechanisms.

You should:

Design and apply fit-for-purpose transparency mechanisms

- Apply visible, user-friendly labelling that suits the context and people consuming your AI-generated content
- Create simple labels such as ‘generated by AI’, or ‘created with AI assistance’ that are easy to identify and understand
- Use icons, colour codes or badges to identify AI-generated content without compromising content quality
- Consider the accessibility of transparency mechanisms for diverse audiences, including people living with disability.

Educate consumers of content about transparency mechanisms

- Tell people who consume your content about AI-generated content risks and the importance of transparency mechanisms
- Help people understand how to interpret and use transparency mechanisms – consider FAQs, videos and explainers.

Provide secure and safe access and storage

- Develop and use tools to safely record and access metadata, ensuring privacy, safety, security and compliance with the model developer’s standards
- Keep metadata in secure, access-controlled storage so you can perform self-audits or comply with auditing requirements.

Create clear verification processes

- Use tools provided by developers to validate watermarks and metadata and include feedback loops if applicable
- Tell people when you can’t verify transparency mechanisms
- Give people clear next steps when they encounter verification issues including support contacts.

Facilitate human oversight and feedback loops

- Facilitate human oversight of AI-generated content
- Participate in feedback loops with AI system and model developers, as well as with experts in relevant forms of harms (e.g. privacy).

Consider transparency mechanisms in procurement processes

- Assess the watermarking, labelling and metadata recording capabilities of AI system suppliers as part of the procurement process.

AI system developers

You are an AI system developer if you are a business or person who designs, builds and tests a system that uses AI. This might look like:

- integrating AI models into applications
- creating a platform that uses existing AI in new ways, like a chatbot or AI-augmented visual design app
- creating user interfaces for AI models
- customising AI models for specific uses.

Examples of AI system developers include software companies and app developers incorporating AI models in their product. These products might include AI image and video editors, or tools that create music samples using AI.

If a company takes an off-the-shelf AI system and tailors it for specific use or context they are considered a system developer. For example, if IT development company Fortunesoft tailors ChatGPT for the financial services sector, we consider it to be an AI system developer.

Other examples of AI system developers include Open AI (ChatGPT), Amazon (Amazon Rekognition) or Microsoft (Copilot). Businesses buy these AI systems off-the-shelf and often deploy them directly.

You should:

Design and apply fit-for-purpose transparency mechanisms for deployers to use

- Design intuitive labels that suit context and people consuming the AI-generated content
- Prioritise user experience for users with varying levels of technical knowledge
- Offer configurable options for transparency mechanisms to be applied by deployers

Provide secure and safe access and storage

- Develop and use tools to safely record and access metadata, ensuring privacy, safety, security and compliance with the model developer's standards
- Keep metadata in secure, access-controlled storage so you can perform self-audits or comply with auditing requirements.

Develop or adopt watermarking techniques

- Develop or adopt post-generation (after content is generated) watermarking techniques at the system or application level
- When available use tools from your model developer to embed model-level watermarks into AI-generated content

- Create ways to verify watermark persistence during your system's operations.

Facilitate human oversight and feedback loops

- Facilitate human oversight of AI-generated content
- Participate in feedback loops with AI system deployers and AI model developers, as well as with experts in relevant forms of harms (e.g. privacy).

AI model developers

You are an AI model developer if you are a business or person who creates, tests, trains and validates AI models. This might look like:

- designing and building AI models
- training AI models on specific datasets
- testing and checking model outputs
- researching new ways to improve model abilities such as by changing model parameters and fine-tuning.

Examples of AI model developers include Open AI, Anthropic and Google DeepMind.

You should:

Develop effective watermarking and metadata techniques

- Develop model-level watermarking techniques by embedding traceable identifiers during model training or integrating them into the content generation process
- Ensure watermark resilience against common transformations and attacks while maintaining utility as technology matures
- Contribute to or use existing standards where applicable to define secure metadata formats, including model version, parameter counts, and training data sources
- Provide tools, software development kits, application program interfaces, or frameworks to system developers and other stakeholders for implementing and verifying watermarks and metadata.

Collaborate with AI system developers and deployers

- Ensure watermarking techniques and metadata standards are effective and fit-for-purpose
- Test and validate that transparency features survive content workflows and user interactions
- Design mechanisms that flag when content has been modified or tampered with
- Create solutions for content tracking throughout the distribution channels

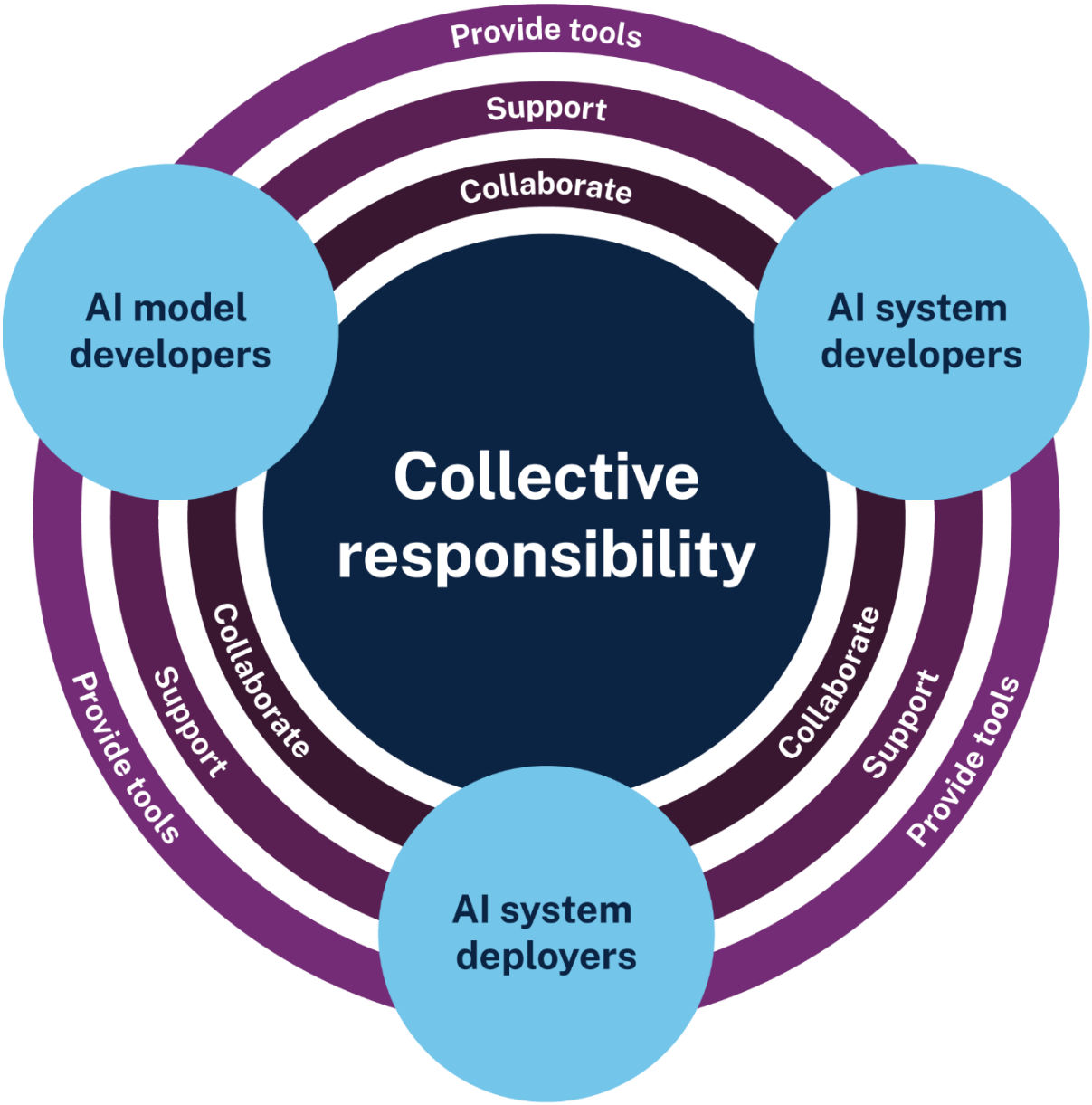
- Participate in feedback loops with AI system deployers and AI system developers, using deployment insights to continuously improve transparency mechanisms.

Working together to improve transparency across the AI lifecycle

Wherever possible, developers and deployers should work together to ensure they build in transparency mechanisms from the start and that these work as intended. Developers should commit to improving transparency mechanisms and sharing best practices, processes and technologies, including publishing this information to improve transparency.

Developers and deployers should also ensure that these mechanisms are accessible to people with disabilities. This audience has diverse ways of communicating and may use a range of technologies to access information.

Figure 2: Opportunities for organisations and individuals to improve transparency of AI-generated content across the AI lifecycle through collaboration



When to use transparency mechanisms

You may not need to use transparency mechanisms every time you use AI generated content. Which transparency mechanisms you use will depend on your context and how much risk your content poses.

You need to decide the best ways to be transparent about how your business uses AI to generate content.

The transparency mechanisms you use should be proportionate to your content's:

- **potential negative impact**, or how the AI-generated content might adversely affect people
- **AI involvement level**, or how involved AI has been in creating the content.

This will help you choose how and when to use transparency mechanisms such as watermarking, labelling and metadata. You can apply mechanisms individually or combine them. When combined, they provide a greater level of transparency.

You should consider how using AI-generated content and being transparent about it will affect employees, customers and the broader public.

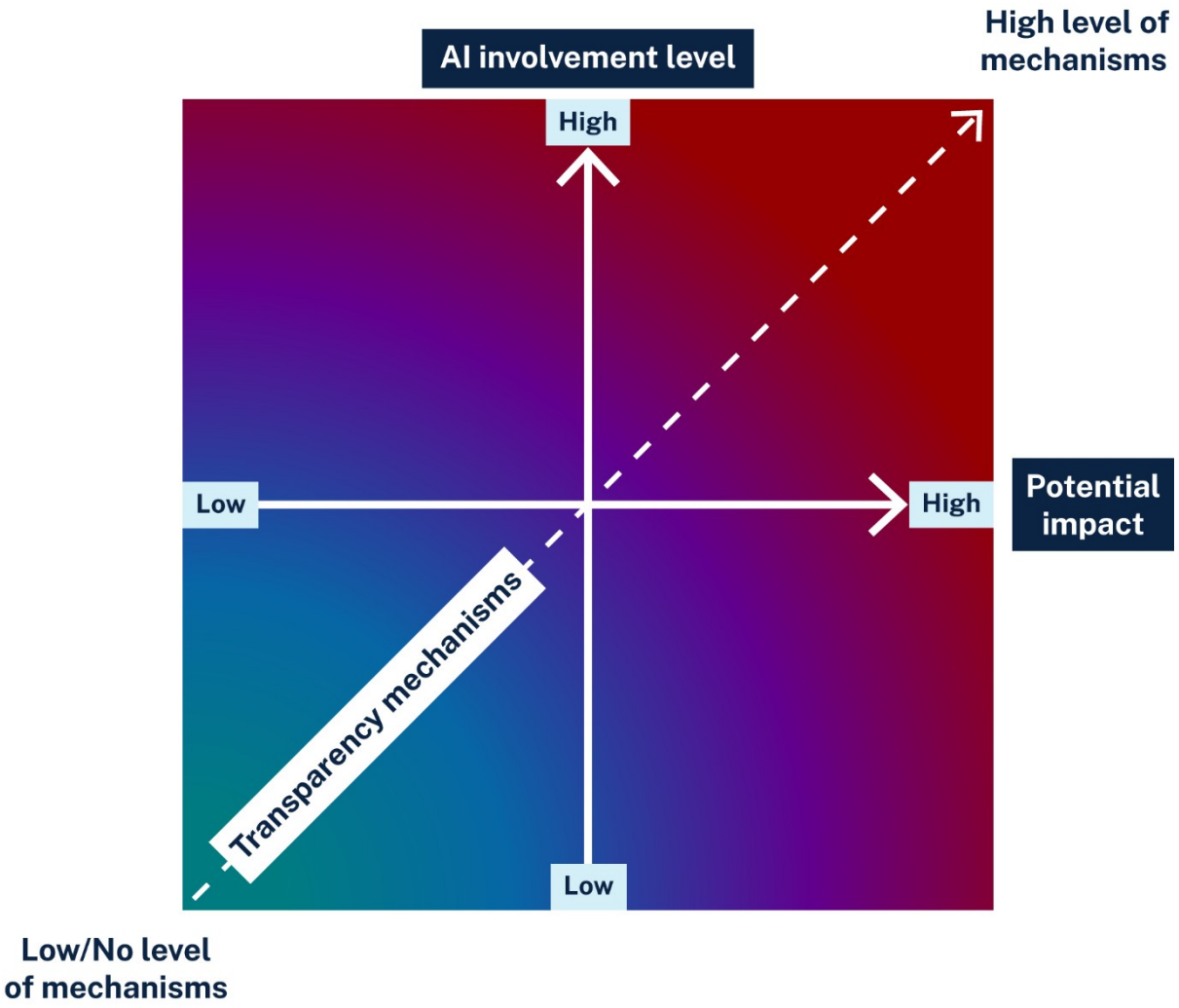
Transparency becomes more important if your AI-generated content has the potential to create more negative impacts, with less human oversight or involvement.

This section will be most useful for businesses who understand the context in which their AI-generated content is going to appear. This usually means AI system deployers and AI system developers.

However, this section can also be useful to AI model developers, as they should be aware of the potential impacts of their models. Deployers and developers should communicate about these impacts and what transparency mechanisms may help to mitigate them.

This section also shows examples of when you might use transparency mechanisms and which types you might use.

Figure 3: Framework for assessing risks of AI-generated content



Assess your AI-generated content

Step 1: assess potential negative impact

The potential impact of AI-generated content depends on:

- its context
- how you use it
- the overall nature of your product or service more broadly.

If the potential negative impact of your AI-generated content is high, you need to use robust transparency mechanisms to make sure you can accurately communicate how you've used AI.

It is important to consider the potential for the content to be seen as authoritative.

For example, using AI-generated content in a clinical setting could lead to misdiagnosis. In recruitment processes it could lead to a breach of employment law. In these circumstances, you'd need to use strong transparency mechanisms for AI-generated content.

You should consider how AI might negatively impact different aspects of people's lives including (though this is not exhaustive):

- people's rights
- people's safety
- collective cultural and societal interests, particularly First Nations peoples
- Australia's economy
- the environment
- the rule of law
- the context in which it is shared
- impacts to the business (such as commercial, reputational impacts).

Your business may already have risk assessment tools or frameworks that you can use to assess the likely impact of your AI-generated content. You can also refer to the principles outlined in the government's proposals paper for introducing mandatory guardrails for AI in high-risk settings, which may help you assess your risk levels. You may also have regulatory obligations, for example under the *Online Safety Act 2021, Australian Consumer Law a*nd the *Privacy Act 1988*.

Step 2: assess AI involvement level

How much AI is involved in making content affects the level of risk. When thinking about AI involvement in creating content, ask:

2 a) How automated is the AI system that's generating or modifying content?

- Systems with lower levels of responsible human oversight need stronger transparency mechanisms.
- Systems where AI serves as an assistive tool with substantial human oversight may not need such strict approaches.
- Fully automated systems, where human input is minimal or absent, may require more extensive transparency mechanisms. Those with human review and oversight may require less.

2 b) Has AI substantially modified or generated the content?

- If an AI system has substantially modified content, the content needs more transparency mechanisms.
- What counts as a substantial modification depends on context and the needs and expectations of your users.

2 c) Is there potential for AI to substantially change the meaning of the content?

- Even minor changes to content can significantly change its meaning, potentially leading to users misinterpreting the content. For example, an AI editor omitting the word ‘not’ could substantially change the meaning of the content it’s editing.
- Low AI involvement can still lead to big changes in a piece of content’s meaning.
- Assess if AI has altered the meaning of your content in ways which may affect your stakeholders’ interpretation.

Step 3: choose the right transparency mechanisms

When AI-generated content has high potential for adverse impacts and high levels of AI involvement in the generation process, it needs more extensive transparency mechanisms. These could include a combination of labelling, watermarking and metadata recording. Types of content that may need this level of transparency include:

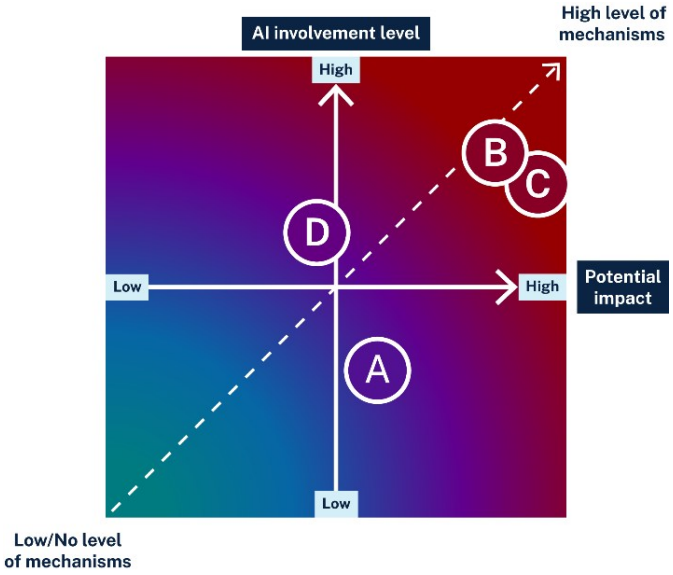
- AI-generated medical reports used in treatment decisions
- when AI changes the spoken language of a high-profile speaker in an online video, often known as a ‘deepfake’.

When AI-generated content has low potential impact and low AI involvement it may need only minimal transparency mechanisms, or none at all. These scenarios could include:

- grammar corrections in casual emails
- automated brightness adjustments to personal photos.

When AI-generated content has high potential impact and low involvement, or low potential impact and high involvement, you’ll need to tailor the transparency mechanisms depending on context.

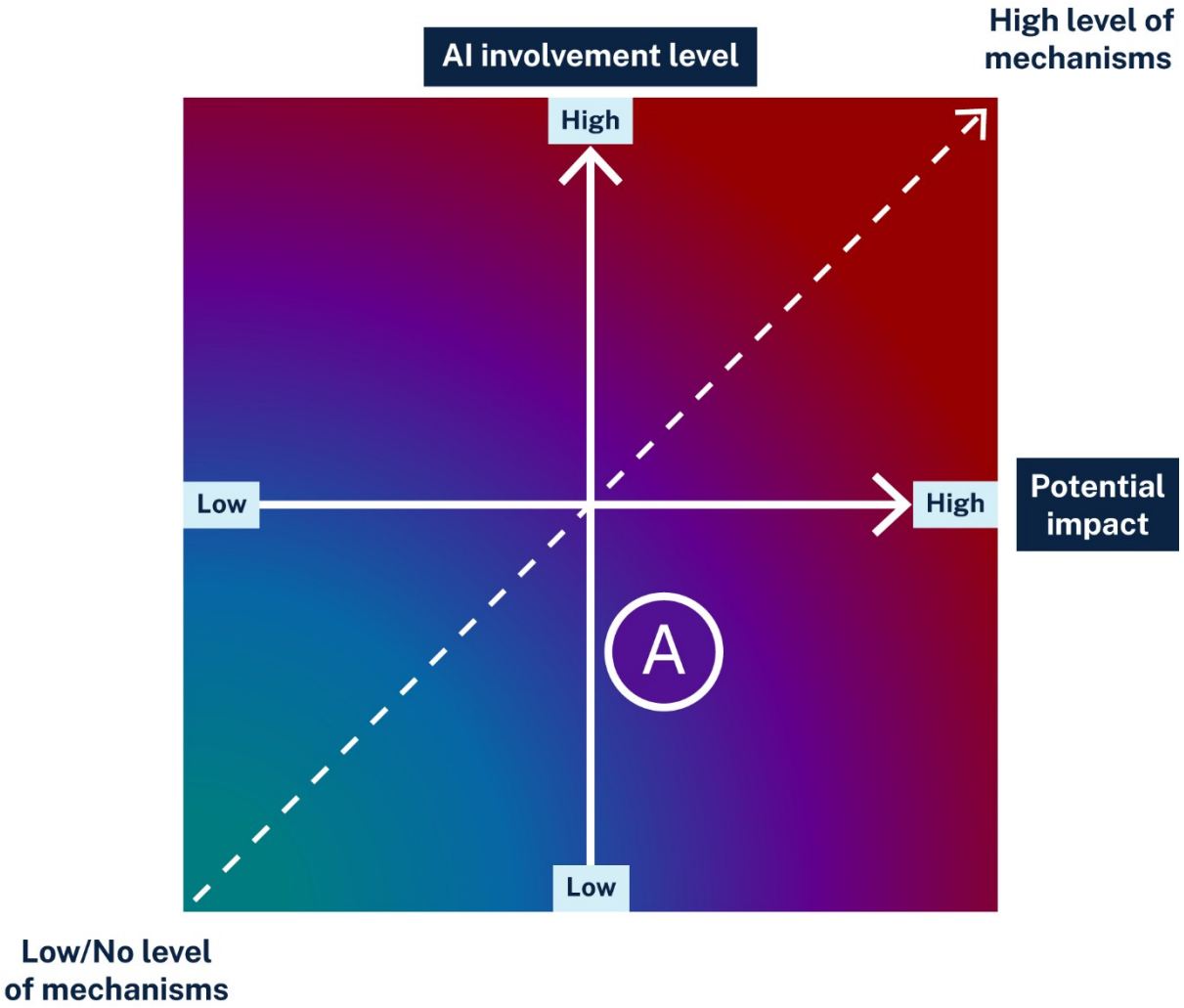
Read the examples to learn more. These are illustrative only.



Examples

Example A: AI-enhanced editorial assistance

You are a journalist using an AI-enabled word processer to write a news article. Your word processor does more than spelling and grammar checks. It also suggests improvements such as adjusting tone for consistency, performs basic fact checking, recommends examples and you include sections of AI-generated text. You review and have editorial control over the final content.



Overall content risk level: Moderate

Potential negative impact: Moderate

News articles have the potential to reach broad audiences and influence public opinion. Readers may expect to be informed about AI's role in the writing process. Publications may also need to take reasonable steps to adhere to relevant industry Standards of Practice.

Extent of AI involvement: Moderate

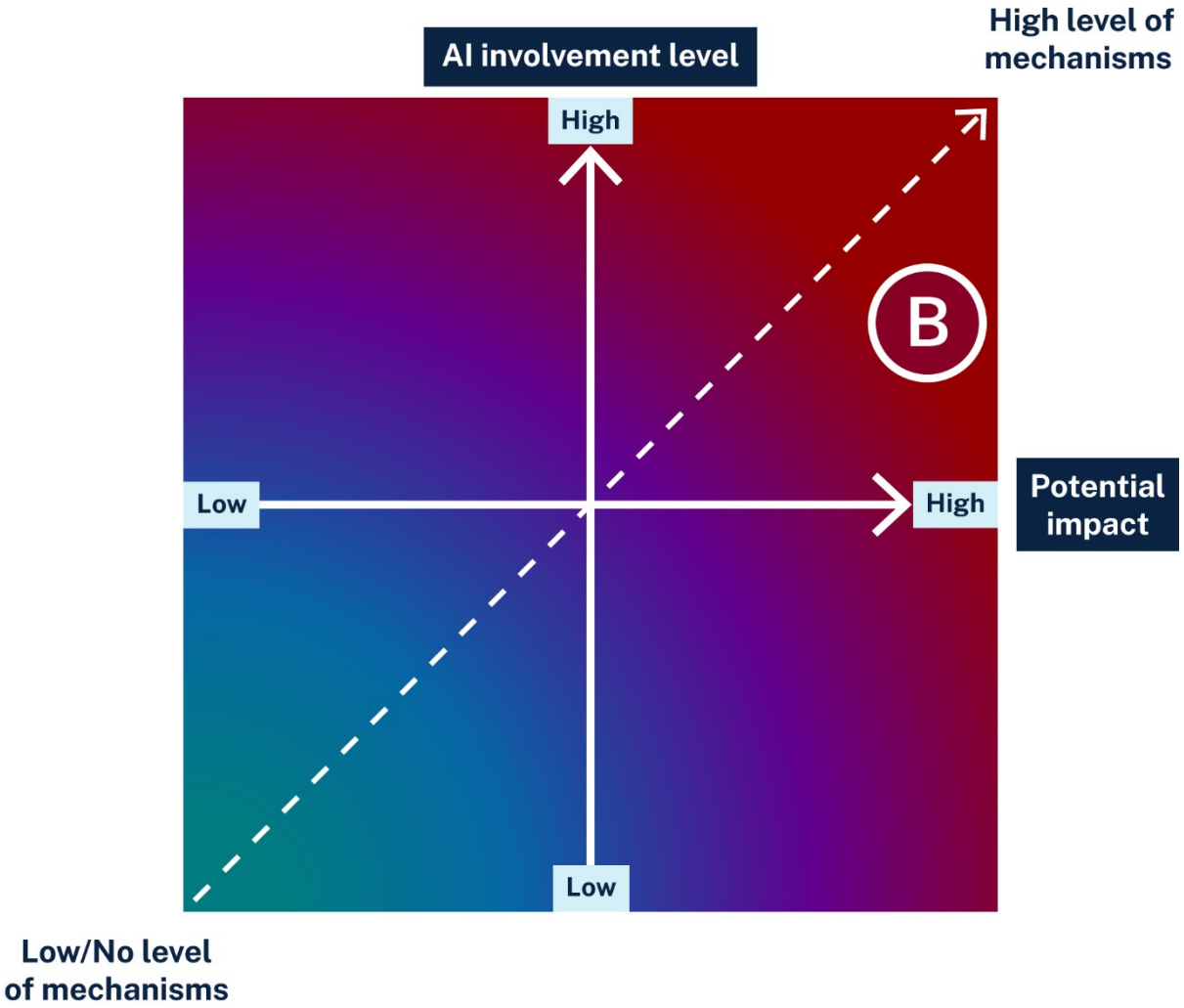
- **Extent of automation:** Low to moderate – humans still have oversight and full editorial control of the content.
- **Content modification:** Moderate – the AI system is significantly generating new content.
- **Meaning alteration:** Moderate – content modification has the potential to significantly change the meaning.

Recommended actions

Depending on the nature of the article and your audience, you may add a label 'Article enhanced by AI'. This ensures readers are aware of AI's supportive role. You may not need to use other transparency mechanisms, as the human author has primary responsibility.

Example B: AI-enhanced images for clinical diagnosis

You are developing an AI system to enhance medical images for diagnostic insights. This could include synthesising specialised imaging views or highlighting potential abnormalities in X-rays, MRIs or other scans. These images directly inform clinical decisions and patient care pathways, so accuracy and reliability are paramount. Without thorough human oversight, incorrect or misleading images could lead to misdiagnoses or improper treatments, or have life-threatening consequences.



Overall content risk level: High

Potential negative impact: High

This is a clinical setting with scope to adversely impact the health and wellbeing of cancer patients.

Extent of AI involvement: Moderate to high

- **Extent of automation:** Moderate – mixed approach with some human oversight with fully automated AI image modification.
- **Content modification:** Moderate to high – AI-created or enhanced images are significantly changing the original content.

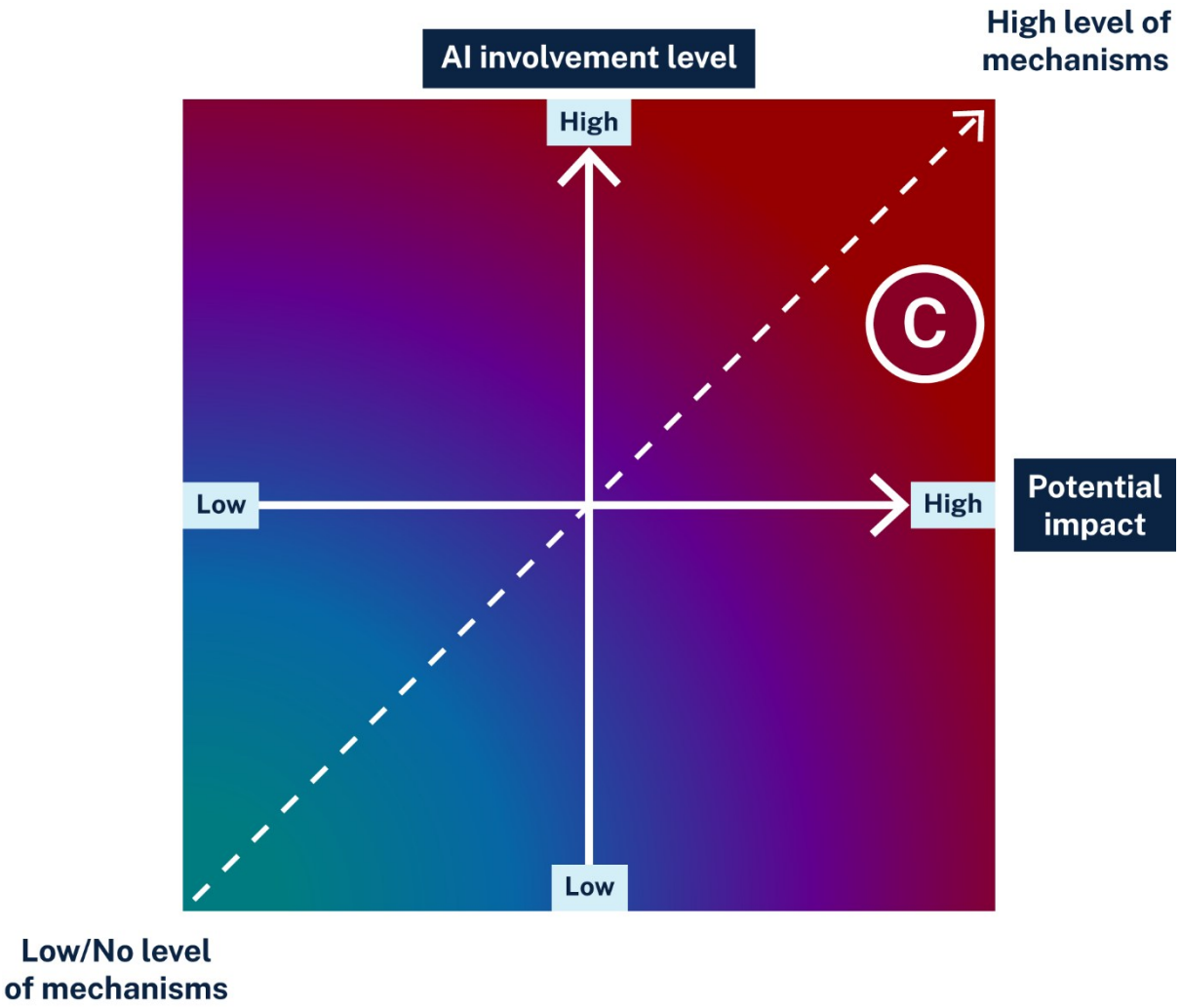
- **Meaning alteration:** High – small changes in an AI-enhanced image could mean the difference between a patient receiving a cancer diagnosis or a negative result.

Recommended actions

Co-design labelling user interfaces with system users and clinicians. Use system-level labelling of content as 'Image enhanced by AI' to disclose AI-modified content to clinicians. Develop and maintain secure, accessible and complete metadata logs (for example, version of AI model used, date/time of generation, confidence scores). All watermarking-related actions mentioned in the AI system developers section should be in place here. These steps should be on top of standard transparency measures and other requirements appropriate within medical settings.

Example C: AI-generated draft of legal contract

You are a lawyer using an AI system to produce fully AI-generated drafts of legal contracts.



Overall content risk level: Moderate to high

Potential negative impact: High

There is the potential for material legal, financial and reputational harm to individuals or organisations relying on AI-generated contracts. There may also be risks to privacy if personal information is an input to the AI system.

Extent of AI involvement: Moderate to high

- **Extent of automation:** Moderate – mixed approach with some human oversight by legal professionals, and some parts of the content generation process are automated.
- **Content modification:** Moderate to high – substantial content modification as AI is generating the contract.

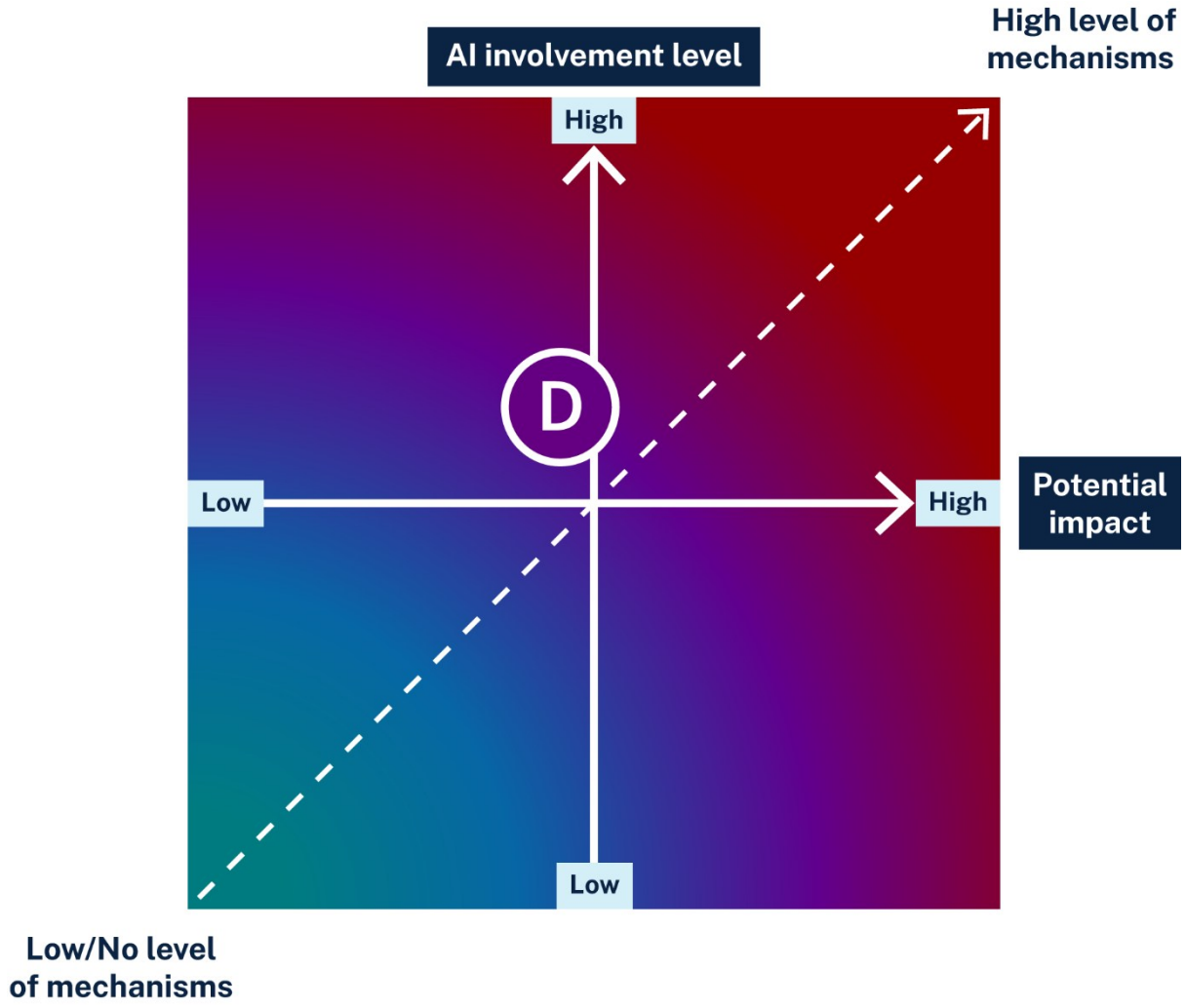
- **Meaning alteration:** High – small changes to contract text can have a significant impact on interpretation and meaning.

Recommended actions

Label contracts with ‘Initial draft generated by AI’ within the law firm. Make sure the lawyer retains human oversight and responsibility for accuracy. All metadata-related actions for system developers listed above should apply here, including maintaining metadata logs (for example, date/time of generation).

Example D: AI-generated marketing content for household products

You run an advertising agency that is putting together a marketing campaign asset for a client advertising a new line of household products. Your creative team decides to use AI-enabled applications to modify and curate images, short videos and messaging text. When content is marketed to consumers, Australian Consumer Law obligations apply, including that representations must be truthful and accurate.



Overall content risk level: Moderate

Potential negative impact: Moderate

In marketing contexts, businesses need to ensure they remain compliant with Australian Consumer Law in all forms of advertisements, promotions, websites and statements.

Extent of AI involvement: Moderate to high

- **Extent of automation:** Moderate – creative teams are using the AI-assisted applications to modify and create content for the campaign. The teams maintain human oversight.

- **Content modification:** Moderate – AI is significantly modifying content from the original.
- **Meaning alteration:** Moderate – the meaning of content is likely to change from the original where AI introduces new visual elements or modifies contexts.

Recommended actions

The advertising agency should label content produced for the client as 'Enhanced by AI.' The advertising agency should ensure human oversight from design teams and copyrighters. The client may choose to label advertisements 'Enhanced by AI'.
The client is also responsible for ensuring that the overall general impression made by representations of the product or its characteristics is accurate to comply with Australian Consumer Law.

Figure 4: Examples of transparency mechanisms for AI-generated image content

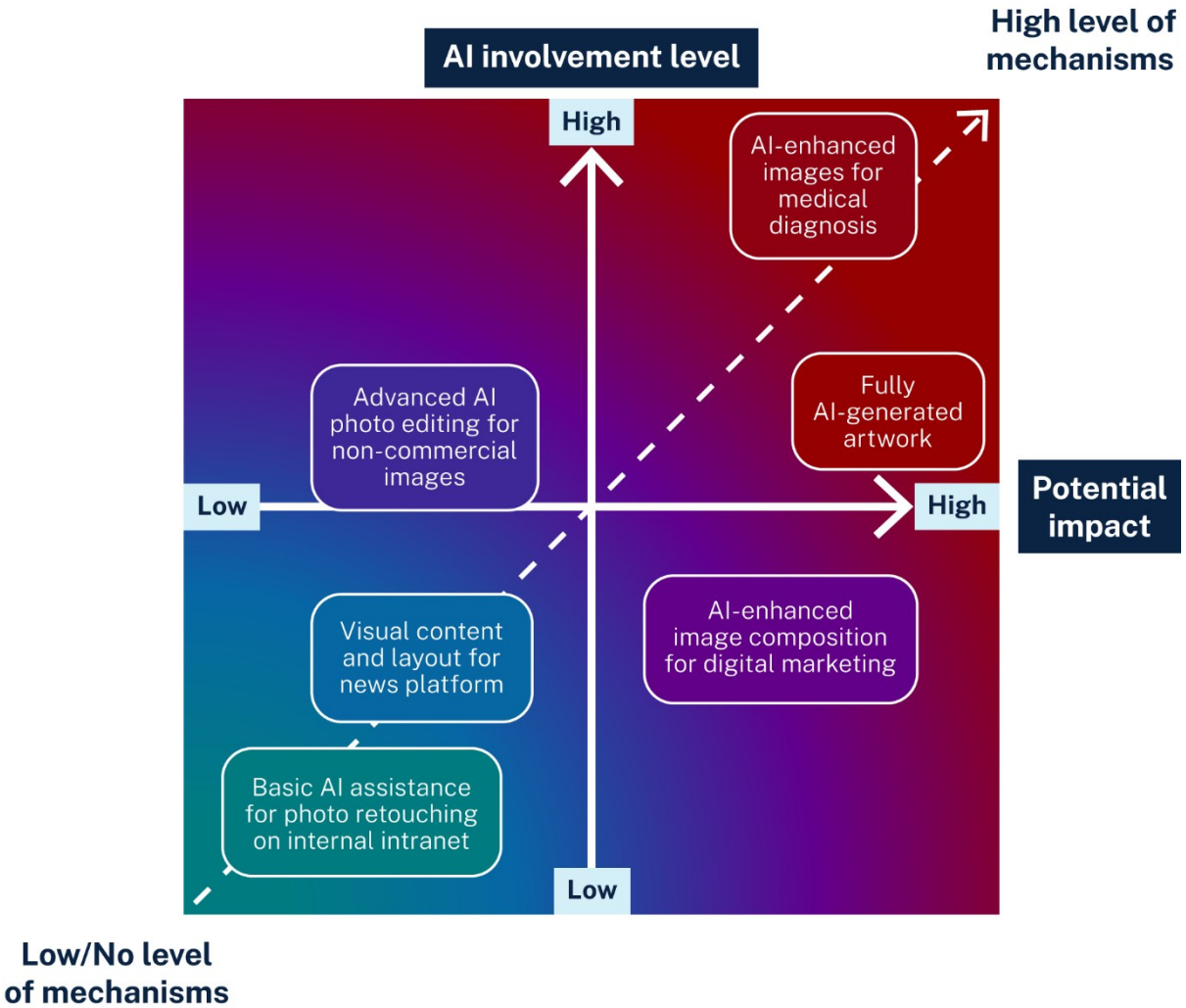


Table 2: Transparency mechanism options for AI-generated image content

Scenario	Labelling	Metadata	Watermarking
Basic AI assistance for photo retouching on internal intranet	Not needed	Not needed	Not needed
Advanced AI photo editing for non.commercial images	May need	May need	Not needed
AI-enhanced image composition for digital marketing	May need	May need	Not needed
Visual content and layout for news platforms curated using AI	Likely need	Likely need	May have
Fully AI-generated artwork	Likely need	Likely need	Likely need
AI-enhanced images for medical diagnosis	Likely need	Likely need	Likely need

Why did we write this?

Background

The Australian Government is taking an integrated approach to mitigating risks in the development and deployment of AI while supporting innovations in the sector.
In January 2024, in its interim response to the Safe and Responsible AI consultation, the government committed to:

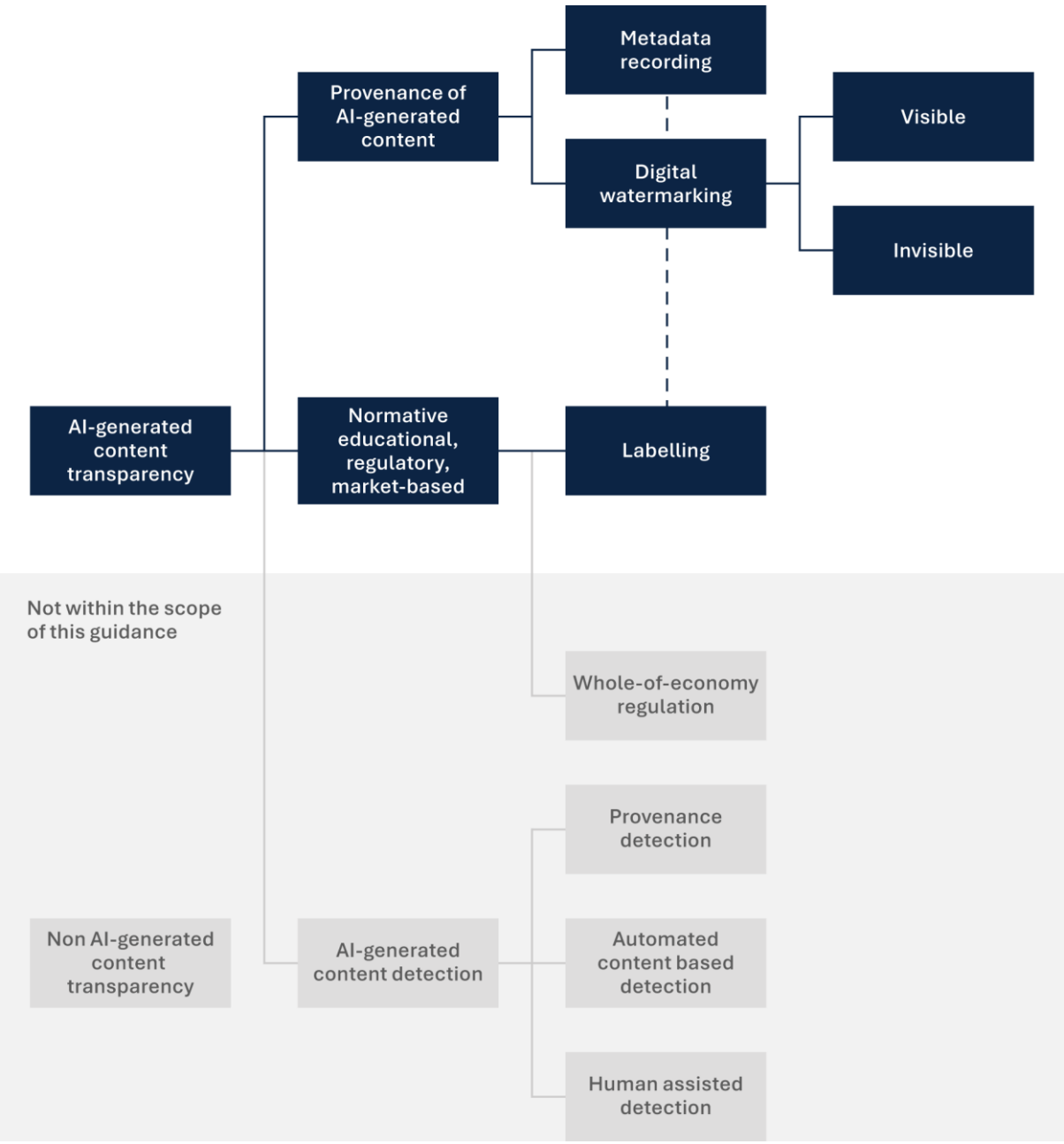
- consulting on the establishment of mandatory guardrails for high-risk AI
- working with industry to develop a Voluntary AI Safety Standard
- working with industry to develop options for voluntary labelling and watermarking of AI-generated content
- establishing an expert advisory group to support the development of options for mandatory guardrails.

This document represents a snapshot in time of best-practice guidance of transparency mechanisms for AI-generated content. This is an area that sits at the cutting edge of research. Because of this, what is ‘best practice’ is evolving. We intend to revise this guidance with updates reflecting any major changes in the current state of the art.

This guidance does not cover transparency mechanisms for:

- non-AI-generated content
- whole-of-economy regulatory frameworks for digital content transparency
- AI-generated content detection mechanisms (see Figure 5).

Figure 5: Spectrum of digital content transparency measures



Watermarking and labelling – or using transparency mechanisms, as we've referred to it in this guidance – is an emerging field of AI governance. Relevant global industry-led approaches such as open-source internet protocol (C2PA) and information security controls (ISO 27002) continue to emerge. In November 2024, in the United States, the National Institute of Standards and Technology released its first synthetic content guidance report. This looked at the existing standards, tools, methods, practices and potential for development of further ways to deliver digital content transparency.
In April 2024, the EU introduced mandatory regulatory requirements for providers of general-purpose AI systems. These require providers to ensure their output is 'marked in a machine-readable format and detectable as artificially generated or manipulated'.

Widespread adoption of digital content transparency measures will be important to achieve broader economic and societal benefits as well as improve digital literacy.

A coordinated approach can incentivise larger model developers to incorporate effective tools. It can also encourage sharing and collaboration on best practices, process and technologies.

The International Network of AI Safety Institutes has identified managing the risks from synthetic content as a critical research priority that needs urgent international cooperation. Australia is co-leading the development of a research agenda focused on understanding and mitigating the risks from synthetic content, including watermarking and labelling efforts. The network aims to incentivise research and funding from its members and the wider research community, and encourage technical alignment on AI safety science.

Fostering greater transparency of AI-generated content requires global collaboration across sectors and jurisdictions to create integrated and trusted systems that promote digital content transparency.

Limitations of transparency mechanisms

Approaches to digital content transparency continue to develop in industry and academia. While the benefits of being transparent about AI-generated content are clear, some limitations remain:

- A lack of standardisation in watermarking technologies means that one watermarking system is not interoperable with another system.
- A lack of standardisation in approach across the economy can be a barrier to behaviour change. For example, a range of different content notifications may confuse the public. Or, in the case of widespread voluntary content labelling, users may assume that the *absence* of a label indicates that content is human-generated. Measures to determine authenticity of human-generated content are out of scope for this guidance, but are closely related.
- Watermarking and labelling techniques are vulnerable to being used maliciously and manipulated or removed, undermining trust in the system. As a result, we should also be pursuing tools of provenance to assist with more persistent trust capabilities in AI content.

It is also critical that systems that display or sell AI-generated content (for example, social media platforms or retailers) make transparency mechanisms and information visible.

Further work across government

This voluntary guidance gives Australian businesses access to up-to-date best practice approaches to AI-generated content transparency. These are based on the latest research and international governance trends – both of which are rapidly evolving. This guidance complements and recognises other Australian Government initiatives which impact AI-generated content transparency including:

- consultation on the establishment of mandatory guardrails for high-risk AI
- the work of the Copyright and Artificial Intelligence Reference Group
- OAIC Guidance on privacy and the use of commercially available AI products
- OAIC Guidance on privacy and developing and training generative AI models
- NSA / ASD's ACSC Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era.

Under Australia's *Online Safety Act 2021*, the eSafety Commissioner regulates mandatory industry codes and standards. These set out online safety compliance measures to address certain systemic online harms. There are a range of enforcement mechanisms for services that do not comply, including civil penalties. Watermarking, labelling or equivalent measures are part of the Online Safety (Designated Internet Services – Class 1A and Class 1B Material) Industry Standard 2024 (DIS Standard). This requires that certain generative AI service providers – those with a risk of being used to produce high impact material like child sexual abuse material – put into place systems, processes and technologies that differentiate AI outputs generated by the model, as well as other obligations. The Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material) requires search engine providers to make it clear when users are interacting with AI-generated materials, among other obligations.

Taking action to address harmful deepfakes

As well as initiatives to improve the transparency of AI-generated content, the Australian Government continues to act against technology-facilitated harms associated with deepfakes.

Deepfakes are digital images, videos or sound files of a real person that have been edited to create an extremely realistic but false depiction of them doing or saying something that they did not actually do or say.

The non-consensual sharing of sexual or intimate material online, including artificially generated deepfake material, is a serious form of technology-facilitated abuse. It often occurs in the context of gender-based and family, domestic and sexual violence.

The eSafety Commissioner's complaints schemes all apply to deepfake material. This includes its scheme applying to image-based abuse. This is when a person shares, or threatens to share, an intimate image or video of someone without their consent.

Last year, the Australian Parliament passed the Criminal Code Amendment (Deepfake Sexual Material) Act 2024. The amendment strengthens existing Commonwealth criminal offences and creates new offences targeting the creation and non-consensual sharing of sexually explicit material online. This includes material that has been created or altered using technology, such as deepfakes.

These civil and criminal schemes work in a complementary manner to provide choice and redress for victim-survivors.

The current phase of online safety code development focuses on ensuring safety measures are in place to prevent children in Australia from accessing or being exposed to Class 1C and Class 2 material (such as online pornography). This includes through risk-appropriate safety measures for AI-generated material. The 'Phase 2' codes also aim to ensure online services have safety measures and tools in place to allow all end-users to manage their online experiences with Class 1C and Class 2 material. eSafety published a Position Paper in July 2024 to guide the development of safety measures. This seeks to ensure the Phase 2 safety measures support a meaningful uplift in online safety practices and responsibilities in respect of AI-generated content, particularly to protect children in Australia. In early 2025, industry associations representing the online industry submitted 9 codes to the eSafety Commissioner for consideration of whether they create appropriate community safeguards.

More information about code development is available in the industry codes section of the eSafety website.

Additionally, the Online Safety (Basic Online Safety Expectations) Determination 2022 sets out the Australian Government's expectations that digital providers keep Australians safe online. This includes social media services, messaging, gaming and file sharing services, apps, websites and other services. There are specific expectations around generative AI:

- that providers will take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of generative AI capabilities on their services
- that providers take reasonable steps to proactively minimise the extent to which generative AI capabilities may produce material or facilitate activity that is unlawful or harmful.

While the expectations are not backed by civil penalties, the eSafety Commissioner can require providers to report on how they are meeting the expectations (with civil penalties available for non-compliance). eSafety regularly publishes transparency reports summarising provider's responses to improve transparency and promote greater accountability for user safety.

The Australian Communications and Media Authority (ACMA) oversees the operation of the voluntary Australian Code of Practice on Disinformation and Misinformation. The code includes measures that support transparency around the steps that signatories take to empower consumers to make better informed choices about digital content. This may include information about digital literacy interventions and the use of new technologies to signal the credibility of information online. The ACMA's oversight role includes reporting to government the effectiveness of the code.

Resources

UNESCO – Ethics of Artificial Intelligence

Australia's AI Ethics Principles

Australia's Voluntary AI Safety Standard

Australia's e-Safety – Safety by Design initiative and principles

Australian Signals Directorate – paper on Content Credentials

OAIC Guidance on privacy and the use of commercially available AI products

OAIC Guidance on privacy and developing and training generative AI models

Partnership on AI's Responsible Practices for Synthetic Media

Online misinformation | ACMA

ISO/CD 22144 – Authenticity of information – Content credentials

ISO/IEC 21617-1:2025 – Information technology – JPEG Trust – Part 1: Core foundation