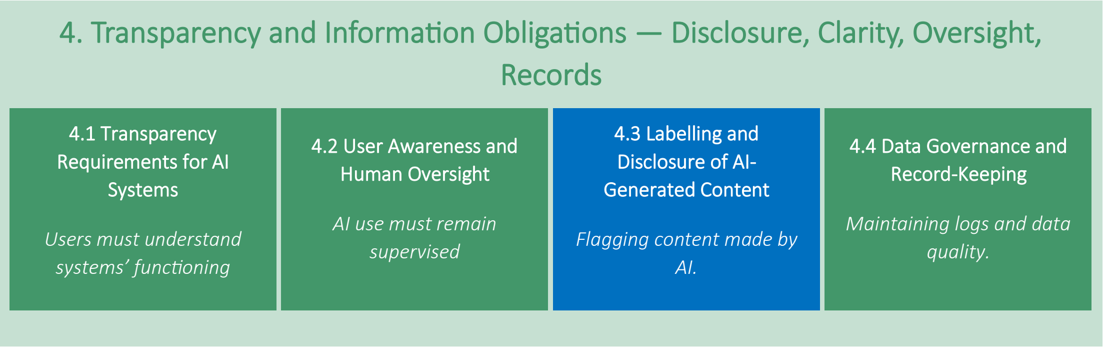
# Labelling and Disclosure of AI-Generated Content
## Flagging content made by AI

Co-authored with **Jessica Mendoza**, *AI Safety Advocate*

1 September 2025



4. Transparency and Information Obligations — Disclosure, Clarity, Oversight, Records

4.1 Transparency Requirements for AI Systems

*Users must understand systems' functioning*

4.2 User Awareness and Human Oversight

*AI use must remain supervised*

4.3 Labelling and Disclosure of AI-Generated Content

*Flagging content made by AI.*

4.4 Data Governance and Record-Keeping

*Maintaining logs and data quality.*

## Introduction: The Rise of AI-Generated Content

Artificial intelligence is increasingly used to generate text, images, audio, and video that are nearly indistinguishable from human-created material. From automated news articles to deepfake videos and synthetic voices, AI-generated content is becoming pervasive across digital platforms. While such content offers innovative possibilities in media, entertainment, and communication, it also raises pressing questions about authenticity, trust, and transparency.

To address these concerns, the EU AI Act introduces explicit obligations to label and disclose AI-generated content. These requirements aim to ensure that users understand when content has been created or manipulated by AI systems, reducing the risk of deception and reinforcing informed engagement in digital spaces.

## The Importance of Disclosure

AI-generated content can be persuasive and realistic. In some cases, it may intentionally or unintentionally mislead audiences, particularly when it replicates human likenesses, mimics real-world events, or conveys false narratives. Without proper disclosure, such content can erode public trust, fuel misinformation, and compromise democratic processes.

Labelling AI-generated content serves multiple purposes:

- It empowers users to interpret content with appropriate scepticism.
- It upholds transparency in digital communication.
- It ensures accountability for content creators and distributors.
- It helps regulators identify and respond to misuse.

By clearly flagging synthetic material, the Act promotes a more honest and navigable information environment.

## What Must Be Disclosed

Under the AI Act, any content that is generated or manipulated by an AI system in a way that may reasonably be mistaken for authentic or human-made must be labelled accordingly. This includes, but is not limited to:

- Text created by large language models
- Synthetic images or videos of people or events
- AI-generated voices or music
- Deepfake media that alters or fabricates real individuals' likenesses

The disclosure must be clear, prominent, and understandable to the average user. Labels such as "AI-generated," "synthetic media," or "created using artificial intelligence" are examples of acceptable indicators. These labels should be included directly within the content or its immediate presentation—such as subtitles, watermarks, or metadata—depending on the format.

The aim is to ensure that users are made aware of the content’s nature at the point of consumption, not buried in technical details or behind user interactions.

## Scope of Application

The labelling requirement applies broadly across sectors and platforms. Any provider or deployer of AI systems used to generate public-facing content is responsible for ensuring that disclosure is implemented. This includes:

- Media organizations using AI for article generation
- Marketing agencies deploying synthetic influencers
- Entertainment companies creating digital characters or avatars
- Platforms that distribute user-generated content involving AI tools

Where content is created for internal use, artistic expression, or non-public settings, the obligation may be less stringent. However, if there is potential for public misunderstanding or reputational harm, labelling remains strongly encouraged.

For public figures or institutions, the threshold for mandatory disclosure may be higher, especially if the content could influence public opinion or affect electoral integrity.

## Deepfakes and Synthetic Personas

Deepfakes—AI-generated media that convincingly imitates real individuals—are a particular focus of the Act’s labelling provisions. These technologies can create video or audio that shows people saying or doing things they never actually said or did.

The AI Act mandates that deepfake content must be explicitly labelled, except in limited cases where disclosure would compromise legitimate public interests, such as journalism or law enforcement. Even in such cases, appropriate safeguards must be in place.

Synthetic personas, such as AI-generated influencers or virtual news presenters, must also disclose their non-human nature. Users should not be left to assume they are interacting with or viewing a real person when, in fact, the character is entirely artificial.

This rule reflects the principle that consent and comprehension require knowing whether one is engaging with a person or a machine.

## Exceptions and Special Circumstances

While the default requirement is to disclose AI-generated content, the AI Act allows for narrowly defined exceptions. These include:

- Law enforcement or intelligence operations where disclosure would compromise effectiveness
- Research or artistic expression, provided the content is not misleading in a harmful context
- Content where the synthetic nature is obvious and does not risk deception

Even in these cases, the burden is on the provider or deployer to demonstrate that an exemption is justified. Where applicable, alternative measures—such as context-based warnings or time-limited exceptions—may be required to balance transparency with other legitimate aims.

## Technical Implementation and Best Practices

To support consistent disclosure, the AI Act encourages the development of technical standards and best practices for labelling. These may include:

- Watermarks embedded in images or videos
- Metadata fields indicating AI authorship
- Visual indicators or tags on social media platforms
- Audible disclaimers for synthetic speech

Providers are expected to adopt solutions that are proportionate to the content type and platform. The labeling must be resistant to tampering and remain intact as the content is shared or modified. Digital platforms may also be required to detect and preserve such indicators during upload or distribution.

These technical safeguards are essential to preventing the erasure or manipulation of labels during downstream use.

## Responsibilities of Providers and Platforms

The obligation to label AI-generated content falls primarily on the provider of the AI system or the entity deploying it to generate content. However, digital platforms that host or distribute such content also bear responsibilities. They may be required to:

- Detect unlabelled AI-generated material
- Enforce labelling standards through moderation or automated tools
- Notify users when they have encountered synthetic content
- Take corrective action when content is found to be deceptive or non-compliant

This shared responsibility model helps to ensure that transparency is maintained throughout the content lifecycle—from creation to dissemination.

Providers must also maintain documentation showing that they have implemented appropriate labelling mechanisms. This documentation may be reviewed as part of conformity assessments, audits, or post-market investigations.

## Enforcement and Penalties

Failure to properly disclose AI-generated content can result in regulatory penalties under the AI Act. These may include:

- Administrative fines proportional to the seriousness of the violation
- Orders to remove or relabel content
- Public disclosure of non-compliant behaviour
- Suspension of the system's deployment in the EU market

Enforcement is managed by national supervisory authorities, with coordination from the EU AI Office to ensure consistency across Member States. Individuals may also have recourse to complaints or remedies if they are misled by unlabelled content.

The penalty framework is designed not only to deter misconduct but also to encourage proactive compliance by providers and platforms.

## Conclusion

As AI systems gain the ability to produce ever more convincing synthetic content, transparency becomes a cornerstone of responsible deployment. The EU AI Act addresses this need through clear, enforceable obligations to label and disclose AI-generated material.

These requirements are not aimed at stifling creativity or innovation. Rather, they promote trust, integrity, and user awareness in an increasingly complex digital environment. By ensuring that people can recognize when content is created by machines, the Act helps preserve the credibility of public discourse and supports informed engagement with emerging technologies.

In the long run, clear labelling of AI-generated content is not just a legal duty—it is a vital step toward sustaining truth and authenticity in the digital age.

## Glossary

**Act or EU AI Act**: European Union Artificial Intelligence Act

**AI**: Artificial Intelligence

**Board**: European Union Artificial Intelligence Board

**EU**: European Union

**SME**: Small and Medium-Sized Enterprise

## How can we help?

AI & Partners
Amsterdam - London - Singapore

AI & Partners ' –AI That You Can Trust'

At AI & Partners, we're here to help you navigate the complexities of the EU AI Act, so you can focus on what matters—using AI to grow your business. We specialize in guiding companies through compliance with tailored solutions that fit your needs. Why us? Because we combine deep AI expertise with practical, actionable strategies to ensure you stay compliant and responsible, without losing sight of your goals. With our support, you get AI you can trust—safe, accountable, and aligned with the law.

To find out how we can help you, email contact@ai-and-partners.com or visit [https://www.ai-and-partners.com](https://www.ai-and-partners.com).