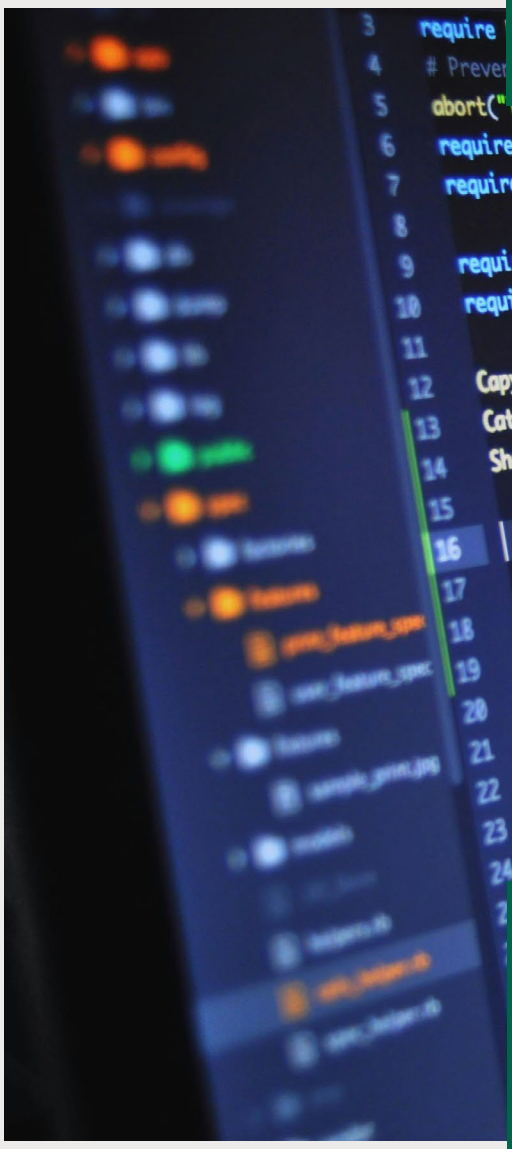
House of Commons Library

Research Briefing

20 January 2026

By Elizabeth Rough,
Jack Clift

AI content labelling



Summary

1 What is AI content labelling and why is it being used?
2 How does AI content labelling work?
3 What information should be in a label?
4 Are there requirements to label AI-generated content?

Image Credits

Close Up Photo of Programming of Codes by luis gomes on Pexels, published 4 August 2017

Disclaimer

The Commons Library does not intend the information in our research publications and briefings to address the specific circumstances of any particular individual. We have published it to support the work of MPs. You should not rely upon it as legal or professional advice, or as a substitute for it. We do not accept any liability whatsoever for any errors, omissions or misstatements contained herein. You should consult a suitably qualified professional if you require specific advice or information. Read our briefing 'Legal help: where to go and how to pay' for further information about sources of legal advice and help. This information is provided subject to the conditions of the Open Parliament Licence.

Sources and subscriptions for MPs and staff

We try to use sources in our research that everyone can access, but sometimes only information that exists behind a paywall or via a subscription is available. We provide access to many online subscriptions to MPs and parliamentary staff, please contact hoclibraryonline@parliament.uk or visit commonslibrary.parliament.uk/resources for more information.

Feedback

Every effort is made to ensure that the information contained in these publicly available briefings is correct at the time of publication. Readers should be aware however that briefings are not necessarily updated to reflect subsequent changes.

If you have any comments on our briefings please email papers@parliament.uk. Please note that authors are not always able to engage in discussions with members of the public who express opinions about the content of our research, although we will carefully consider and correct any factual errors.

You can read our feedback and complaints policy and our editorial policy at commonslibrary.parliament.uk. If you have general questions about the work of the House of Commons email hcenquiries@parliament.uk.

Contents

Summary **4**

1 What is AI content labelling and why is it being used? **6**

2 How does AI content labelling work? **7**

2.1 Visible disclaimers 7

2.2 Invisible watermarks 8

2.3 What are the limitations of AI content labelling? 11

3 What information should be in a label? **12**

3.1 How do users interpret AI content labels? 13

4 Are there requirements to label AI-generated content? **16**

4.1 Labelling rules and regulations 16

4.2 What are the labelling policies of social media companies? 18

4.3 What are the labelling policies of news organisations? 29

4.4 What are the labelling policies of search engines? 33

4.5 What are the labelling policies of video game services? 37

Summary

‘Generative artificial intelligence’ (AI) refers to systems that use machine learning to create new content such as text, images, audio, video and code. Content produced by generative AI is becoming increasingly realistic. This makes it harder for people to distinguish between content created with, and without, AI.

What is AI content labelling and why is it being used?

AI content labelling is used to alert people when they are engaging with content that has not been created by humans. It involves marking content that has been generated or altered by AI to help people understand its origins and assess its reliability.

AI content labelling is particularly relevant in the context of ‘deepfakes’, which are AI-generated videos, images or audio files deliberately designed to appear real. Deepfakes can spread disinformation and be used to commit fraud and depict individuals in misleading ways. Labels applied to deepfakes are sometimes described as “impact-based labels”, as they highlight the potential for harm.

Not all AI-generated or AI-altered content sets out to deceive or is inherently misleading. “Process-based labels” aim to communicate to users how a particular piece of content was created, including whether AI was involved, rather than comment on its potential to cause harm or other consequences.

How does AI content labelling work?

AI content labelling can take different forms. Visible disclaimers are explicit labels placed on content, such as text overlays, captions, watermarks or audio prompts. For example, AI chatbots may be labelled as “simulation” or “parody” to clarify that users are interacting with software rather than a real person. Tags may also be embedded in the metadata for a piece of content to provide information about its origins and to indicate how AI was involved in its creation.

The Coalition for Content Provenance and Authenticity (C2PA), for example, has developed “content credentials”, an open-source protocol using cryptography to encode details about a piece of content’s origin and editing

history. A watermark label, usually a speech bubble containing “cr”, accompanies the content and provides provenance information when it is clicked. Companies such as Adobe have adopted this system and integrated it into their software, and platforms like LinkedIn and Meta use it to label AI-generated content.

Invisible digital watermarks are also being developed for a range of digital content. Invisible digital watermarks embed technical signals in content that reveal its origin and/or composition. While these signals are invisible to people viewing the content, they can be detected by specialised algorithms.

There is not yet a consensus on the most effective design for an AI content label. The answer will partly be determined by what a label is setting out to achieve. For example, is it to highlight the use of generative AI in the creation of a piece of content (a ‘process-based’ label), or is to alert people that a piece of content could be misleading (an ‘impact-based’ label)?

Are there requirements to label AI content?

In the UK, there is no legislation requiring AI-generated content to be labelled. The UK Government’s consultation on Copyright and Artificial Intelligence, which closed in December 2024, acknowledged the benefits of clear AI labelling, but noted that there were “technical challenges involved”. Proposals for wider AI regulation in the UK have been delayed, and it is unclear whether future legislation will include labelling requirements.

In the European Union, article 50 of the EU AI Act sets transparency rules for content produced by generative AI. For example, providers of AI systems that interact with humans must alert users when they are interacting with AI, while providers of AI systems that generate or manipulate content must mark outputs in a machine-readable way. These rules are due to take effect in August 2026, although the European Commission has proposed delaying implementation until 2027.

Company policies on AI content labelling

Social media companies, news organisations, search engines and gaming platforms have adopted varying approaches to labelling AI-generated content. Social media companies variously rely on automatic detection technologies and user disclosures to label AI content, and in some cases differentiate between AI-edited and AI-generated content. Many news organisations include guidelines on the use of AI-generated content in their editorial policies, although published, publicly available guidelines are not universal. The policies of search engines and gaming platforms on AI-generated content vary significantly and are often partly determined by the policies of the companies that own or operate them.

1 What is AI content labelling and why is it being used?

Generative artificial intelligence (AI) applies machine learning techniques to large datasets to create new content in the form of text, images, audio, video and code.

Content produced by generative AI is becoming increasingly realistic. This makes it harder for people to distinguish between content created with, and without, AI.

AI content labelling is a broad term that is used to describe different methods of marking content as AI-generated or AI-altered. Labelling can alert users when the content they are engaging with online contains elements that have not been created by humans.

Marking images, video and audio in this way has been prompted by concerns that AI-generated content could mislead and deceive users, particularly through the circulation of 'deepfakes'. Deepfakes are "AI-generated videos, images and audio content that are deliberately created to look real".[1] While deepfakes can spread disinformation, they can also be used to commit fraud and scams, and to depict people in ways that undermine their character. These types of labels have been described as "impact-based" and are applied to content to denote the potential for harm.[2] The aim of impact-based labelling is to decrease the likelihood that AI-generated or AI-altered content "misleads or deceives its viewers".[3]

However, not all AI-generated or AI-altered content sets out to deceive. Nor is all AI-generated content inherently misleading. "Process-based" labels aim to communicate to users how a particular piece of content was created, including whether AI was involved, rather than comment on its potential to cause harm or other consequences. Process-based labels aim to increase transparency and enable users to make informed decisions about the authenticity of a piece of content and whether to trust it or not.

1 Ofcom, How can tech firms help users spot deepfakes?, July 2025
2 Chloe Wittenberg and colleagues, Labeling AI-Generated Content: Promises, Perils, and Future Directions, November 2023, p3
3 Chloe Wittenberg and colleagues, Labeling AI-Generated Content: Promises, Perils, and Future Directions, November 2023, p3

2 How does AI content labelling work?

There is not yet a consensus on the 'best' approach to AI content labelling, and a variety of approaches are being trialled.[4] This section outlines some of the different types of labelling.

2.1 Visible disclaimers

Visible disclaimers are explicit labels placed on AI content itself. The aim of visible disclaimers is to inform people that content may be AI-generated or AI-altered. Images, videos and texts, for example, may feature a written overlay (a title, caption or annotation on the content) or a watermark that alerts a viewer to the use of AI in their production and/or delivery. Audio content can be labelled with voice prompts. Tags may also be embedded in the metadata for a piece of content to provide information about its origins and to indicate how AI was involved in its creation.[5]

The exact wording featured in an AI content label will vary from platform to platform and may depend on the type of content and what the label is trying to achieve (see section 3 below). AI chatbots, for example, are a software application that uses AI capabilities, like natural language processing, to provide human-like conversation through text or voice. AI chatbots are increasingly used on websites to provide basic customer support functions and may feature an image (sometimes referred to as an 'avatar') of a person. To show a user that they are interacting with an AI chatbot and not a real person, the chatbot may be labelled as a 'simulation' or 'parody'.[6]

Placing a visible disclaimer on a piece of AI-generated content assumes that the origins of the content, and the use of AI in its creation, can be identified. This may be achieved by content creators actively self-disclosing their use of AI and labelling the content themselves. Labels may also be applied by platforms after using 'AI content detectors'. These are tools that use statistical analysis and machine learning to determine whether content (typically text) is generated by a human or by AI.

4 University of Graz, Options for AI labeling - Teaching and Learning with Artificial Intelligence, accessed 15 December 2025

5 Metadata typically provides information about, and the characteristics of, the data such as its author and the date it was created.

6 US National Telecommunications and Information Administration, AI Output Disclosures: Use, Provenance, Adverse Incidents, March 2024

2.2 Invisible watermarks

Other approaches are being developed for marking and identifying the provenance of a piece of content.[7] Broadly speaking, these provenance-identifying mechanisms involve embedding 'technical signals' in a piece of content that provide details of its origin and/or composition. While these signals are invisible to viewers, they can be detected by specialised algorithms.[8] This approach has been referred to as digital or AI 'watermarking', as the International Telecommunication Union (ITU) explains:

> AI watermarking involves embedding markers into multimedia content for it to be accurately identified as AI-generated. The technology is composed of two parts: the watermark-embedding algorithm and the watermark-detection algorithm. This combination creates a unique, identifiable signature – invisible to humans but detectable through algorithms and traceable back to the original AI content-generation model.
>
> Watermarks are created during the model-training phase by teaching the model to embed a specific signal or marker in the content it generates. Then, after the AI model is deployed, specialised algorithms can detect the presence of that embedded "watermark. Alongside identifying content as AI-generated, well-designed watermarking should also make clear the provenance of such content.[9]

Some platforms are developing watermarking tools. Google's SynthID is an invisible watermark that is automatically added to media generated by Google's AI tools.[10] This includes audio that is generated or published through Lyria (Google's AI music-generation model), text generated by Gemini, (Google's AI assistant) and photos edited using Google Magic Editor (Google's AI photo editing tools).

To detect a SynthID, users can upload an image to Gemini and ask if it has been created or altered by Google AI. Gemini will then look for a SynthID.[11] For other types of content, Google is trialling a SynthID detector (verification portal) with media professionals which can check if a piece of content was watermarked with a SynthID. According to Google, the purpose of SynthID is to help "foster transparency and trust in generative AI".[12]

Google and other companies (including Meta, OpenAI, Adobe and the BBC) are also part of the Coalition for Content Provenance and Authenticity (C2PA). C2PA has developed an approach to content authentication called 'content credentials'. Content credentials is an "open-source protocol that relies on

7 The focus of this section is on watermarking content. However, methods have been developed to watermark AI models (to prevent the theft of the models) and to watermark the data sets used to train AI models (to address any unauthorised use of the data).

8 European Parliamentary Research Service, Generative AI and watermarking, December 2023, PDF

9 ITU, AI watermarking: A watershed for multimedia authenticity, May 2024

10 IT Business Today, Google Launches SynthID for AI Content Detection, May 2025

11 SynthID - Google DeepMind, accessed 17 December 2025

12 SynthID - Google DeepMind, accessed 17 December 2025

cryptography to encode details about the origins of a piece of content".[13] A label (the 'content credential') remains with the piece of content throughout its lifespan, containing information about its provenance such as "when the content was created, where it came from, and its editing history".[14] This can include whether AI was used to generate or modify the content.

In its content credentials explainer, the C2PA states that content credentials "do not provide value judgments about whether a given set of provenance data is 'true'", but rather that the provenance information for a piece of content is an accurate reflection of how the content has been created and edited, and that this provenance information is free from tampering.[15]

Content credentials are an 'opt-in' technology.[16] Companies can opt in by integrating content credentials into their software. Adobe products (such as Photoshop) have included content credentials as a feature that users can choose to opt into themselves. Adobe automatically applies content credentials, however, to media that has been generated using its AI tools, like Adobe Firefly.[17] Sony and Nikon are also adding content credentials to new models of their cameras. In addition, individuals can choose to use C2PA's approach: where content credentials has not been integrated into a piece of software, content creators can still apply the information using the Adobe Content Authenticity app, regardless of how or where the content was created.

Section 9 of the C2PA's technical specification for content credentials outlines how content credentials use a combination of "hard" and "soft" bindings to provide information about the provenance of a file or a stream of data (known as an "asset").

Hard bindings typically use a "cryptographic hashing algorithm" to place "hashing" over all or part of the asset's data.[18] Hashing is a form of encryption that associates an asset with a string of unique, nondescript text.

Soft bindings used by content credentials include "a fingerprint computed from the digital content or an invisible watermark embedded within the

13 University of Maryland, The Inside Scoop on Watermarking and Content Authentication — NSF Institute for Trustworthy AI in Law & Society (TRAILS), November 2023

14 Content Credentials, About Content Credentials | Synthetic Media Detection, accessed 26 November 2025

15 Coalition for Content Provenance and Authenticity (C2PA), C2PA and Content Credentials Explainer, not dated

16 Coalition for Content Provenance and Authenticity (C2PA), C2PA and Content Credentials Explainer, not dated

17 Adobe, Content Credentials overview, updated September 2025. Adobe also has an 'inspect tool' so users can view the content credentials.

18 Coalition for Content Provenance and Authenticity (C2PA), Content Credentials: C2PA Technical Specification, not dated. Section 13.1 of the C2PA Technical Specification provides further details about the use of hashing in content credentials.

digital content".[19] A 'fingerprint' is a collection of the inherent properties (or metadata) of a piece of digital content, including things such as the pixels in an image or information about the device which took it.

Using soft bindings alongside hard bindings means that even if an asset's provenance data (or 'manifest') is removed, its provenance can still be traced.

For example, if someone uploads an image with C2PA content credentials to a platform that removes its metadata (such as information about the device which took the image), an invisible C2PA watermark (a soft binding) will survive. Someone else who views the image can then use a tool to decode the C2PA watermark and fetch the matching manifest from a repository that is stored remotely in cloud storage. The tool can use the manifest to check if the hashing (a hard binding) in the image that was uploaded to the platform matches the hashing in the original image that is stored in the repository.

In practice, when C2PA content credentials are attached to an asset such as an image or a video, it displays a label. The default label is a watermark of a white speech bubble containing "cr" (a cropping of the word "credentials"), which is usually placed in the top right-hand or top left-hand corner of the asset. Examples of the default content credentials label can be found on the Content Credentials official website and in an explainer published by Adobe, which adopted the use of content credentials across its products including Photoshop and Lightroom in 2023.[20] When a user clicks on the "cr" label, a pop-up window shows the asset's available provenance data. As discussed in section 4.2 below, some companies, such as LinkedIn, use the default C2PA "cr" label for assets that carry content credentials; other companies, such as Meta and TikTok, use their own labels for assets that carry content credentials for AI-generated content.

The International Press Telecommunications Council (IPTC) has also published best practice guidelines on "how to use embedded metadata to signal 'synthetic media' content that was created by generative AI systems".[21] The IPTC guidelines are specifically for AI-generated images, and Google has published information on how IPTC metadata for photos can be shown in the results of a Google Image search.

19 Coalition for Content Provenance and Authenticity (C2PA), Content Credentials: C2PA Technical Specification, not dated. Section 18.10 of the C2PA Technical Specification provides further details about the use of soft bindings in content credentials.

20 Content Credentials, About, not dated; Adobe, Content Credentials overview, updated 3 September 2025

21 International Press Telecommunications Council (IPTC), IPTC publishes metadata guidance for AI-generated "synthetic media", 9 May 2023

2.3 What are the limitations of AI content labelling?

AI content labelling faces several technical challenges, including the need to coordinate industry-wide standards and to ensure widespread adoption. Without common standards, a watermark "generated by one technology may not be readable by a system using different technology", thus requiring multiple detection systems.[22]

Both visible and invisible watermarks may be removed, manipulated or mimicked by malicious actors seeking to undermine the authenticity of a piece of content or to try and spread disinformation. The Information Technology Industry Council notes that watermarks on video are much harder to remove but adds that they can be "distracting and make content consumption less appealing".[23] In other words, the stronger the watermark, and the harder it is to remove, the more noticeable it may be to users.[24] In addition, embedding watermarks in the metadata of an image, for example, can increase the file size and require additional amounts of storage. This, in turn, can "sometimes make it harder to download and access certain types of content".[25]

22 European Parliamentary Research Service, Generative AI and watermarking, December 2023, PDF

23 Information Technology Industry Council (ITI), Authenticating AI-Generated Content (PDF), January 2024, page 11

24 DataCamp, AI Watermarking: How It Works, Applications, Challenges, February 2025

25 Information Technology Industry Council (ITI), Authenticating AI-Generated Content (PDF), January 2024, page 11

3 What information should be in a label?

There is not yet a consensus on the most effective design for an AI content label. The answer will partly be determined by what a label is setting out to achieve.

As outlined in section 1.1, there can be different reasons for labelling a piece of content as AI-generated or AI-altered. Sometimes, AI content labelling is to warn people about the potential for harm and the risk of a piece of content deceiving or misleading a user (an 'impact-based label'). At other times, it is to promote transparency and communicate how a piece of content was created (a 'process-based label').

Process-based labels could be applied to all forms of AI-generated or AI-altered content, regardless of whether the content risks misleading the user. Process-based labels also tend to place responsibility on the user to decide how they should interpret the label and engage with the content.[26]

Studies have examined what information a content label should convey, how detailed it should be, and what language/terminology people find helpful. For example, research examining the public understanding of AI content labels, conducted with over 5,000 participants across five countries, found that the amount of description used in a label can influence how it is interpreted.[27]

The researchers concluded that if the goal of a label is to identify AI-generated content that is misleading, the terms "AI-generated", "Generated with an AI tool" and "AI-manipulated" were insufficient because they were not automatically associated with content that aims to mislead.[28] In contrast, using terms like 'deepfake' and 'manipulated' on labels were "effective at identifying misleading content" but performed poorly for identifying AI-generated content.[29]

Other studies have considered whether labels like 'AI-generated' should be accompanied by further elaboration. This could include providing a brief description of the content in question and specifying the AI model used to

26 Chloe Wittenberg and colleagues, Labeling AI-generated media online, PNAS Nexus, May 2025

27 Brian Eastwood, MIT Sloan, How should AI-generated content be labeled?, November 2023

28 Ziv Epstein and colleagues, What label should be applied to content produced by generative AI? November 2023

29 Ziv Epstein and colleagues, What label should be applied to content produced by generative AI? November 2023

generate it. It could also involve a disclaimer explaining that the content may not accurately represent real-world events. [30]

Using terminology like 'AI-generated' assumes that someone will automatically know what this means and, therefore, how to interpret the label. This may not be the case. One study found that participants were confused by what the term 'AI-generated' meant, and whether the label implied that all the content was AI-generated, or whether AI had been used as part of the creative process or to enhance human-created content.

3.1 How do users interpret AI content labels?

There is currently a limited amount of information on exactly how users interpret AI content labels and how they may influence online behaviours. Some early findings have indicated that adding a label can reduce belief in, and trust of, the content, regardless of its underlying veracity or provenance.

For example, a research study published in October 2024 examined the impact of labelling news headlines and articles as AI-generated content. The experiments run by the researchers showed that a headline labelled as AI-generated reduced their perceived accuracy:

> labeling headlines as AI-generated lowered their perceived accuracy and participants' willingness to share them, regardless of whether the headlines were true or false, and created by humans or AI.[31]

Thus, while a process-based label may help a user to understand the provenance of a piece of content, it is potentially less informative when it comes to the accuracy or reliability of the content.

Similarly, a study involving over 7,500 people, published in 2025, found that placing a warning label on an AI-generated image (which depicted a scene that did not occur) reduced the "believability" of the content and led participants to rate the image as less credible.[32] Participants also said that they would be less likely to share the image with others.

Adding a process-based label stating that AI was involved in the generation of an image, but not using language that 'warned' users about the content, was also found to decrease the believability of the content. It did not,

30 Jingruo Chen and colleagues, Examining the Impact of Label Detail and Content Stakes on User Perceptions of AI-Generated Images on Social Media, CSCW Companion '25: Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing, Pages 270 - 275, October 2025

31 Sacha Altay and colleagues, People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation, PNAS Nexus, October 2024

32 Chloe Wittenberg and colleagues, Labeling AI-generated media online, PNAS Nexus, May 2025

however, noticeably affect participants' propensity to engage with the content (such as sharing, 'liking' and/or commenting on it on social media).[33]

AI content labels may not influence people's views on uncontroversial topics

In other contexts, however, content labels have been found to have no impact on the persuasiveness of a piece of content. A team of researchers at Stanford University in California surveyed over 1,500 people in the US to investigate how adding authorship labels affected the perception of AI-written policy arguments. [34] The policy arguments were on topics the researchers defined as "less polarized" policy matters, such as allowing colleges to pay student athletes.[35] The researchers found that people's judgements of the accuracy of the policy arguments were not significantly affected by an AI label:

> … we found no significant difference between people's support for a policy argument when told the argument had been generated by an expert AI model, a human policy expert, or told nothing about its authorship. The labels also had no significant effects on people's judgments of the content's accuracy or people's intentions to share the policy argument with others.[36]

The researchers concluded that while AI content labels may improve transparency, their inability to affect persuasiveness "highlights the need for complementary safeguards", such as media literacy education "which aims to equip people with the tools to discern AI-generated writing from human-authored writing as well as the validity of the information itself".[37]

The effect of AI content labels may depend on the type of content being labelled

Other studies have indicated that user reactions to, and engagement with, labelling depend on the nature of content. Several have distinguished between "high-stakes" content (like news reporting, images of police and protests) and "low-stakes" content (such as images of scenery, content designed for entertainment).[38]

They have found that users show greater levels of scepticism toward AI-generated content in high-stakes contexts.

---

33 Chloe Wittenberg and colleagues, Labeling AI-generated media online, PNAS Nexus, May 2025

34 Stanford University, Labelling AI-Generated Content May Not Change Its Persuasiveness, July 2025

35 Until recently there were limits on how student athletes in the US could be compensated and how much they could receive.

36 Stanford University, Labelling AI-Generated Content May Not Change Its Persuasiveness, July 2025

37 Isabel O. Gallegos and colleagues, Labeling AI-Generated Content May Not Change Its Persuasiveness, Policy Brief, Stanford University Human-centred AI Policy & Society, July 2025, p5, PDF

38 See for example Olivia Burrus and colleagues, Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content, Interactions, July-August 2024, p38-42

By contrast, users demonstrated more engagement with, and trust in, AI-generated content in low stakes contexts.[39] One study suggested that this may stem from "reduced perceived risk, increased familiarity and lower emotional investment" in low-stakes content.[40]

In addition, users have been found to link the authenticity of a piece of content to the extent to which AI is being used, as one study explained:

> edits [of images] perceived by viewers to be examples of simple autoenhancements, such as blemish removal or color corrections (whether accurate or not), have less impact on perceptions of authenticity than those perceived as complex GenAI production, such as turning a frown into a smile, or adding, removing, or replacing an object.[41]

39 Jingruo Chen and colleagues, Examining the Impact of Label Detail and Content Stakes on User Perceptions of AI-Generated Images on Social Media, CSCW Companion '25: Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing, Pages 270 - 275, October 2025

40 Jingruo Chen and colleagues, Examining the Impact of Label Detail and Content Stakes on User Perceptions of AI-Generated Images on Social Media, CSCW Companion '25: Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing, Pages 270 - 275, October 2025

41 See for example Olivia Burrus and colleagues, Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content, Interactions, July-August 2024, p38-42

4 Are there requirements to label AI-generated content?

Requirements to label content as AI-generated are emerging globally, with a mixture of legislative developments and platform policies either mandating, or encouraging, labelling. This section examines the regulations, and company policies, that could affect AI content labelling in the UK. The list of company policies is not exhaustive and should be taken as correct at the time of writing.

4.1 Labelling rules and regulations

UK

There is no UK legislation that requires AI-generated content to be labelled as such. The government's consultation on Copyright and Artificial Intelligence, which closed in December 2024, said that the government was in favour of clear AI content labelling:

> The government believes that clear labelling of AI outputs would be beneficial to right holders and the public, but acknowledges the technical challenges involved. The government welcomes views on how to achieve this outcome, including how we can support the development of emerging tools and standards.[42]

It added that "regulation may be needed to help ensure labelling happens consistently" and that the government will "consider the case for it as it further develops its approach to AI regulation".[43] In July 2025, the UK communications regulator, Ofcom, also published a discussion paper on how different labelling tools and techniques could be used to identify deepfakes.

In response to a parliamentary question in March 2025 about regulating AI, the government said that it was currently refining its legislative proposals:

> The government is clear in its ambition to bring forward legislation which allows us to safely realise the enormous benefits and opportunities of the most

[42] Intellectual Property Office, Department for Science, Innovation and Technology and Department for Culture, Media and Sport, Copyright and Artificial Intelligence, December 2024

[43] Intellectual Property Office, Department for Science, Innovation and Technology and Department for Culture, Media and Sport, Copyright and Artificial Intelligence, December 2024

> powerful AI systems for years to come. The government is continuing to refine its proposals and will launch a public consultation in due course.[44]

At the time of writing, the government had not launched a consultation on regulating AI. The Guardian newspaper reported in June 2025 that proposals to regulate AI more broadly had been delayed "by at least a year".[45] It added that the government "intends to introduce a 'comprehensive' AI bill in the next parliamentary session to address concerns about issues including safety and copyright".[46] At this stage, it is unclear whether any future AI bill will include measures relating to content labelling.

European Union

The situation is different in the EU. Under article 50 of the EU AI Act, there are transparency rules for content produced by generative AI:

- Providers of AI systems that interact with humans (such as AI chatbots) must alert users that they are interacting with AI and not another human.
- Providers of AI systems that generate or manipulate content (audio, image, video or text) must mark those outputs in a way that is machine-readable and detectable.
- Deployers of emotion recognition or biometric categorisation systems must inform users that they have been exposed to this type of system. Emotion recognition systems are AI systems which infer a person's emotions (for example, happiness, sadness, surprise, shame) or intentions from their biometric data. For example, a smart watch may infer excitement or fear if it measures a person's heart rate increasing.
- Deployers of AI systems that generate deepfakes must disclose that content has been artificially generated/manipulated.

Article 50 is due to become effective from August 2026. However, in November 2025, the European Commission proposed delaying the implementation of article 50 (and other parts of the AI Act that govern 'high-risk' AI systems) until 2027.[47]

The EU AI Act tasks the European Commission with adopting guidelines on article 50 and developing a code of practice. The latter will focus specifically on how to comply with the obligation to make AI-generated or AI-manipulated content detectable. The guidelines and code of practice are currently being developed.

---

44 PQ 41098 [on Artificial Intelligence: Regulation], 31 March 2025

45 The Guardian, UK ministers delay AI regulation amid plans for more 'comprehensive' bill, 7 June 2025

46 The Guardian, UK ministers delay AI regulation amid plans for more 'comprehensive' bill, 7 June 2025

47 European Commission, Digital Omnibus on AI Regulation Proposal, November 2025

The EU AI Act applies to all AI systems placed on the market or used in the EU, regardless of where the provider, or the AI system, is based. It also covers AI systems used outside the EU if the output produced by it is used in the EU.

4.2 What are the labelling policies of social media companies?

Social media companies' policies on the use and labelling of AI-generated content on their platforms vary. Some companies, such as Meta and TikTok, have different labelling policies for:

- content that has been edited or modified using generative AI tools
- content that has been wholly generated using AI tools

In some cases, social media companies' policies on manipulated or misleading content overlap with, or stand in place of, explicit policies on AI-generated content (see, for example, the discussion of Reddit below). X (formerly known as Twitter) is the only social media company discussed below that has not published any policies on the use and labelling of AI-generated content on its platform.

Social media companies have different rules about how users should disclose the use of AI. Companies including Meta, TikTok and YouTube have specific policies on users disclosing the use of AI-generated content, although these requirements sit alongside the use of technologies that identify and automatically label AI-generated content on these platforms. Other companies, such as LinkedIn, Pinterest and Snapchat, have published guidance on how they identify and label AI-generated content using technologies such as the "content credentials" developed by the Coalition for Content Provenance and Authenticity (C2PA), but do not specify how users should self-disclose the use of AI-generated content.

Similarly, social media companies make differing use of technologies for identifying and labelling AI-generated content. Some companies, such as Pinterest, use International Press Telecommunications Council (IPTC) metadata to help identify and label AI-generated content (discussed in section 2.2 above). Others, including LinkedIn, TikTok and YouTube, use the C2PA's "content credentials" technology (discussed in section 2.2 above). Meta uses a combination of IPTC and C2PA technologies, while Reddit and X do not specify which, if any, technologies they use to identify and label AI-generated content.

Ofcom, the UK's media regulator, has not issued any specific guidance on how social media companies should label AI-generated content. In June 2025, Ofcom issued its strategic approach to AI for 2025/26. Ofcom says that the industries it regulates are free to use AI as they see fit, though it also says that it will monitor and mitigate any associated risks:

> In general, our regulation is technology-neutral, which means regulated companies are essentially free to deploy AI as they see fit, without needing our permission, helping to enable faster innovation and growth.
>
> That said, while AI affords new opportunities and benefits for businesses and consumers, it is important for Ofcom to stay ahead of any associated risks, and take action to mitigate them.[48]

LinkedIn

LinkedIn uses the C2PA's "content credentials" to label AI-generated content. When the metadata of an image or video uploaded to the platform contains C2PA content credentials, LinkedIn applies the generic "cr" watermark to the content. When a user clicks on the "cr" watermark, LinkedIn displays the content credentials and any available metadata.

LinkedIn's guidance on content credentials notes that the platform is not able to identify all AI-generated content. It advises users to "remain mindful of the content you share and interact with" and reminds users that they can report any content that "appears to be misinformation or [that] otherwise violates [its] Professional Community Policies".[49]

LinkedIn's professional community policies on false or misleading content note that "deceptive synthetic or manipulated media, that does not clearly disclose the fake or altered nature of the content" may be removed from the platform, although AI-generated content is not mentioned specifically.[50]

LinkedIn provides a number of generative AI tools that help its users to create content including job descriptions and cover letters. While LinkedIn sets out the privacy policies for its generative AI tools for both members and non-members, it does not include any guidance for labelling content produced using its generative AI tools. In addition, LinkedIn does not provide an example wording for users to disclose that AI-generated content has been used in their posts.

Meta (Instagram, Facebook, Messenger, WhatsApp, Threads)

Meta owns and operates the social media platforms Instagram, Facebook, Messenger, WhatsApp and Threads.

How Meta identifies AI-generated content

In February 2024, Meta published information on how it identifies AI-generated content. This guidance differentiates between content produced

48 Ofcom, Ofcom's strategic approach to AI, updated 13 October 2025
49 LinkedIn, Content credentials, not dated
50 LinkedIn, False or misleading content, not dated

using Meta's own generative AI tool (Meta AI) and content produced using other generative AI tools or on other platforms.

The guidance states that Meta adds International Press Telecommunications Council (IPTC) metadata and "invisible watermarks" to photorealistic images generated through its Meta AI tool, and that these are "in line" with best practice guidelines developed by the Partnership on AI.[51]

In addition, the guidance states that Meta is contributing to the development of "common standards" for identifying AI-generated content on other platforms through the C2PA's "content credentials" and the IPTC's technical standards.[52]

Separately, Meta's guidance on labelling AI-generated content says that a label will be applied "based on our detection of industry-shared signals of AI images or people self-disclosing that they're uploading AI-generated content".[53] Meta has not published any further guidance on how users can self-disclose that they have used AI-generated content.

How Meta labels AI-generated content

Meta's guidance on labelling AI content distinguishes between content that has been edited or modified with AI tools and content that has generated by AI tools:

- Meta uses the "AI info" label for posts with content that has been edited or modified using AI tools. It displays the "AI info" label within a post's "menu", which users can access by clicking the three dots in the top right-hand corner of the post.

- Meta also uses the "AI info" label for posts with content that has been generated using AI tools. However, the "AI info" label is displayed more prominently in these posts and is located below the name of the user who has posted the content.

In both cases, clicking on the "AI info" label displays a disclosure which states that "content in this post may have been modified with AI". Sample images of the "AI info" label and the disclosure are included within Meta's guidance.

The "AI info" label replaced the previous "Made with AI" label in July 2024. Meta states that it made this change because the previous label was not "always aligned with people's expectations and didn't always provide enough context".[54]

---

51 Meta, Labeling AI-Generated Images on Facebook, Instagram and Threads, 6 February 2024

52 Meta, Labeling AI-Generated Images on Facebook, Instagram and Threads, 6 February 2024

53 Meta, Our Approach to Labeling AI-Generated Content and Manipulated Media, updated 12 September 2024

54 Meta, Our Approach to Labeling AI-Generated Content and Manipulated Media, updated 12 September 2024

In addition, Meta guidance published in 2025 specifies that the "AI info" label "may be required" on Meta's Facebook, Instagram and Threads platforms when content contains "photorealistic video or realistic-sounding audio that was digitally created, modified or altered, including with AI".[55] The guidance notes that users may incur "penalties" for sharing such content without a label, although the guidance does not specify the nature of these penalties.[56] This requirement does not apply to images, although images may receive a label if Meta detects that they were modified with or generated by AI. This guidance links directly to Meta's Community Guidelines on misinformation, which state that Meta will remove content, including AI-generated content, that promotes physical harm or violence, "harmful health misinformation", or "voter or census interference"; for all other misleading content, Meta says that it may "place an informative label on the face of the content" or "reject content submitted as an advertisement".[57]

Meta has different policies for labelling advertising content created using its generative AI tools. Guidance published in February 2025 states that Meta adopts a "multi-pronged approach" to labelling content produced by its "in-house generative AI tools":

- When the use of generative AI tools does not result in significant edits to an image or video, or does not include a photorealistic human, Meta does not apply an AI label.
- When the use of generative AI tools results in a significant edit to an image or video, Meta applies an AI label either within the post's "menu" or next to the "Sponsored" label that is included with advertising content.
- When the use of generative AI tools results in the inclusion of a photorealistic human, Meta applies an AI label next to the "Sponsored" label.[58]

In July 2025, Meta's Oversight Board, which adjudicates on content moderation issues, published recommendations that Meta "apply a relevant label to all content with the same manipulated media" and "ensure that the informative labels for manipulated media on Facebook, Instagram and Threads are displayed in the same language that the user has elected for its platform". Meta has not responded to the Oversight Board's recommendations and it has not published any new guidance on labelling manipulated media.

Pinterest

In April 2025, Pinterest announced that it would begin labelling AI-generated content as "AI modified". The "AI modified" label is shown in the bottom left-

55 Meta, How to identify AI content on Meta products, 2025
56 Meta, How to identify AI content on Meta products, 2025
57 Meta, Community Standards: Misinformation, 2025
58 Meta, Expanding GenAI Transparency for Meta's Ads Products, 3 February 2025

hand corner of a post, or "Pin", when a user views the post in close-up; the label is not shown when a user is browsing images using the grid view. Clicking on the "AI modified" label opens a disclosure that states that "We believe [this Pin] may have been modified with AI".[59]

Pinterest's guidance on generative AI labels states that the "AI modified" label is not displayed in the close-up view for advertisements. Instead, users must click on the three-dot menu next to the Pin and select "Why am I seeing this ad?". This opens a disclosure which states that "We believe this ad was modified with AI".[60]

Pinterest's guidance on generative AI labels also outlines how the platform identifies AI-generated content, stating that it uses IPTC metadata and its own generative AI detection tools:

> To do this, we not only analyse an image's metadata following the IPTC Metadata Standard, but also developed classifiers that automatically detect Gen AI content, even if the content doesn't have obvious markers.[61]

Pinterest has not published additional information on its generative AI detection tools.

Pinterest has not published any guidance on whether users can self-disclose AI-generated content, although users whose content has been mistakenly labelled as "AI modified" can appeal the label through Pinterest's help centre. Pinterest has noted that it is "working on ways to expand capabilities to better identify Gen AI content in the future through additional technologies".[62]

In October 2025, Pinterest introduced filters that allow users to opt out of seeing AI-generated content on the platform.

Reddit

Reddit does not have platform-wide rules on labelling AI-generated content. Instead, online communities within the platform, known as "subreddits", can make their own rules on how users share AI-generated content. A 2025 study of over 300,000 subreddits found that, in November 2024, only 1.2% had rules governing the use of AI-generated content, although this number had doubled since July 2023. The authors found that rules on AI-generated content "are more common in larger subreddits and communities focused on art or celebrity topics, and less common in those focused on social support".[63]

59 Pinterest, Introducing Gen AI Labels: Pinterest is Taking a New Step in Transparency, 30 April 2025
60 Pinterest, Gen AI labels, not dated
61 Pinterest, Gen AI labels, not dated
62 Pinterest, AI at Pinterest, not dated
63 Travis Lloyd and colleagues, AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content, Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 25 April 2025

Reddit's guidance on "manipulated content and misleading behaviour" contains platform-wide rules about the use of generative AI tools to create "manipulated content". The guidance prohibits "AI-generated content that deliberately misleads others about real-life events":

> Content created or modified using generative AI technologies is generally allowed on Reddit – subject to each community's specific rules and the Reddit Rules. However, this policy prohibits sharing AI-generated content that deliberately misleads others about real-life events or the actions of real-life individuals, or that presents itself as human-generated. When posting permissible AI-generated content, be transparent and include a tag (or other form of indication) disclosing that the content was generated or modified by AI to reduce confusion.
>
> We use a combination of reports by moderators, users, and internal systems to aggregate signals and determine whether someone's behaviour may constitute a violation under this policy.[64]

The guidance does not specify what types of labels for AI-generated content are acceptable.

The Reddit Rules include guidelines that could be applied to AI-generated content. Rule 2 asks users not to "engage in disruptive behaviours (including content manipulation) that interfere with Reddit communities" and rule 6 asks users to "ensure people have predictable experiences on Reddit by properly labelling content and communities".[65]

Snapchat

Snapchat has published guidance on the use of generative AI tools on its platform, which notes that it may use a number of approaches to indicate that a feature is powered by generative AI:

> We may indicate that a feature in Snapchat is powered by generative AI in a number of ways, including using the sparkle icon [...], specific disclaimers, Context Cards, or tool tips. When you see these contextual symbols or other indicators in Snapchat, you should know that you are interacting with AI and not a human, or viewing content that has been produced using AI and does not depict real world scenarios.[66]

The guidance also provides details of the "Snap Ghost with sparkles" watermark that "may be added" to images generated through some of Snapchat's AI tools. The guidance says that the watermarks "provide transparency":

64 Reddit, Manipulated content and misleading behaviour, updated October 2025
65 Reddit, Reddit Rules, accessed 21 December 2025
66 Snapchat, Guidance on the use of generative AI, not dated

> The purpose of these watermarks is to provide transparency that the image was created with generative AI and is not real or based on real events, even if it is a realistic style.[67]

The guidance outlines how users can report AI content that they do not like.

In addition, Snapchat's community guidelines say that the company reserves the right to terminate the accounts of users who abuse its AI products:

> Snap offers generative AI features through our services. We implement safeguards designed to help keep generative AI content in line with our Community Guidelines. We also expect Snapchatters to use AI responsibly. We reserve the right to take appropriate enforcement action against accounts that use AI to violate our Community Guidelines or abuse our AI products. This can include termination of an account.[68]

Snapchat does not provide any information on how it detects AI-generated content created on other platforms and shared on Snapchat, or how this content would be labelled. Similarly, Snapchat does not specify whether users can self-disclose the use of AI-generated content in what they share.

TikTok[69]

TikTok's guidance on AI-generated content states that images and videos that have been "significantly edited by AI" should be labelled on the platform. TikTok defines this as content that has been "modified by AI beyond minor corrections":

> We consider content that's significantly edited by AI as that which uses real images/videos as source material, but has been modified by AI beyond minor corrections or enhancements, including synthetic images/videos.[70]

TikTok's guidance states that users can disclose AI-generated content within their posts "by adding text, a hashtag sticker, or context in the post's description".[71] The social media consultancy firm Napolify has published a guide to TikTok's policies on AI-generated content that contains examples of acceptable labels, including "AI-generated", "synthetic content", "created with AI" and "#AIgenerated".[72]

TikTok's guidance also states that the platform offers two labels that can be applied to AI-generated content:

---

67 Snapchat, Guidance on the use of generative AI, not dated

68 Snapchat, Community guidelines, not dated

69 Since March 2023, access to TikTok, including its corporate website, has not been permitted on the UK Parliament network; BBC News, TikTok banned from UK Parliament over security concerns, 23 March 2023. The social media consultancy firm Napolify is an external website that provides a summary of TikTok's approach to labelling AI-generated content; Napolify, What are TikTok's AI content guidelines for 2026?, updated 26 November 2025

70 TikTok, About AI-generated content, not dated

71 TikTok, About AI-generated content, not dated

72 Napolify, What are TikTok's AI content guidelines for 2026?, updated 26 November 2025

- "Creator label". Users can add a "creator labelled as AI-generated" disclosure to their content during the upload process. Content that has been created using TikTok's own generative AI tools will be labelled automatically (see below). However, if a user has made further alterations to AI-generated content after using TikTok's generative AI tools, the user should add the "creator label" to their content. The guidance provides step-by-step instructions for users to apply the "creator label".[73]

- "Auto label". TikTok automatically applies an "AI-generated" label to content that it identifies as "completely generated or significantly edited with AI". TikTok uses the C2PA's "content credentials" to identify AI-generated content based on its metadata.[74]

In November 2025, TikTok announced updates to its policies on labelling AI-generated content. TikTok notes that its use of the C2PA's "content credentials" has helped to label more than 1.3 billion videos as AI-generated content to date.[75] TikTok also states that its own system of invisible watermarks mitigates the challenge of content credentials labels being removed when content is reuploaded:

> [...] a common industry challenge is that these kinds of [content credential] labels may be removed when content is reuploaded or edited elsewhere. "Invisible watermarks" add another layer of safeguards with a robust technological "watermark" that only we can read, making it harder for others to remove.[76]

TikTok also announced that it is trialling a new tool that enables users to select how much AI-generated content they would like to see in their "For You" feed using the "Manage topics" feature.[77] Users can access this feature by clicking on "Settings", then "Content Preferences", then "Manage Topics" and using the slider button to set the amount of AI-generated content they would like to see for different topics.[78]

X (formerly known as Twitter)

X (formerly known as Twitter) has not published any policies on labelling AI-generated content.

However, X has published guidance on managing what it calls "inauthentic content" on the platform. Under the "inauthentic content" heading within this guidance, X says that users may not share "synthetic and manipulated media":

---

73 TikTok, About AI-generated content, not dated

74 TikTok, About AI-generated content, not dated

75 TikTok, More ways to spot, shape and understand AI-generated content, 24 November 2025

76 TikTok, More ways to spot, shape and understand AI-generated content, 24 November 2025

77 TikTok, More ways to spot, shape and understand AI-generated content, 24 November 2025

78 Cybernews, TikTok gives its users more control over the AI-generated content they see, 21 November 2025

> You may not share inauthentic media, including, manipulated, or out-of-context media that may result in widespread confusion on public issues, impact public safety, or cause serious harm ('misleading media'). We may use our own technology or receive reports through partnerships with third parties in order to determine if media have been manipulated or presented out of context. In situations where we are unable to reliably determine if it is misleading media, we may not take action.[79]

None of the enforcement actions that X mentions in this guidance relate to labelling AI-generated content.

X's "community notes" function enables users who have registered as "contributors" to "collaboratively add context to potentially misleading posts" by adding notes to these posts.[80] Community notes only become publicly visible when enough contributors have marked the note as "helpful".

Ofcom investigation into Grok (January 2026)

Grok is a generative AI chatbot. It is available to download as an app for devices that use the iOS and Android operating systems and is fully integrated into the X social media platform. This means that X users can interact with Grok while using X by asking Grok questions or prompting it to generate AI content. Grok was developed by xAI, an artificial intelligence and technology company that was founded in 2023 by Elon Musk, the owner and executive chairman of X.

In January 2026, Ofcom announced it had launched a formal investigation into X over its Grok AI chatbot.[81] This followed media reports that Grok had been used to alter images of people to undress them without their consent and to create sexually explicit images of children.[82] Ofcom noted that the "undressed images of people" might amount to "intimate image abuse or pornography" and that the "sexualised images of children" might constitute "child sexual abuse material (CSAM)".[83]

Ofcom's investigation will focus on AI-generated content that is created and shared through Grok, and not on how or whether this content is labelled as AI-generated. Ofcom said that its investigation would assess whether X had failed to comply with its legal obligations under the Online Safety Act 2023, including taking down illegal content swiftly when the platform become aware of it and protecting children in the UK from seeing pornography. Further information on the provisions of the Online Safety Act can be found in the Commons Library briefings Online Safety Bill: progress of the Bill (October 2023), Implementation of the Online Safety Act (February 2025) and E-petition relating to the Online Safety Act (December 2025).

---

79 X (formerly Twitter), Authenticity, X Help, April 2025

80 X (formerly Twitter), About Community Notes on X, not dated

81 Ofcom, Ofcom launches investigation into X over Grok sexualised imagery, 12 January 2026

82 BBC News, Ofcom asks X about reports its Grok AI makes sexualised images of children, 5 January 2026

83 Ofcom, Ofcom launches investigation into X over Grok sexualised imagery, 12 January 2026

In a post responding to media reports about Grok, X said that it routinely removes illegal content including CSAM and permanently suspends accounts that share such content:

> We take action against illegal content on X, including Child Sexual Abuse Material (CSAM), by removing it, permanently suspending accounts, and working with local governments and law enforcement as necessary.
>
> Anyone using or prompting Grok to make illegal content will suffer the same consequences as if they upload illegal content.[84]

The post included links to the X Rules and X's range of enforcement options, although these had not been updated following the media reports.[85] The post reiterated an earlier post from Musk in which he said that anyone who used Grok to make illegal content would "suffer the same consequences" as those who upload the illegal content.[86]

X subsequently restricted access to Grok's image creation tools to paying subscribers.[87] Paying subscribers have to provide personal details, including credit card information, and can therefore be identified if Grok has been misused. However, some non-paying users of the Grok app said they could still "generate sexualised imagery of women and children". X later announced that users would not be able to edit photos of real people to show them in revealing clothing in jurisdictions where this is illegal.[88]

In a statement to the Commons on 12 January 2026, the government said that it supports Ofcom to use its "full powers" in its investigation, including applying for a court order to stop UK users accessing X if the platform does not act to prevent illegal content being shared.[89] The government also said that the criminal offence of creating or requesting the creation of non-consensual intimate images would be brought into force "this week":

> The Data (Use and Access) Act 2025 made it a criminal offence to create or request the creation of non-consensual intimate images, and today I can announce to the House that this offence will be brought into force this week and that I will make it a priority offence in the Online Safety Act, too. That means that individuals are committing a criminal offence if they create or seek to create such content, including on X, and anyone who does that should expect to face the full extent of the law.[90]

The government added that it would legislate in the Crime and Policing Bill to "criminalise nudification apps", making it a criminal offence for companies to

---

84 Safety (@Safety), X (Twitter), 4 January 2026 [accessed 14 January 2026]

85 X (formerly Twitter), The X Rules, accessed 14 January 2026; X (formerly Twitter), Our range of enforcement options, accessed 14 January 2026

86 Elon Musk (@elonmusk), X (Twitter), 3 January 2026 [accessed 14 January 2026]

87 The Guardian, Grok turns off image generator for most users after outcry over sexualised AI imagery, 9 January 2026

88 BBC News, X to stop Grok AI from undressing images of real people after backlash, 15 January 2026

89 HC Deb 12 January 2026 c607

90 HC Deb 12 January 2026 c607

"supply tools designed to create non-consensual intimate images".[91] In a letter to the chair of the Science, Innovation and Technology Committee (PDF), the government stated that apps whose "one despicable purpose" is to use AI to create fake nude photos or images would be banned; media reports noted that the ban may not therefore apply to Grok, because it is a "general-purpose" AI tool and is not designed specifically as a nudification app.[92]

The US State Department said that "nothing is off the table" in how it might respond if access to X in the UK were blocked following the conclusion of Ofcom's investigation.[93]

YouTube

YouTube's guidance on "disclosing use of altered or synthetic content" states that the platform requires video creators to "disclose meaningfully altered or synthetically generated content that seems realistic".[94]

Users can proactively disclose the use of AI-generated content during the video-upload process on YouTube by selecting the "Altered Content" option, including for content created within the platform's YouTube Studio workflow. By contrast, YouTube's guidance notes that videos featuring AI-generated content that are created through the platform's YouTube Shorts tool will be labelled automatically.

Additional guidance from YouTube notes that the platform relies on three types of information for labelling AI-generated content:

- A user manually discloses the use of AI-generated content within the YouTube Studio workflow. The guidance notes that when a user does not disclose AI-generated content, "in some cases, YouTube may take action to reduce the risk of harm to viewers by proactively applying a label that creators will not have the option to remove".[95]
- A user creates content using YouTube's generative AI tools. In this case, YouTube will automatically apply a label.
- A user uploads content that contains C2PA "content credential" metadata. In this case, YouTube will automatically apply a label.

For most AI-generated content on YouTube, a label in the video's expanded description field will tell viewers that it was made from "altered or synthetic content". Viewers must click on the description field to expand it to make the label visible. However, for "sensitive content", a more prominent label is displayed in the video player. YouTube defines "sensitive content" as videos

91 HC Deb 12 January 2026 c607

92 Department for Science, Innovation and Technology, Intimate Grok AI Deepfakes (PDF), 12 January 2026; POLITICO, UK nudification app ban won't apply to Elon Musk's Grok, 14 January 2026

93 POLITICO, US State Department threatens UK over probe into Elon Musk's X, 13 January 2026

94 YouTube, Disclosing use of altered or synthetic content, not dated

95 YouTube, Understanding 'How this content was made' disclosures on YouTube, not dated

that contain "information about elections, ongoing conflicts, natural disasters, finance, or health" that "can greatly affect the well-being, financial security, or safety of people and communities".[96] Example screenshots of these labels are visible in YouTube's guidance on disclosing altered or synthetic content.

The guidance notes that when a user discloses the use of AI-generated content in a video, it "won't limit a video's audience or impact its eligibility to earn money".[97]

4.3 What are the labelling policies of news organisations?

In the UK, there are different regulators for television news broadcasters and print and online news publications.

Ofcom regulates the communications industry, including news broadcasting organisations.[98] There are two regulators for print and online news publications: the Independent Press Standards Organisation (IPSO) and Impress. Unlike Ofcom, which has a statutory duty to regulate news broadcasting organisations, membership of both IPSO and Impress is voluntary and not all news publications are members of them. The Commons Library briefing How is the press regulated? (July 2024) provides more detail on how news publications are regulated.[99]

Ofcom's approach to the use of AI in the industries that it regulates, including news broadcasting organisations, is discussed in section 4.2 above.

In January 2026, IPSO published guidance on the use and labelling of AI-generated content for the news publications it regulates. The guidance emphasises news organisations' "editorial responsibility" in using and labelling AI-generated content and says that human editorial oversight should be used to maintain accuracy when publishing AI-generated content:

> Human editorial oversight upholds journalistic standards, even as large language models and other AI tools become more prevalent.
>
> Those regulated by IPSO are accountable for their content and must satisfy themselves that they are taking care over accuracy.[100]

96 YouTube, Disclosing use of altered or synthetic content, not dated
97 YouTube, Disclosing use of altered or synthetic content, not dated
98 House of Lords Library, Regulation of news broadcasting companies, 11 March 2024
99 House of Commons Library, How is the press regulated?, 16 July 2024
100 Independent Press Standards Organisation (IPSO), Artificial intelligence, accessed 19 January 2026

The guidance builds on IPSO's response to an independent review conducted by Sir Bill Jeffrey (PDF), in which IPSO said that editorial responsibility should apply to published material created with generative AI.[101]

In April 2025, Impress issued guidance on the use and labelling of AI-generated content by news publications (PDF). The guidance recommends that publishers should consider disclosing the use of AI and prominently displaying their AI policies:

> Publishers using AI tools in their newsrooms should consider disclosing their use of the technology by posting terms and conditions on the front page of their website or publication, with a link to any specific AI policies that detail practices and procedures about the use of AI in their organisation.
>
> [...]
>
> Publishers should attribute their use of AI in content generation of all types unless there is good reason not to do so. The attribution could be in the form of an AI logo, AI assisted reporter byline or other appropriate labelling method.[102]

The guidance also recommends that publishers should disclose how they use AI and any associated potential risks:

> Publishers should disclose how they use AI tool(s) in their news reporting, including the purpose and scope of the tool(s) and any limitations and risks which they may carry. This disclosure should be concise and primarily address the impact which AI tool(s) have on the publisher's newsgathering process, editorial decision-making and published content.[103]

Impress does not specify a wording that should be used to disclose the use of AI-generated content.

BBC News

The BBC has published editorial guidance on the use of AI, which states that generative AI should not be used to create news content unless it is the subject of such news content:

> Generative AI should not be used to directly create news content published or broadcast by BBC News/Nations, current affairs or factual journalism unless it is the subject of the content and its use is illustrative.[104]

The BBC has also published guidance on AI transparency. The guidance states that the use of AI can either be disclosed directly ("labelling or otherwise revealing the use of AI at the point of consumption") or indirectly ("labelling or other interventions not present at the point of consumption").

---

101 Independent Press Standards Organisation (IPSO), Implementation update for the independent external review of IPSO by Sir Bill Jeffrey, December 2024, p.14

102 Impress, Best Practice Note: The Use of Artificial Intelligence (PDF), April 2025, p.6

103 Impress, Best Practice Note: The Use of Artificial Intelligence (PDF), April 2025, p.10

104 BBC, Editorial guidance on the use of AI, updated 24 June 2025

The guidance includes a "disclosure formula" that should be used when indicating that content is AI-generated.[105]

Channel 4 News

Channel 4 has published guidance on its "AI principles" (PDF). This guidance states that Channel 4 will "be transparent" about its use of AI and "will keep people informed of how we use AI and [...] will always feel comfortable explaining our use of AI to others".[106] The guidance does not specify how Channel 4 labels AI-generated content.

The guidance also contains a list of ways in which Channel 4 is currently using generative AI. Only one of these relates to the use of AI-generated content in its programming:

> 6. Trialling, transparently, the use of AI in our programming to disguise anonymous interviews, reconstruct events and at the centre of new programme formats like The Honesty Box.[107]

In October 2025, Channel 4's Dispatches aired "Will AI Take My Job?". The programme featured what the broadcaster said was the UK's first "entirely AI-generated" TV presenter.[108]

The Daily Telegraph and the Sunday Telegraph

The Telegraph's editorial and commercial guidelines do not mention AI or AI-generated content.[109] However, the guidelines state that the Telegraph adheres to IPSO's Editors' Code of Practice, which "stipulates that the press must take care not to publish inaccurate, misleading or distorted information".[110]

In November 2024, the Press Gazette reported on the Telegraph's uses of AI-generated content. The Telegraph has begun using AI to generate summaries for some email newsletters and some articles. While the AI-generated summaries of articles are accompanied by a disclaimer, it is not clear whether email newsletters come with a similar disclosure.[111]

GB News

GB News's Artificial Intelligence Policy (updated 24 July 2025) outlines how GB News uses AI. The "Accountability and Transparency" section of this policy

---

105 BBC, BBC AI Transparency, updated 16 January 2025

106 Channel 4, Channel 4's AI Principles (PDF), May 2025, p.2

107 Channel 4, Channel 4's AI Principles (PDF), May 2025, p.3

108 Channel 4, Channel 4 makes TV history with Britain's first AI presenter, 20 October 2025

109 The Telegraph, Editorial and Commercial Guidelines, not dated

110 The Telegraph, Editorial and Commercial Guidelines, not dated; IPSO, The Editors' Code of Practice (2025 edition), 1 January 2025

111 Press Gazette, Telegraph is launching an AI-driven newsroom tool every month, 14 November 2024

states that GB news will disclose where AI has been used "in a substantial way for broadcast content":

> GB News is committed to full accountability regarding our use of AI. Experienced editing staff are responsible for the publication of all of our output, including any which has been assisted using artificial intelligence. Where AI has been used in a substantial way for broadcast content we will include a verbal or visual disclosure, as appropriate to the medium, to enhance transparency.[112]

The policy does not specify the wording of disclosures that AI-generated content has been used.

The Guardian

Section H of the Guardian's Editorial Code (PDF, updated March 2025) states that AI should not be used to generate content for published journalism outside "exceptional and specific circumstances":

> AI systems should not be used to generate text or images intended to be directly inserted into published journalism outside of exceptional and specific circumstances. Any exceptional use must be explicitly approved by the relevant senior duty editor and must be clearly signalled to readers on the article itself.[113]

The Guardian's Editorial Code does not specify the wording of disclosures that AI-generated content has been used.

ITV News

ITV has not published publicly available guidance on its use and labelling of AI-generated content.

In October 2024, ITV was criticised for posting a job advert for a "head of generative AI innovation" to implement "AI-driven innovations in content creation". An ITV spokesperson noted in response that ITV was using AI tools to "optimise our content for viewers":

> At ITV we are using AI tools to enhance and expand our creative and production processes. While nothing can replace the human creativity of our teams, we are exploring how GenAI can help our staff to work more efficiently and creatively and optimise our content for viewers.[114]

112 GB News, Artificial Intelligence Policy, updated 24 July 2025
113 The Guardian, Editorial Code (PDF), updated March 2025
114 BBC News, Derry Girls creator hits out at ITV over AI plans, 21 October 2024

Reuters

Reuters has published guidance on its use of AI, which states that its journalists use AI as a tool in their work, and that it discloses when news content has been produced primarily or solely using AI:

> Reuters journalists at times use generative Artificial Intelligence as one of the tools in our reporting, writing, editing, production and publishing. When we rely primarily or solely on generative AI to produce news content, we clearly disclose this use, providing context about how we have used the tool.[115]

The guidance does not specify the wording of disclosures that AI-generated content has been used.

Sky News

Neither Sky News's policies and standards guidance nor its editorial guidelines mention AI or AI-generated content.[116] Sky News has not published publicly available guidance on its use and labelling of AI-generated content.

The Times and the Sunday Times

The Times has published editorial guidelines on its use of generative artificial intelligence. Section 4 of the guidance outlines the "currently very limited" ways in which the Times uses AI-generated content, including suggesting headlines and summaries for articles that are then considered by an editor.[117]

Section 4 of the guidance also states that the Times is committed to being "transparent in how we are using Gen AI in order to ensure that we can be held to the highest ethical standards".[118]

The guidance does not specify the wording of disclosures that AI-generated content has been used.

4.4 What are the labelling policies of search engines?

Ofcom is the regulator of search engine services in the UK for the purposes of the Online Safety Act 2023, which aims to protect users of online services from

115 Reuters, Reuters and AI, not dated

116 Sky News, Sky News policies and standards, updated 3 March 2025; Sky News, Editorial Guidelines (PDF), October 2025

117 The Times, Generative artificial intelligence: The Times and Sunday Times editorial guidelines, not dated

118 The Times, Generative artificial intelligence: The Times and Sunday Times editorial guidelines, not dated

illegal or harmful content and activity.[119] However, Ofcom has not issued any specific guidance to search engines about how AI-generated content should be labelled in search results.

In November 2024, Ofcom issued an open letter to online service providers about the use of generative AI and chatbots. These included generative AI tools "that enable the search of more than one website and/or database", which constitute regulated "search services" under the Online Safety Act.[120] These "answer engines" differ from traditional search engines in that they generate answers from a variety of sources in response to a user's prompt question.[121]

Bing

Bing is a search engine owned and operated by Microsoft. Bing introduced generative AI search results in July 2024. In October 2024, Bing published guidance on its generative search results labelling, which states that such results will be indicated with a label:

> You will generally see generative search results for informational and complex queries, and it will be indicated under the search box with the sentence 'Results enhanced with Bing generative search'.[122]

There is no publicly available guidance from Bing that clarifies its position on the labelling of AI-generated content within its general search results.

In April 2023, Microsoft updated its guidance on the responsible use of AI in Bing (PDF). This notes that Bing has several different ways to tell users that they are interacting with AI systems:

> The new Bing provides several touchpoints for meaningful AI disclosure where users are notified that they are interacting with an AI system as well as opportunities to learn more about the new Bing.[123]

This reflects Microsoft's Responsible AI Standard, which states that Microsoft AI systems are designed to inform people that they are interacting with or viewing content powered or created by AI:

> Microsoft AI systems are designed to inform people that they are interacting with an AI system or are using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic.[124]

---

119 Ofcom, Online safety – what is Ofcom's role, and what does it mean for you?, 9 November 2023. See also Ofcom, Ofcom's access to services regulated under the Online Safety Act, updated 29 July 2025

120 Ofcom, Open letter to UK online service providers regarding Generative AI and chatbots, 8 November 2024

121 Ofcom, The Era of Answer Engines: Generative AI's impact on search experiences and online safety, 4 November 2025

122 Bing, The next step in Bing generative search, October 2024

123 Microsoft, The new Bing: Our approach to Responsible AI (PDF), updated April 2023, p.8

124 Microsoft, Microsoft Responsible AI Standard, v2: General requirements (PDF), June 2022, p.12

Neither Bing nor Microsoft specify the wording for disclosures that AI-generated content has been included in results for searches made through Bing.

Google

The FAQ section of Google's guidance on AI-generated content notes that using AI appropriately is not against its guidelines:

> Appropriate use of AI or automation is not against our guidelines. This means that it is not used to generate content primarily to manipulate search rankings, which is against our spam policies.[125]

The FAQ section also notes that disclosures are useful where someone might question how something was created:

> AI or automation disclosures are useful for content where someone might think 'How was this created?'. Consider adding these when it would be reasonably expected.[126]

The guidance says that Google's search rankings are based on "the quality of the content, rather than how it is produced". It recommends that the builders of websites ensure that their content demonstrates "expertise, experience, authoritativeness, and trustworthiness" (E-E-A-T) to ensure that it places highly in Google's search ranking system.[127]

Google's guidance on the use of generative AI on websites recommends that the builders of websites listed on Google disclose information about how AI was used during a website's construction:

> consider adding information on how your content was created in a way that makes sense for your audience, such as by providing more background information on how automation was used and adding image metadata.[128]

This is not a requirement, and Google does not specify the wording for disclosures that AI-generated content has been included in results for searches made through Google.

Google has separate policies for AI-generated content on e-commerce websites, which require that some product data is identified as AI-generated. Google's guidance for e-commerce sites says that AI-generated images must contain metadata that show it was created using AI:

125 Google, Google Search's guidance about AI-generated content, February 2023
126 Google, Google Search's guidance about AI-generated content, February 2023
127 Google, Google Search's guidance about AI-generated content, February 2023
128 Google, Google Search's guidance on using generative AI content on your website, updated 10 December 2025

> All images created using generative AI must contain metadata indicating that the image was AI-generated by using the IPTC DigitalSourceType TrainedAlgorithmicMedia metadata tag.[129]

Removal of Google AI-generated health overviews (January 2026)

In January 2026, the Guardian reported that Google's AI overviews, which use generative AI to provide summaries of information in response to a user's search, were in some cases providing "false and misleading health information".[130] Medical experts consulted by the Guardian said that information provided in some AI overviews, including in response to searches about liver disease and pancreatic cancer, "put people at risk of harm".[131]

Google subsequently removed some AI-generated health overviews. However, the Guardian noted that some overviews that included misleading content, which had previously been raised with Google, still appeared in search results.[132]

In a statement shared with the technology news site Gizmodo, Google said that the removal of some AI-generated health overviews was the result of a review by its internal clinicians:

> Our internal team of clinicians reviewed what's been shared with us and found that in many instances, the information was not inaccurate and was also supported by high quality websites. In cases where AI Overviews miss some context, we work to make broad improvements, and we also take action under our policies where appropriate.[133]

In a separate statement shared with the Guardian, Google said that its AI overviews "only show up on queries where it has a high confidence in the quality of the responses".[134]

---

129 Google Merchant Centre, AI-generated content, not dated

130 The Guardian, Google AI Overviews put people at risk of harm with misleading health advice, 2 January 2026

131 The Guardian, Google AI Overviews put people at risk of harm with misleading health advice, 2 January 2026

132 The Guardian, 'Dangerous and alarming': Google removes some of its AI summaries after users' health put at risk, 11 January 2026

133 Gizmodo, Google Pulls Some Health-Related AI Overviews as Industry Goes All-In on Healthcare, 13 January 2026

134 The Guardian, 'Dangerous and alarming': Google removes some of its AI summaries after users' health put at risk, 11 January 2026

Yahoo

Section 6(b) of Yahoo's terms of service outlines the search engine's use of AI and third-party AI providers. The terms of service note that AI content or responses may contain inaccuracies and should be independently verified:

> In some instances, use of AI query features may be governed by the AI Provider's terms of service and privacy policy. You understand and agree that content or responses generated by AI may contain inaccuracies and should never be relied upon without independent verification. Yahoo does not control the content or responses provided by AI Providers, and makes no representations or warranties about the accuracy or completeness of such content or responses (or the sites and sources accessed through such content or responses).[135]

There is no publicly available guidance from Yahoo that clarifies its position on the labelling of AI-generated content within its general search results.

4.5 What are the labelling policies of video game services?

Ofcom also regulates online video game services in the UK for the purposes of the Online Safety Act 2023, including "user-to-user" services that enable users to "generate, share or upload content [...] on the service that may be encountered by other users of the service".[136] However, Ofcom has not issued specific guidance on the labelling of AI-generated content on gaming platforms or online video game services.

Fortnite (Epic Games)

Fortnite is an online "battle royale" survival game in which players compete against up to 100 other players to be the last player standing. Fortnite is developed by Epic Games.

Fortnite's developer rules, community rules and content rules do not mention the use of AI-generated content in Fortnite or in any other game developed by Epic Games.

However, sections 1.6 to 1.8 of the developer rules state that players must be original in the content they create in the game, and that they must include licensing or copyright information in the "attributions" of the Creator Portal where necessary.[137] The community rules tell players that they cannot impersonate anyone else and the content rules prohibit the creation,

---

135 Yahoo, Terms of service, updated 27 October 2025

136 Ofcom, Gaming: know the online safety risks, the rules, and how to comply, updated 11 December 2025

137 Epic Games, Fortnite Developer Rules, updated 11 December 2025

publishing and use of anything that infringes on others' rights, including their intellectual property rights.[138]

In April 2025, the gaming and technology website WCCF Tech reported that Fortnite would not stop players from using AI-generated content in their thumbnails on the platform unless it violates other platform rules. Summarising comments from Epic Games executives in an online interview, the article states that Epic Games will not use generative AI in its own creations:

> ...Epic itself wouldn't be using generative AI in its own creations such as Fortnite skins, but that when it comes to user's content, Epics' concerns are more geared towards whether the content is 'compliant with [its] rules,' rather than being concerned with how the users made that content.[139]

In June 2025, Fortnite announced that players will be able to use generative AI to create non-playable characters (NPCs), who can engage in conversations based on text prompts from players.[140] Fortnite has published guidance on using AI to create NPCs.[141]

In November 2025, Fortnite launched Chapter 7, its latest "season" of the game. Players criticised what they thought was the inclusion of AI-generated content in Fortnite Chapter 7, although Epic Games did not confirm the use of AI-generated content in the game.[142]

At the same time, players noted comments made by Tim Sweeney, the CEO of Epic Games, in which he criticised the games storefront Steam for mandating that developers disclose the use of AI-generated content in their games.[143] Steam requires that any games sold through its platform disclose the use of AI in "pre-generated" content (created with the help of AI tools during development) or "live-generated" content (created with the help of AI tools while the game is running).[144] Epic Games hosts its own games storefront and does not require similar disclosures from games developers.

Minecraft (Microsoft)

Minecraft is a "sandbox" game in which players create and explore a world made up of fully customisable "blocks".

138 Epic Games, Epic Games Community Rules, updated 12 September 2024; Epic Games, Epic Games Content Guidelines, updated 12 September 2024

139 WCCF Tech, Epic Games Won't Stop AI Generated Thumbnails On Fortnite, Says Gen AI Will Become "More And More Difficult To Detect", April 2025

140 bo3.gg, Epic Games Introduces Generative AI NPCs for Fortnite, 5 June 2025

141 Epic Games, AI and NPCs, not dated

142 Metro, Fortnite hit with 'AI slop' complaints after Epic boss says to remove warnings, updated 1 December 2025

143 Futurism, CEO of Fortnite Maker Furious That Steam Is Labeling Games With AI-Generated Assets, 1 December 2025

144 Steam, AI Content on Steam, not dated

Mojang Studios and Xbox Game Studios, which developed Minecraft, are owned by Microsoft. Section 8 of Xbox's community standards, which apply to Minecraft, prohibits players from infringing others' rights, including intellectual property rights, but does not specifically mention AI-generated content.[145]

Neither Minecraft's end-user licence agreement (EULA) nor its usage guidelines mention AI-generated content.[146] Both policies permit players to create and distribute their own modifications ('mods') to the game based on its Minecraft: Java Edition, including player-developed tools and plug-ins. However, players cannot create or distribute a 'modded version' of the game that contains "a substantial part of our copyrightable code or content", or any other "hacked" versions of the game.[147] There are a number of AI agents that enable players to create their own mods for Minecraft. The AI visual content creation platform Apatero has published an explainer on the use of AI in Minecraft.[148]

Roblox

Roblox is an online game platform that enables players to develop and play games created by themselves or other players. Each player creates their own personalised avatar.

In March 2025, Roblox announced the introduction of Roblox Cube, a generative AI system enabling players to create 3D models and environments from text inputs.[149] An article published by the legal consultancy firm Bristows notes that "it would seem that Roblox has not implemented any specific safety or other controls around the potential use of prompts for generating infringing, inappropriate or illegal content" in Roblox Cube.[150]

In September 2025, Roblox announced the development of a "photo-to-avatar technology" that will use AI to generate a player's in-game avatar from a selfie. The announcement notes that this feature is currently limited to users who are aged 13 and over, but it does not specify whether avatars generated from selfies will need to be labelled as AI-generated content.[151] Roblox has also published additional guidance on the creation and use of avatars.[152]

Roblox's AI-Based Tools Supplemental Terms and Disclaimer outlines acceptable uses of the generative AI-based tools available on the platform.

---

145 Microsoft/Xbox, Xbox Community Standards, updated May 2024

146 Minecraft, Minecraft End(er)-User License Agreement ("EULA"), not dated; Minecraft, Minecraft Usage Guidelines, not dated

147 Minecraft, Minecraft End(er)-User License Agreement ("EULA"), not dated

148 Apatero, Generate 3D Minecraft Structures with AI Prompts: Complete Guide 2025, 13 November 2025

149 Roblox, Introducing Roblox Cube: Our Core Generative AI System for 3D and 4D, 17 March 2025

150 Bristows, Not Child's Play: AI and online safety for Roblox, 8 April 2025

151 Roblox Creator Hub, Early Preview: Alpha Release of Photo-to-Avatar APIs, updated 25 September 2025

152 Roblox, Create avatar items, not dated

Section 2(e) of the terms obliges users not to "represent that the Output was human-generated when it was not". However, the terms do not require users to label AI-generated content or disclose the use of generative AI tools.[153]

Roblox has published guidance on how its uses AI large-language models (LLMs) to moderate AI-generated text on the platform and how it uses AI to moderate content more generally.[154]

[153] Roblox, AI-Based Tools Supplemental Terms and Disclaimer, not dated

[154] Roblox, State-of-the-Art LLM Helps Safeguard Unlimited Text Generation on Roblox, 22 July 2025; Roblox, How Roblox Uses AI to Moderate Content on a Massive Scale, 9 July 2025

The House of Commons Library is a research and information service based in the UK Parliament. Our impartial analysis, statistical research and resources help MPs and their staff scrutinise legislation, develop policy, and support constituents.

Our published material is available to everyone on commonslibrary.parliament.uk.

Get our latest research delivered straight to your inbox. Subscribe at commonslibrary.parliament.uk/subscribe or scan the code below:

commonslibrary.parliament.uk

@commonslibrary