# Deepfake Defences 2

## The Attribution Toolkit

**Report**

Published 11 July 2025

# Contents

## Section

## Annex

# Overview

## The deepfakes challenge

Deepfakes are AI-generated audio-visual content that are deliberately designed to misrepresent someone or something. They pose a significant threat to online safety and can cause harm in myriad ways. The last year has seen numerous incidents where deepfakes have promoted financial scams, depicted people in non-consensual sexual imagery, and targeted politicians. From the Facebook adverts that falsely showed the Financial Times journalist Martin Wolf promoting a scam investment group, to the deepfake intimate images of Taylor Swift that were shared over X and Telegram, it is not hard to find examples of the real-world damage that deepfakes can inflict in the UK and around the world.

This is not just a problem for high profile individuals. Ordinary members of the public can just as easily be the victims of a deepfake. A survey undertaken by the Alan Turing Institute in 2024 found that 15 percent of UK adults have been exposed to harmful deepfakes, while 90% are either very or somewhat concerned about this issue. Children are also being targeted, sometimes by their own peers. A study undertaken by Internet Matters revealed that 10 percent of children aged 13-16 had either directly experienced, or knew of someone who had experienced, being featured in fake nude images or videos.[1]

One reason for the increase in prevalence of deepfakes online is the advent of easily accessible generative AI models ('GenAI'). These have enabled anyone with basic technical literacy to create sophisticated and convincing deepfakes. Moreover, many of these models have been made available on an open-source basis, allowing bad actors to 'fine tune' the technology for the specific purpose of creating particular types of deepfake. Another important trend is the emergence of what might be called a 'deepfake economy', made up of professional creators offering to generate deepfakes for a fee, as well as websites dedicated to hosting such content.

Deepfakes are relevant to Ofcom's work because the sharing of certain types of deepfake is regulated under the Online Safety Act 2023 ('the Act'). Regulated user-to-user and search services are required to assess the risk of harm to users that is posed by illegal content or content that is harmful to children on their services.[2] This could include some forms of deepfake (for example, deepfake fraud content and deepfake child sexual abuse material). Regulated services also have a duty to take steps to prevent or minimise the risk of their users encountering this kind of content.[3]

## Deepening our deepfake defences

Against this backdrop, Ofcom has set out to better understand what is driving the growth of deepfakes online and to identify what can be done to stop their spread. Last year, we published a discussion paper called Deepfake Defences that summarised the results of our first phase of research on this topic.

[1] It is important to note that all nude deepfakes of children are illegal and classified as child sexual abuse material (CSAM).

[2] The illegal content and children's risk assessment duties are set out sections 9 and 11 of the Act for user-to-user services and sections 26 and 28 of the Act for search services.

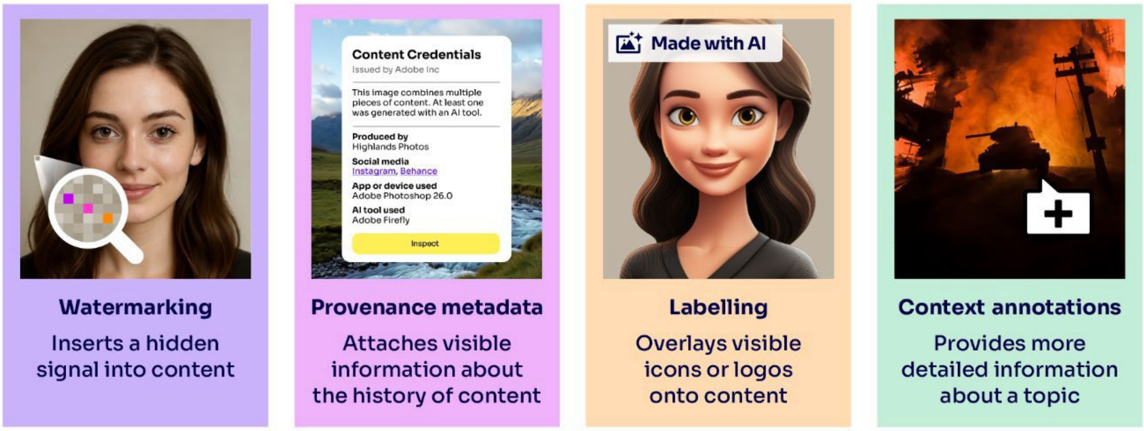
[3] The safety duties about illegal content and the safety duties protecting children are set out in sections 10 and 12 of the Act for user-to-user services, and sections 27 and 29 of the Act for search services.

The paper introduced a three-part typology of deepfakes – those that demean, defraud and disinform – and set out (at a high level) a range of interventions that different actors in the AI supply chain could take to address the sharing of this type of harmful content. Among the measures we identified were:

- Designing prompt filters to prevent AI models being instructed to create deepfakes
- Using red team methods to evaluate the likelihood of a model creating deepfakes
- Embedding signals into content via watermarks and provenance metadata schemes to indicate that content is synthetic
- Deploying machine learning classifiers to identify deepfakes
- Suspending or removing the accounts of users who repeatedly create and share deepfakes
- Establishing an easy way for users to report deepfakes where they encounter them

In this follow up discussion paper, we look more closely at the merits of so-called **'attribution measures'**, which include watermarking, provenance metadata schemes, AI labels and context annotations.[4] These are all designed in one way or another to attribute certain types of information to a piece of content, for example; information about who created it, how and when it was created, and – in some cases – whether the content is accurate or misleading.

**Figure 1: The four attribution measures**


Watermarking
Inserts a hidden signal into content
Provenance metadata
Attaches visible information about the history of content
Made with AI
Labelling
Overlays visible icons or logos onto content
Context annotations
Provides more detailed information about a topic

Content Credentials
Issued by Adobe Inc
This image combines multiple pieces of content. At least one was generated with an AI tool.
Produced by
Highlands Photos
Social media
Instagram, Behance
App or device used
Adobe Photoshop 26.0
AI tool used
Adobe Firefly
Inspect

Attribution measures have garnered increasing interest from policymakers, industry and civil society groups, and many organisations are now making use of them in a bid to tackle deepfakes. This includes Google DeepMind and Meta (which have recently released tools for watermarking content), Adobe (which has been a leading player in establishing the Coalition for Content Provenance and Authenticity's 'C2PA' specification), TikTok and YouTube (which have begun to apply AI labels to content shared on their platforms), and X (which set up one of the first context annotation initiatives on a social media platform).

However, despite this increase in interest, there is still much that we don't know about attribution measures, including how they function, their strengths and weaknesses, and what it would take to deploy them successfully. This discussion paper aims to provide more substantial answers to these questions and improve our overall understanding of attribution measures. It draws on the findings of

[4] In our Deepfake Defences discussion paper, we referred to these as 'embedding' measures. Following our second phase of research, we decided that 'attribution' was a more appropriate way to describe the collective purpose of these measures: https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/deepfake-defences

several research activities, including a review of existing literature, interviews with experts, a survey and a series of interviews with users of online platforms, and our own internal technical evaluation of openly available watermarking tools. See **Box 1** for more details on our research methods.

Below we summarise our findings, explaining first how the attribution measures work in practice, before setting out 8 key takeaways that should guide future action by industry, government and researchers in this area.

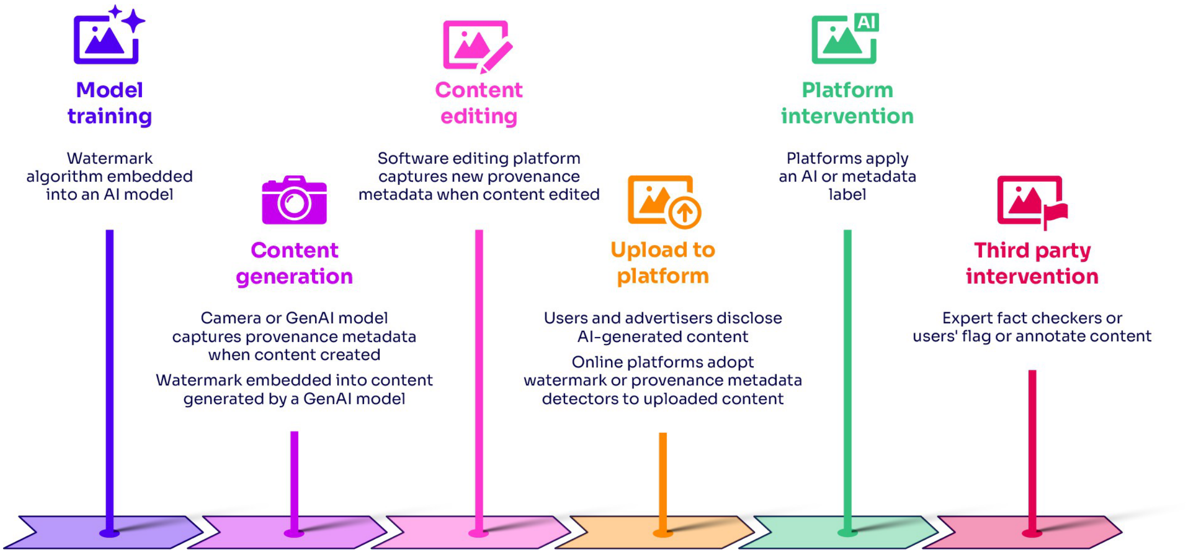
Please be aware that this discussion paper does not constitute formal guidance. Regulated services under the Online Safety Act are not required to adopt any of the measures featured in this paper. It is intended only to shine a light on emerging issues and best practices.

## What do attribution measures involve?

Our research shows that attribution measures can be deployed in numerous ways, with new techniques and practices continuing to emerge. However, the four measures can be broadly summarised as follows:

- **Watermarking** – Watermarking refers to techniques that embed an invisible signal into content. For images, this may involve subtly altering the colour of specific pixels, while for audio, it can mean changing elements of its frequency. Typically, these signals convey information about the AI model used to create the content. The signals can be identified at a later stage by a different party (e.g. a social media platform) using a special detection algorithm. Platforms may also apply a label to this content.
- **Provenance metadata** – Provenance metadata is commonly associated with watermarking; however provenance metadata is not embedded within content but rather attached to it. Another difference is that metadata usually carries more information, such as when the content was created and by whom. Provenance metadata can be updated every time a piece of content is edited (e.g. metadata could be attached to a photo at the point it is taken, and then updated when that photo is subsequently altered using image editing software) and can be shown in a label.
- **AI labels** – AI labels apply a visible and recognisable icon to content that signals something about it, for example whether it is synthetic. Some platforms ask users to voluntarily disclose when they are uploading AI-generated or edited content, at which point a label is automatically applied to it. Alternatively, platforms can detect and label AI content using their own moderation tools (which may involve scanning for watermarks and provenance metadata).
- **Context annotations** – Context annotations provide users with additional information about a piece of content. They usually take the form of a short note that is visible below a piece of content. Different actors can be involved in annotating content, from independent fact-checking organisations to 'superuser' experts and everyday users. Platforms often set thresholds for when annotations appear publicly on content (e.g. they may first require a given number or diversity of annotators to vote that the note is 'helpful').

**Figure 2: The lifecycle of attribution measures**



Model training
Watermark algorithm embedded into an AI model
Content generation
Camera or GenAI model captures provenance metadata when content created
Watermark embedded into content generated by a GenAI model
Content editing
Software editing platform captures new provenance metadata when content edited
Upload to platform
Users and advertisers disclose AI-generated content
Online platforms adopt watermark or provenance metadata detectors to uploaded content
Platform intervention
AI
Platforms apply an AI or metadata label
Third party intervention
Expert fact checkers or users' flag or annotate content

## 8 key takeaways for attribution measures

While each attribution measure is unique, with its own merits and limitations, our analysis reveals several overarching findings that can inform how we think about and make use of these tools and techniques:

1. **Evidence shows that attribution measures can help users to engage with content more critically.** Research shows that, when deployed with care and proper testing, attribution measures can improve the capacity of users to spot misleading content, and in some cases can discourage them from sharing that content onwards. Moreover, our qualitative research shows that many users would welcome the roll out of attribution measures to help them make sense of content.

2. **Users should not be left to identify deepfakes on their own.** While attribution measures like AI labels and context annotations can help to empower users, platforms should avoid placing the full burden on individuals to detect misleading content. Some of the information captured by these tools – such as provenance metadata – could be used by platforms to inform their own moderation efforts, including which content to prioritise for moderation or review.

3. **Striking the right balance between simplicity and detail is crucial when communicating information to users.** Too much information can be overwhelming, but too little can result in confusion. With AI labels, for example, there is a danger that these simple icons are misinterpreted by users as meaning that the content is untrustworthy. These effects can frustrate content creators who post legitimate material. It is also possible that if labelling becomes common, the *absence* of a label on a piece of content could lend that content undue legitimacy (though such claims need further investigation).

4. **Attribution measures need to accommodate content that is neither wholly real nor entirely synthetic.** AI models are increasingly being used to augment existing content (e.g. via AI photo filters), not just to create new content from scratch. Attribution measures will be more effective where they can convey this nuance to users, communicating *how* AI has been used, not just *whether* it has been used.

5. **Attribution measures can be susceptible to removal and manipulation.** Our own technical evaluations of publicly available watermarks revealed that they could be removed from image

content following certain types of edits, such as cropping. There is also a risk that bad actors manipulate attribution measures, for example using context annotations to spread disinformation.

6. **Greater standardisation could boost the efficacy and adoption of attribution measures.** Model developers, platforms and others often use different approaches to apply attribution measures, which can increase costs and create confusion. For example, each AI developer has its own method for embedding watermarks, which forces downstream platforms to deploy multiple detection algorithms to pick up on these signals. A degree of standardisation could lower costs and facilitate the take up of attribution measures.
7. **The pace of change means it would be unwise to make sweeping claims about attribution measures.** Many of the measures and related initiatives explored in this paper are relatively new. Platforms have only begun to use AI labels in the last two years, and model developers have only recently started watermarking their content. With so much change afoot, it is important to avoid making sweeping judgements about the merits of particular measures.
8. **Attribution measures should be used in combination to tackle the greatest range of deepfakes.** Rather than view measures like watermarking and metadata as substitutes, we should instead see them as complementary safeguards, whose collective deployment increases the probability of curtailing the spread of deepfakes. AI developers and platforms should also remember there are other actions they can take besides introducing attribution measures, for instance using AI classifiers to detect deepfakes and enabling users to easily report this content.

## Next steps

We hope the findings set out in this paper will have immediate value to those who are researching, building or deploying new methods for tackling deepfakes. Ofcom will be drawing on these insights to:

- Inform our upcoming consultation on our draft Code of Practice on fraudulent advertising – an area where deepfakes are known to be an issue.
- Continue raising awareness among regulated services of their duties to tackle deepfakes that are illegal or harmful to children.
- Explore enforcement action where there is evidence that services are not doing enough to comply with their duties in this regard.
- Conduct further research on best practice solutions for addressing deepfakes, which may include an investigation of machine learning-based deepfake detection classifiers.
- Exchange lessons with other regulators, including via the Digital Regulation Cooperation Forum (see for example our joint horizon scanning work on the future of synthetic media).
- Support the UK government as it draws up new offences to tackle certain forms of deepfake, for example AI-generated CSAM material.

We would welcome feedback on this paper. Contact our Technology Policy team at technologypolicy@ofcom.org.uk.

### Box 1: The research methods that informed this paper

**Literature review** – We reviewed 205 publicly available research papers that discussed watermarks, provenance metadata, AI labels and context annotations.

**Expert engagement** – We followed up by interviewing ten experts, including from academic institutions in the UK and US, civil society groups, an online platform and an industry body.

**User research** – We commissioned YouGov to undertake an online survey on AI labels in December 2024. We asked 2,143 internet users aged 16+ in a UK representative sample about their exposure to, understanding of, and attitudes towards, AI labels. In February 2025, we commissioned YouGov Qualitative to undertake 24 one-hour online depth interviews with internet users aged 18+. In each interview, we presented participants with image and video-based scenarios of deepfake content that included three different visible labels (a provenance metadata label, an AI label, and a context annotation label). Participants were asked to reflect on each scenario, describing what they noticed, what they thought the label meant, and whether it changed their perception of the content presented to them. For further detail on the method and results, please see the survey data tables, qualitative research report, and technical report on both qualitative and quantitative research that are being published alongside this paper.

**Technical evaluation** – Ofcom's Online Safety Technology team evaluated three publicly available watermarking tools against common benchmarks. This involved embedding a watermark into images, and testing whether those watermarks remained detectable when those images were edited. See **Annex 1** for further detail on the method and results.

# Watermarking

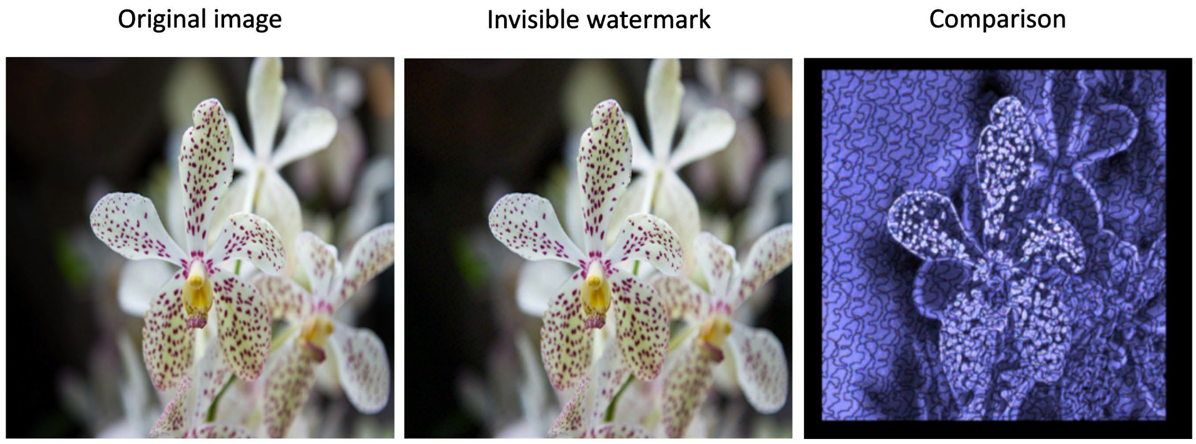
*In the following chapters, we provide more detail on what each attribution measure is and how it is deployed. We then look at its benefits and limitations. Finally, we set out what's next for the measure, including how industry can maximise its impact.*

## What is watermarking?

Watermarking refers to techniques that embed a signal into content. They are used in a range of contexts to authenticate content and individuals, and to prevent counterfeiting and other forms of criminal activity. Visible watermarks, for instance, are applied to banknotes and passports to validate their authenticity. They can also help to protect copyright and other intellectual property rights by proving ownership of content and tracking unauthorised use of copyrighted material.

In an AI context, model developers can use watermarks to signal that the content generated using their tools is synthetic rather than human-made, and to communicate other forms of information, such as the specific model used to generate the content. While some watermarks are visible, in this paper we focus on invisible watermarks that can only be detected by algorithms.[5] Watermarks of this kind subtly alter aspects of the content, such as the colour of specific pixels in an image or particular frequencies in an audio recording, in ways imperceptible to the human eye or ear.

**Figure 3: Comparing watermarked and non-watermarked images**



Watermarking techniques are developing at pace and can now be applied to text, image, video and audio content. The market for such solutions is rapidly expanding, driven by initiatives from technology firms and the growing availability of open-source tools and resources. Google DeepMind released their SynthID toolkit in August 2023, which watermarks image, video, audio – and most recently – text outputs. SynthID is integrated into Google's GenAI tools, such as Gemini and Imagen, and is also available on an open-source basis for text.[6]

[5] We acknowledge that alternative definitions exist for watermarking, see: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf. We use Fernandez' framework which categorises watermarks based on the point at which they are deployed in the content lifecycle, see: https://arxiv.org/abs/2502.05215

[6] However, the complete end-to-end system, including the watermark detection models and unique generation keys, remains proprietary.

Similarly, Meta has rolled out a watermarking tool for images called Stable Signature, which embeds a watermarking algorithm into the weights of a model during the training process. Stable Signature features within many of Meta's AI tools, such as those available on WhatsApp, Facebook and Instagram, and appears to have been used many times via model intermediary platform GitHub.[7] Smaller technology firms have also debuted watermarking applications, including Resemble.AI, whose PerTH tool embeds watermarks within audio content.

Alongside these industry efforts, many academic researchers are now investigating the strengths and limitations of different techniques, with hundreds of papers documenting their analysis. Some academics have gone further and proposed their own watermarking solutions, from the University of Maryland's Tree-Ring to the University of California Berkeley's StegaStamp.

# How are watermarks deployed?

## The three stages of watermarking

Invisible watermarking typically involves three stages:

1. **Embedding.** A watermark is first created and embedded into a piece of content using an embedding algorithm that introduces subtle modifications.[8] In images, this may involve altering the colour of specific pixels; in audio, it can mean changing elements of its frequency. Video watermarking may entail both techniques. In the case of text, watermarks can be embedded by shifting the probability of words, syllables or parts of a sentence generated by a language model.

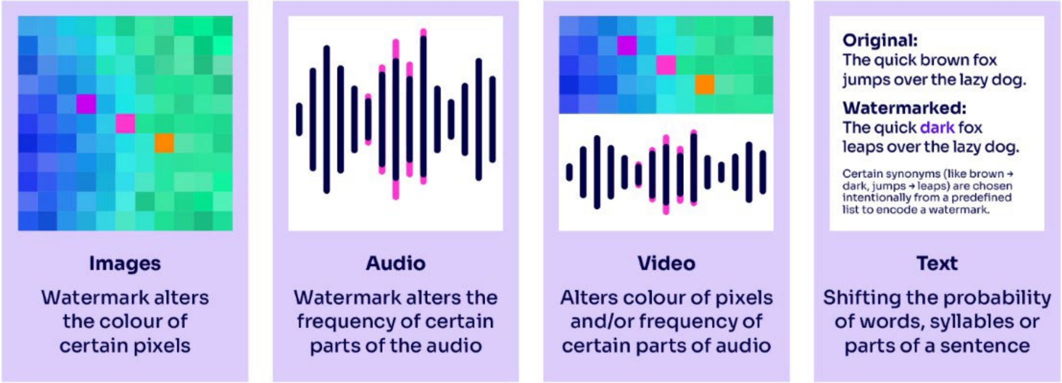
2. **Detection.** A separate algorithm is used to detect and validate the presence of a watermark. This often involves the use of a detection key – a piece of information necessary for watermark extraction and verification. These keys can be public (freely available) or private (restricted to trusted users).

3. **Interpretation and action**. The watermark detector algorithm often provides a probabilistic 'confidence' score indicating the likelihood that a piece of content contains a watermark. DeepMind's SynthID text watermarking tool, for example, provides a percentage probability that a string of textual content is watermarked. Depending on the confidence score, the person or organisation extracting the watermark may then wish to take further action. For example, a platform hosting proven watermarked content may choose to apply a visible label describing it as synthetic or otherwise.

[7] At the time of writing, Stable Signature has been cited 277 times in academic papers and its code has been used approximately 55 times to begin a new project.

[8] Watermarks can be made to be irreversible so that distortions made to the content by the watermark cannot be easily reversed to the content's original form. Alternatively, the content can be reversible so that it can be reconstructed without the watermark.

**Figure 4: How watermarks are applied to different types of content**


Images
Watermark alters the colour of certain pixels
Audio
Watermark alters the frequency of certain parts of the audio
Video
Alters colour of pixels and/or frequency of certain parts of audio
Text
Shifting the probability of words, syllables or parts of a sentence

**Original:**
The quick brown fox
jumps over the lazy dog.

**Watermarked:**
The quick dark fox
leaps over the lazy dog.

Certain synonyms (like brown → dark, jumps → leaps) are chosen intentionally from a predefined list to encode a watermark.

## In-model, out-of-model and post-hoc watermarks

The algorithms used to watermark content can be deployed at different stages of the content generation process.[9] This decision can have a bearing on whether a watermark is vulnerable to removal. There are three main options available to those watermarking content:

- **The 'in-model' option** – In this case, the watermarking algorithm is, in effect, integrated directly into the weights of the model during its training or finetuning.[10] The aim of this is to ensure that the watermarking method is more robust to tampering and watermarks will still be generated even if the model is later modified (finetuned). Meta's Stable Signature tool operates on this basis. Given that this approach requires full access to the underlying architecture of the model, it is only feasible for AI developers rather than those procuring and deploying models 'downstream' (e.g. social media platforms). The in-model approach also usually requires AI developers to undertake additional finetuning or training on their model, which can increase costs and complexity.
- **The 'out-of-model' option** – A second option is to place the watermarking algorithm 'on top' of the model so that any output it produces includes a watermark. It manipulates the output process without altering the model's weights. This is the approach taken by Stability.AI. Unlike with in-model watermarking, the out-of-model option can be deployed not just by the AI developer but by any actor who has access to the model's API. This could include an online platform hosting the model. However, the same actors can also remove the watermarking algorithm should they wish. In the case of open-source models, anyone can disable an out-of-model watermarking feature.[11]
- **The 'post-hoc' option –** A third option is to apply a watermark to content after it has been generated, for example by using the Content Authenticity Initiative's (CAI's) TrustMark. This model-agnostic approach allows creators to apply their own watermarks to existing digital content. The main drawback of post-hoc methods is that their use is discretionary, and in an open-source context an actor can easily bypass this step.

[9] We use Fernandez' framework which categorises watermarks based on the point at which they are deployed in the content lifecycle. See: https://arxiv.org/abs/2502.05215

[10] This involves making a small but permanent adjustment to the part of the GenAI model that generates an image. This teaches it to build a hidden watermark into every picture it creates. Because this change is integrated so deeply, it's difficult to reverse without damaging the GenAI model's ability to generate high-quality images.

[11] Some open-source model licenses *may* seek to prohibit a third party from removing the watermarking module, but it would still be technically possible to remove it.

The above options should not be seen as exhaustive. Watermarking techniques continue to evolve, with innovative methods emerging periodically from academia and industry. One idea that has picked up interest in recent months is that of 'dataset watermarking'. This involves manipulating training datasets in such a way that any model that is trained on that data produces outputs with a distinct pattern.[12] Researchers at Facebook and INRIA have applied this approach in the context of image generators, using what they describe as a 'radioactive' image training dataset to train a model that creates images with a unique invisible marker. Academics at the University of Maryland found that they could later identify these watermarks in deepfake video content.

**Table 1: Types of watermarking approaches**

| | Overview | Actor involved | Example |
|---|---|---|---|
| **In-model** | Watermarking algorithm is embedded directly into a model's weights during its training or finetuning process. | Model developer | Stable Signature |
| **Out-of-model** | Watermarking algorithm is embedded 'on top' of the model so that any output it produces contains a watermark | Model developer, online platforms with API access to GenAI models | SynthID, Stability.AI, TreeRing (University of Maryland) |
| **Post-hoc** | Watermarking algorithm applies a watermark after content has been generated | Third parties, including online platforms, academic or civil society researchers, content creators | TrustMark, PostMark, WatermarkAnything |

# What are the benefits of watermarking?

## Watermarks are imperceptible and unlikely to diminish the user experience

A key advantage of invisible watermarks is that they are designed to be unseen and unheard, meaning they should not diminish the quality or aesthetics of the content. This is particularly important when generating synthetic content to be used in the creative and entertainment sectors (e.g. videos in film production) or in commercial settings (e.g. images for visual presentations or videos for training simulations). Imperceptibility allows for a natural and uninterrupted user experience, allowing benign forms of synthetic content to be shared across online platforms without requiring special handling.

A related benefit is that invisible watermarks avoid the 'scarlet letter' effect. This describes the potential for content to be unfairly dismissed by users or viewed as low quality simply because it is

[12] This type of watermarking may also help users to assert copyright claims. For example, a content creator could watermark their own content and monitor if it appeared in a GenAI model's output, to help trace where their content has been used to train GenAI models.

visibly marked as being AI-generated (something we will explore further in the chapter on AI labels). Invisible watermarks allow content to be evaluated without this prejudgment.

### Watermarks could allow deepfakes to be traced back to a specific model

Beyond being imperceptible, invisible watermarks can attribute content to the specific AI model that generated it, offering the potential to trace deepfakes to their source. For example, a social media service using a variety of watermark detection algorithms could find that many of the deepfake fraud adverts appearing on its platform are linked to the models of a single developer. The service could then pass this information onto that model developer, potentially resulting in the introduction of more robust model safeguards or enforcement action by the developer.

In principle, this form of model traceability could equally apply to open-source models, so long as a watermarking algorithm has been integrated 'in-model' (in the open-source model weights). While the watermarking algorithm *can* be removed through finetuning, the finetuning process is a high-effort and computationally expensive attack that requires technical expertise by an adversarial actor. A study undertaken by researchers at the Oxford Internet Institute found that open-source models have formed the basis for many thousands of new model variants dedicated to creating non-consensual sexual deepfakes. The study recommends that model developers adopt more robust watermarking techniques to allow for deepfakes to be more easily traced back to their source.

### Watermarks can be detected automatically and at scale

Invisible watermarks are designed to be identified automatically by algorithms, rather than relying on human perception. This allows for content to be analysed at scale and at speed, so long as the platform using the detection algorithms has sufficient technical expertise and computing resources. This is an essential quality for social media platforms processing enormous volumes of content. Researchers and industry are also continuing to pursue improvements in the speed at which watermarks can be extracted from content. Meta, for example, claims that the detection algorithms used within its AudioSeal tool allow for 'real time' detection of watermarks in audio streams. Its researchers note that AudioSeal achieves 'two orders of magnitude faster detection' than previous state-of-the-art-models.

A related benefit of watermarks is that they are designed to be identified and acted on by the platforms that host the content, as opposed to users. Compared to some other attribution measures, they don't require people to interpret a signal or weigh up evidence about a piece of content, which demands time, bandwidth and media literacy that may not always be commonplace. That said, platforms could still make errors in the process of identifying and processing watermarked content, for example with genuine content potentially being misidentified as synthetic, and vice versa.

## What are the limitations of watermarks?

### Watermarking detectors are not interoperable

Many types of watermarking tool exist today, and many AI developers now add watermarks to their model's outputs. This includes Google DeepMind for Imagen and Gemini, Meta for its AI applications, Microsoft for Bing Image Creator, and Amazon for Titan. However, each AI developer tool uses a different approach to embed a watermark in content and each requires a different algorithm or key to detect those watermarks. This means that an online platform, journalist or law enforcement agency that seeks to identify watermarked content would need to operate multiple different detectors, which could be costly and complex.

Moreover, AI developers are often selective in who they share detection algorithms and keys with. While many organisations have an interest in detecting watermarks, only some of these will be allowed access to the necessary tools. Several developers, including Meta, don't allow the public to independently check for the existence of their watermarks.

Deciding who should have access to detection algorithms and keys is not straightforward. Some AI developers have responded requests for publicly available watermark solutions. Amazon, for example, has opened up its Titan watermark detection API such that anyone can check for the existence of a Titan watermark on a piece of content. Yet opening up access to watermarking solutions introduces security risks. Not only could it make it easier for bad actors to remove watermarks from content, it could also lower the barriers to people creating counterfeit watermarks to be applied to real content. One way of managing these risks is for AI developers to provide limited access to their detection algorithms and keys. Google DeepMind, for instance, announced that it will be providing partial access to its SynthID detector to a select group of journalists, media professionals and researchers.

## Some watermarks are not robust to content edits

Industry has made efforts to improve the robustness of watermarks, yet some still fail to survive simple content edits of the kind that many people engage in every day. As part of our research on attribution measures, we investigated how resistant three popular watermarks were to removal. We found that some types of edits, such as cropping the image, could result in the watermark being removed from content (see **Box 2** for more details). These results align with evaluations undertaken by academic researchers.

As well as being removed unintentionally, watermarks have been shown to be susceptible to malicious attacks that deliberately seek to remove, reverse engineer or damage the watermark. OpenAI, for instance, paused their roll out of a text watermarking tool last summer as they felt it could be manipulated by bad actors. Researchers from Harvard University, George Mason University and Sapienza University of Rome, meanwhile, were successfully able to remove model watermarks, even where they lacked access to the underlying model and watermark detector. Reflecting on their findings, the researchers argue that "strong watermarking is impossible to achieve". More recently, several independent technology experts identified that they could use GenAI models themselves to remove watermarks from images.

### Box 2: Evaluating publicly available watermarking tools against common benchmarks

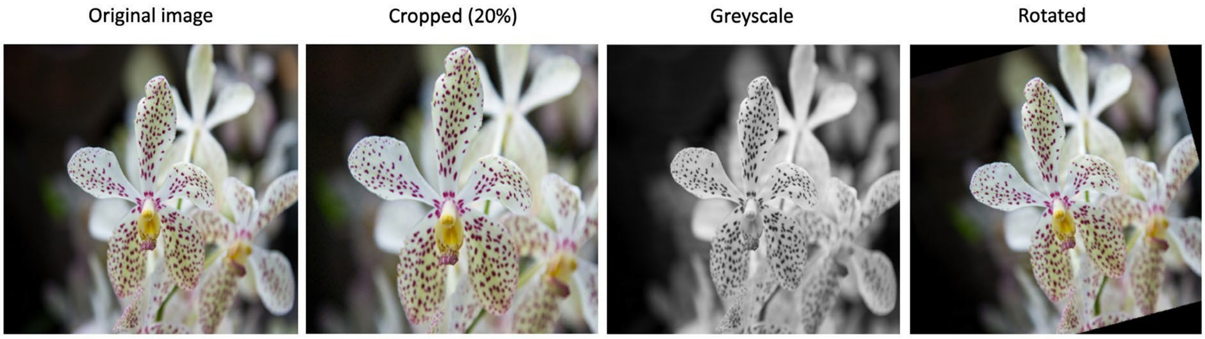
Ofcom's Online Safety Technology team ran several experiments in 2025 to test the robustness of three well-known and openly available watermarking tools. We began by using the tools to add watermarks to hundreds of images, which featured a mix of objects and animals. We then edited those images and looked at whether the watermarks remained in place. In all, we tested 26 distinct types of content edits, some of which entailed minor adjustments that most people could make, with others requiring more skill and time. Where possible within each type of edit, we tested different intensities of the edit to give us a total of almost 100 different edits, for example cropping 1% of the image and then cropping 10%.

The experiments showed that some edits did not dislodge the watermarks. For example, erasing a small part of an image or overlaying text onto an image all had no noticeable effect on the presence of any of the three watermarking types. We found that post-hoc watermarks performed significantly better than the in-model watermark across many of the edits, with post-hoc watermarks very often remaining in the images after they had undergone physical or colour-based alterations.

However, the watermarks weren't robust against all of the content edits. Even simple adjustments such as cropping an image resulted in the removal of the watermarks. While the watermarks weren't always robust against edits to the images, it's also important to consider the robustness of the watermarking algorithm to withstand attempts to remove it from the image generation architecture. In-model watermarking algorithms, for example, are considered to be more robust against attempts to remove them, when compared with other types of watermarking approaches.

As intended for invisible watermarks, we found that the watermarked images retained their original quality, and the watermark itself could not be seen by the human eye. See **Annex 1** for more information about our evaluation method, and the edits we applied to the images.

**Figure 5: Example of watermarked images and image edits**



## Watermarking may be less effective for audio and text content

Several factors continue to hinder the effectiveness of tools used for watermarking audio and text content. In the case of audio watermarking, some of the experts we spoke with commented on the difficulty of watermarking speech content. This was in part due to the relatively narrow range of frequencies in human voices, and the fact that this type of audio is sparser than others such as music, which has many sounds layered on top of one another. Another reason is the 'uncanny valley' effect, with listeners often noticing even minor alterations to speech audio. There is also a risk that watermarks are stripped from audio that is shared via phone calls and messaging services, given that providers will often filter or compress this content.

Text watermarking tools can suffer from similar challenges. For example, tools that work by deliberately including or prohibiting certain 'tokens' in the output of a model could result in the generation of lower quality or even nonsensical content (e.g. incorrect answers to maths sums). In turn, this could create a disincentive for AI developers to adopt such tools. In a bid to address this weakness, a group of researchers recently proposed a 'soft' watermarking approach, which works by biasing the model to prefer certain tokens (e.g. some words over others). This allows for the creation of a watermark without blocking the generation of certain content.[13]

[13] This method reportedly makes synthetic text detectable from short spans of words (as few as 25 tokens), while false positives were "statistically improbable".

### Watermarks may pose risks to privacy

Another potential weakness of watermarking is that it could pose risks to people’s privacy, depending on how watermarking tools are designed. Researchers at the University of California highlight the example of email-based watermarks being used to identify whistleblowers. This is possible in circumstances where recipients of an email each receive a unique version (e.g. with different whitespace), allowing for the later identification of recipients who leak that material. Some of the experts we interviewed also raised concerns that large scale watermark detection – which could involve scanning all content for the presence of watermarks – could amount to a form of surveillance.

## What’s next for watermarking?

Watermarking techniques will continue to evolve in the coming years. To maximise the benefits of these solutions, we would encourage watermark developers, together with industry and academic researcher communities to:

- **Develop interoperable watermarking standards for embedding, detecting and auditing watermarks.** This would make it easier for platforms and other third parties to detect watermarks. The Brookings Institution has suggested that realising such standards would require international coordination between model developers, third party services, policymakers and standards bodies. Among other agreements, these organisations would need to reach a consensus on how to share watermark detector capabilities securely.[14]
- **Carefully design and finetune all elements of the watermark tool.** It is important that watermark developers design robustness into both their watermarking algorithm (which embeds watermarks within content) and its corresponding watermark detector so that they may perform better across different types of attack. While one watermarking algorithm appears to be difficult to remove from its associated model, our technical evaluations found that the watermarks in the images produced by the model could be removed through basic content edits.
- **Evaluate the robustness of watermark tools against publicly available benchmarks before deployment.** This would go some way towards revealing how resistant those watermarks are to removal via several dozen edit types, including those used by bad actors and everyday users. A US NIST report identifies several publicly available benchmarks that could be used for this purpose, including WAVES,[15] StirMark, NIST’s Watermarking Benchmark Suite and Break Our Watermarking System (BOWS). Those developing and testing watermarking tools should make this information available so that users can compare solutions and choose the one that is most reliable for their purpose.

[14] The Brookings Institution also advocates for a registry of watermarking models and services, and the establishment of a ‘continuity plan’ for a watermark detection services where the associated model or watermark developer discontinues their service.

[15] The WAVES paper proposed a standardised framework for testing for distortion, regeneration and adversarial attacks.

# Provenance metadata

## What is provenance metadata?

Almost all software applications and online platforms deploy some form of metadata. Metadata provides information about content, for example its title, captions, colour schemes, access controls or provenance. Metadata has many uses, ranging from supporting brand recognition among publishers and creatives, to helping enforce disclaimer laws, to providing content quality indicators that enable online platforms to make better moderation decisions.

**Provenance metadata** is a specific type of metadata that describes the history and source of a piece of content. This information can encompass the:

- Authors who created or edited the content
- Editing and version history, which sometimes includes details on the type of edits made
- Time history on when the content was made, edited or processed
- Geographic locations from the camera or device used to create or edit the content
- Software and devices (such as the AI models) used to create or edit the content

This information can empower a reader – be they a user or a platform that hosts the content – to assess if and how the content has been manipulated.

To be effective, provenance metadata schemes require the involvement of many actors across the content creation, editing and dissemination stack. Their success is also contingent on the widespread adoption of shared standards. This is because the creation and editing of content takes place across multiple settings, including digital cameras, mobile phones, AI models and editing software. Several openly available provenance metadata standards have been established.[16] The most well-known is the Coalition for Content Provenance and Authenticity (C2PA)'s specification, which outlines a technical approach for embedding and detecting provenance metadata. Many organisations have signed up to this standard, including international camera and mobile phone manufacturers, AI developers, and online platforms.[17] Qualcomm recently announced that smartphones using its chips will be able to embed provenance metadata on content created on those devices.

## How is provenance metadata deployed?

### The four stages of embedding provenance metadata

There are four key stages involved in embedding provenance metadata.

**1. Generating provenance metadata**

Metadata may be generated manually or automatically, and either at the point at which content is created or when it is edited at a later stage. Camera manufacturers (including Fujifilm, Sony, Canon, Nikon and Leica), AI developers (including OpenAI and Amazon), and some mobile phone companies (Samsung) are now embedding metadata into the content their products produce, for example, when a user takes a photo. The same is true of many content editing tools, from mainstream

[16] The US technology agency, NIST, identifies nine provenance metadata standards, which includes the International Press Telecommunications Council (IPTC) Photo Metadata which has been updated to reflect GenAI models: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf

applications like Microsoft's Paint and Designer, to more novel AI editing tools like the voice impersonator app Respeecher, which adds provenance metadata to downloaded audio recordings.

Under the C2PA specification, actors will digitally 'sign' provenance metadata at the point of its creation using a key that only they have access to, to help prevent information being tampered with. This results in a **'manifest'** which represents a single event in the history of the content and means that actors or software involved in the creation or editing of content can be distinguished from one another. Content that has undergone multiple edits will have several manifests which are compiled into a 'manifest store'.

Some types of metadata are allowed to be removed by trusted users. It may be necessary, for example, to redact the name of content creators in cases where this poses a risk to their safety.

### 2. Binding provenance metadata to content

Once generated, metadata can be attached to the content or be stored externally. There are two main techniques for attaching provenance metadata to content: hard binding and soft binding, which can also be used in combination.[18] The key difference between the two is how well they tolerate a piece of content being edited:

- Hard binding creates a strong link between the provenance metadata and the content, breaking when content has been altered. This is best to use when the intention is to authenticate an original piece of content (e.g. whether video footage has been taken on a given camera).
- In contrast, soft binding links the provenance metadata and the content in such a way that allows for the content to be edited without the metadata breaking. This is preferable to use when the intention is to trace content throughout its creation and editing. The soft binding method has the weakness of being vulnerable to intentional removal by bad actors.

### Box 3: How does hard and soft binding work?

Hard binding uses a cryptographic hash to uniquely identify the entire asset, or a portion of it. The process generates a unique, fixed string of characters, so that if the content is altered, the newly generated hash will not match the original, indicating that the content has been tampered with (note that even benign edits, such as compression, will break the hash match).

Soft binding can include perceptual hashes or watermarks to link metadata to its content. The former creates a fingerprint of the contents' visual or audio characteristic, while the latter creates an imperceptible mark that is embedded directly into the content.  Both methods are designed to recognise content even after it has undergone certain types of edits, enabling the content to be re-linked to its provenance metadata, even if the original link has been lost.

### 3. Detecting provenance metadata within content

Once provenance metadata has been generated and attached to content, it is ready to be verified. This is done using applications that confirm the integrity of a piece of content and the authenticity of its manifest, by comparing its cryptographic hash to the manifest (in the case of hard binding) or using a perceptual hash fingerprint or watermark (in the case of soft binding). Verification applications can be used by a variety of interested parties, including online platforms, journalists and members of the public.

[18] The C2PA specification proposes that organisations apply both hard and soft binding approaches.

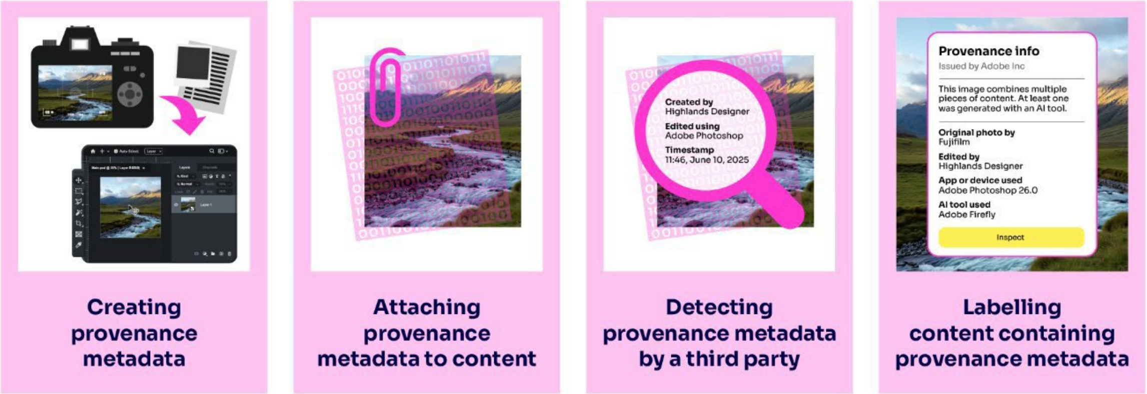
One such tool is the C2PA's Content Credentials verification service which extracts C2PA metadata from images, video and audio files where it is embedded. Some provenance metadata detectors now operate at the internet browser level, including the IPTC Photo Metadata Inspector extension for Chrome and Mozilla Firefox, and Digimarc's C2PA Content Credentials checker Chrome browser extension.

**4. Applying a provenance metadata label**

Platforms have the option of applying a user-facing label to content once metadata has been identified. Google Search, for example, has an 'About this Image' feature that can display metadata alongside images as they appear on the pages of search results. TikTok and LinkedIn, meanwhile, automatically detect metadata in audiovisual content as it is uploaded, at which point a label with that information is made available to users. Similarly, Google's Photos App shows metadata in an 'AI info' section explaining how photos have been edited with its Magic Editor and Magic Eraser. The next chapter provides more detail on the merits of visible content labels.

As well as showing users the information contained in content metadata, some platforms will use it to inform their internal content moderation decisions. For example, Google announced last year that their advertising systems have started to integrate C2PA metadata, with a view to using that information to enforce their policies.

**Figure 6: Deploying provenance metadata**


Creating provenance metadata
Attaching provenance metadata to content
Created by
Highlands Designer
Edited using
Adobe Photoshop
Timestamp
11:46, June 10, 2025
Detecting provenance metadata by a third party
Provenance info
Issued by Adobe Inc
This image combines multiple pieces of content. At least one was generated with an AI tool.
Original photo by
Fujifilm
Edited by
Highlands Designer
App or device used
Adobe Photoshop 26.0
AI tool used
Adobe Firefly
Inspect
Labelling content containing provenance metadata

# What are the benefits of provenance metadata?

## Metadata carries detailed information

Provenance metadata can capture detailed information about a piece of content's history and origin. This includes information about the creator, the tools used to produce it (including AI models) and the location of its creation. In contrast, watermarks typically embed only a limited amount of data, often just an identifier that relates to a given AI model. Depending on the type of system used, metadata can also record the trail of edits made to content, whereas watermarks are static signals that do not provide a history of changes.

The depth of information provided by metadata may be valued by both users and the platforms that host content, potentially enabling them to make more informed judgements about whether the content can be trusted. Take the example of metadata showing that a photo was real content taken by a digital camera, and then edited later using AI software known to be used by professionals. A user or content moderator presented with this level of detail would likely make a different judgement than they would if they were notified with a simple alert that AI was used at some point in the overall creation of the content.

Provenance schemes may help to verify information in environments where there is significant exposure to disinforming deepfakes, for example, those that purport to show real events. YouTube automatically recognises C2PA metadata in videos captured by a participating camera, software or mobile app, and includes this information in its metadata label, 'captured with a camera'. In theory, this should make it harder to fake footage of real-world events captured on these devices. Truepic set up Project Providence which uses C2PA metadata to help authenticate images and videos related to the Ukraine-Russia conflict, with Ukrainian prosecutors relying on this metadata in legal proceedings.

## Common standards make metadata easier to adopt

Provenance metadata standards are now well established, particularly the C2PA specification. Many organisations have voluntarily adopted this approach, and the standard may shortly be confirmed as an ISO international standard, potentially driving further adoption. Added to these formal efforts is a groundswell of bottom-up activity from researchers and civil society groups, who have developed discussion forums and guides to aid the adoption of metadata standards.

The relative maturity of the provenance metadata ecosystem has made it easier for organisations to participate. While it is difficult to generalise, organisations may find that adopting a metadata scheme requires fewer resources than implementing watermarking tools. As previously discussed, there is no common standard for watermarking, which means that organisations seeking to detect watermarks must employ multiple detection algorithms.

This does not mean that establishing a provenance metadata framework is pain-free. Audio generator app Respeecher, for example, found it took several weeks to prove that they were a legitimate organisation so that they could access a key to embed provenance metadata. They also reported challenges in implementing the cryptographic methods required for hard binding metadata. Yet these barriers were ultimately able to be overcome.

### Box 4: Resources required to deploy a common provenance metadata framework

The resources involved in implementing a provenance metadata scheme will vary depending on the complexity of an organisation's systems and whether they are embedding the metadata or detecting it. For large social media platforms that have complex infrastructure, and which host significant volumes of content, or for model developers seeking to embed metadata at source, the costs involved in deploying metadata are likely to be greater. The camera manufacturer Leica explained that it took around a year from initially demoing the Content Credentials scheme to rolling it out fully across their camera hardware.

# What are the limitations of provenance metadata?

## Metadata labels may not be well understood by ordinary users

While metadata can communicate a rich amount of detail about content, this information is not necessarily understood by every user who sees it. The qualitative study we commissioned to inform this paper found that many ordinary internet users did not recognise provenance metadata labels and were not confident in their ability to correctly interpret the information contained in the label. In fact, some did not know why certain provenance metadata information was included or what it was meant to signify. Although participants generally found it helpful to know which AI tool or device was used to create the content, they said they would also like to know how they were used.

The study also revealed that some users found it difficult to make sense of the headline label that was shown to them (i.e. not just the detail included in the metadata). Some of those who were shown the Content Credentials label assumed that it was either a logo, an indication that the content was created on a mobile or web app, or part of the content itself. Despite this lack of understanding, the research participants said they appreciated the additional transparency provided by the provenance metadata, which they believed would enable them to do further research should they wish. Many were also able to correctly interpret that certain provenance metadata indicated that content had been manipulated or tampered with and could be untrustworthy.

Researchers elsewhere have shared similar concerns about people's ability to critically assess metadata. Academics at the University of Washington, for instance, have warned that internet users may mistakenly believe that the presence of metadata means that a piece of content is authentic. As a recent paper from NIST makes clear, users could accidentally or deliberately enter incorrect metadata at the point at which it is recorded and signed. It is also possible that users come to view the absence of metadata as a signal that the content is less trustworthy, which could discredit legitimate content sources.[19] Such issues may be resolved in time as people become more aware of provenance metadata schemes.

## Provenance metadata could be removed

As with watermarks, it is possible in some cases for bad actors to intentionally remove or manipulate metadata. At its simplest, hard binded metadata can be invalidated by editing the content or removed by taking a photo or video of a piece of content – a practice known as air gapping.[20] Soft binding approaches can also be vulnerable to manipulation; bad actors can create a manipulated file that is visually similar enough to an original to share its perceptual hash, thereby allowing the fake content to be incorrectly linked to the original content's metadata. More elaborate techniques to manipulate soft binding approaches can involve bad actors extracting and copying a digital fingerprint to another piece of content.

While it is generally desirable to ensure provenance metadata cannot be removed from content, there will be circumstances where it is legitimate for platforms to do so. For example, the names of journalists, political activists and other users may need to be stripped from metadata where this poses a risk to their safety. Indeed, the harm assessment undertaken by the CAI found that metadata containing personal information could be misused for 'targeted exposure and harassment'. Similarly, it may be necessary to remove geolocation metadata from content to preserve the privacy of content creators and editors, particularly for groups at a heightened risk of harms like stalking and harassment, such as women and girls.[21]

[19] The Content Authenticity Initiative's harm assessment suggested that content which lacked metadata could be downranked or otherwise made less visible on platforms, which could end up penalising users who interact with older devices or software from firms who lack the resources to embed metadata in content. However, there isn't evidence to suggest that this is happening yet. See: https://c2pa.org/specifications/specifications/1.4/security/_attachments/Initial_Adoption_Assessment.pdf

[20] In this scenario, soft-binding methods such as perceptual hashing could be used to rematch the metadata to content with a similar hash.

[21] Ofcom's draft guidance on protecting women and girls outlines how platforms can prevent location data being inadvertently shared.

There are also more prosaic reasons for removing or manipulating metadata, such as when trusted organisations and journalists need to redact inaccurate information. Several platforms choose to strip provenance metadata from content uploaded to their site to conserve bandwidth.

The CAI is exploring how to address several harms outlined in their harm assessment for the C2PA specification, which includes providing a means to users to challenge inaccurate or misleading provenance data that relates to them.

## What’s next for provenance metadata?

Advocates of provenance metadata are taking steps to respond to the above limitations. To maximise the benefits of provenance metadata solutions, we would encourage the architects and signatories of provenance metadata schemes to:

- **Update provenance metadata frameworks –** The C2PA has set up working groups to help update and scale its standard, which includes supporting metadata bindings for live video streams, 3D images formats, and audio use cases. It is also developing additional approaches for signing metadata to expedite the authentication process.
- **Improve interpretability of metadata labels –** The C2PA specification includes User Experience Guidance for Implementers, while the BBC has outlined several visual principles for communicating provenance information to public audiences. However, more research is needed to understand how metadata labels are understood and interpreted. Platforms could do more to improve the efficacy of provenance metadata labels, for example by making this information visible to internal teams to support content moderation decisions, and by testing labels with users ahead of deploying them.
- **Prevent the wrongful removal of provenance metadata** – Hard- and soft-binding approaches are not mutually exclusive, and stakeholders – including Adobe – advocate for using both in combination to increase the overall strength of content provenance systems. Securing assurances from platforms not to strip provenance metadata unnecessarily would be another way to increase the likelihood of metadata remaining in place throughout the content lifecycle.
- **Review the implications of provenance metadata for privacy and freedom of expression –** The CAI’s harm assessment provides important analysis on potential and unintended impacts of provenance metadata schemes. In doing such an exercise, organisations put themselves in a good position to consider how they can respond appropriately to those risks. WITNESS has recommended that provenance metadata organisations go further to explore how to redact or remove metadata information for individual use cases.

# AI labels

## What is AI labelling?

A label consists of a visible and recognisable icon that signifies a given characteristic or a risk to users. It is designed to be easily accessible and is commonly used to help users understand and navigate online content. Many social media platforms use a tick icon to indicate that something about a user has been verified, or a maturity label to show whether content is age appropriate. A lock sign on a browser, meanwhile, indicates that a connection is secure. For many years, academics have studied the impacts of disinformation-related warning labels on users.

These established labelling practices are now being adapted for AI-generated content. Several platforms – including YouTube, Meta and TikTok – are exploring AI label strategies to alert users to synthetic content. The same is true of upstream model developers, with the EU's AI Act requiring AI developers to ensure that synthetic outputs are marked and detectable.

AI labels can be applied to images, videos, audio and text. They tend be known as process labels, in the sense that they communicate the technical processes through which content has been created or modified, e.g. 'Made by AI' labels. These labels are 'value neutral' and cover a broad range of AI generation and edits, with the user left to reach their own conclusions about how AI might impact the content's legitimacy and meaning. [22]

## How are labels deployed?

Platforms face three choices when deploying AI labels:

### Choosing how to detect AI content

A platform first needs to decide how to detect synthetic content for labelling. They can do this in one or more of the following ways:[23]

- **Self-disclosure** – Some platforms require users to disclose that their content is synthetic at the point of upload. On TikTok, users can switch a toggle that applies a standardised label to the content. Disclosure requirements are not limited to platform users, however; Meta also requires advertisers to label content in certain circumstances, e.g. where they create or alter an advert relating to political or social issues. This is the simplest way to detect synthetic content but relies on the end user acting honestly.
- **Automated detection of content produced by a platform's own AI tools –** Where content is generated by a platform's own AI model, the platform can automatically recognise that the content is synthetic and immediately apply a label. Snap, for example, adds a Snap Ghost sparkle icon to images that have been generated by its own GenAI tools, including its MyAI chatbot and AI Lenses feature. Google detects its SynthID watermark across its GenAI consumer products, from Imagen to Magic Editor.

[22] The EU AI Act requires developers to ensure that synthetic outputs are marked and detectable.

[23] Platforms may also choose to deploy deepfake detection tools. See our first paper on Deepfake Defences: https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754

- **Industry detection standards** – Some platforms have adopted detection capabilities that enable them to identify content that has been embedded with relevant provenance metadata (as explained in the last chapter). LinkedIn announced that it would apply a Content Credentials label to content containing C2PA metadata: This can include information about whether it has been generated or edited by AI. This is a more complex system to establish than one based on the other two approaches above.

## Choosing which AI content to label

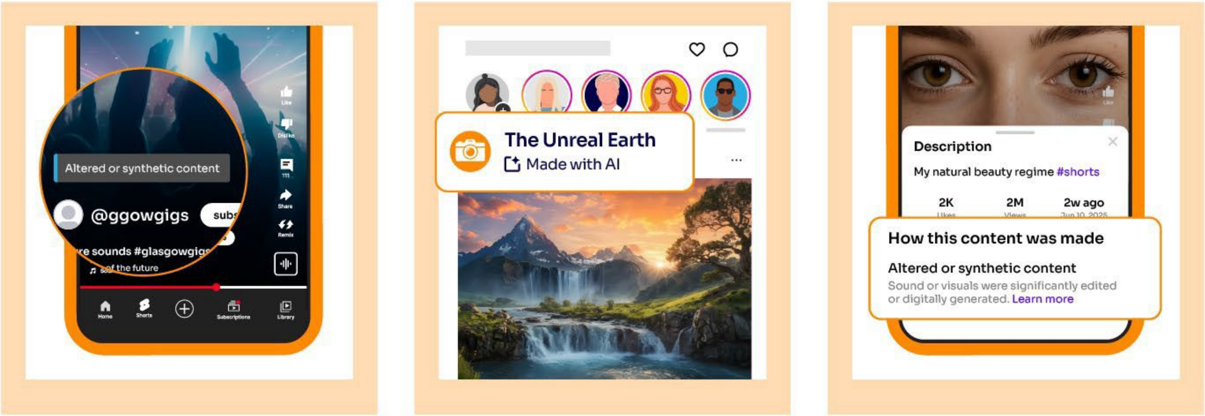
Platforms have the option of labelling all the synthetic content they detect or just a subset of synthetic content that poses a higher risk of deception. An example of the latter approach comes from YouTube's Creator Studio app, which requires creators to disclose only AI-generated or edited content that is realistic, which the platform then labels. Specifically, users are required to disclose content that is generated or edited to mimic the likeness of a real person (e.g. a politician), alters footage of real events or places, or generates realistic scenes. In the same vein, Meta has chosen to make its AI labels more prominent on content that poses a high risk of deceiving the public about important issues. All photorealistic images created with Meta's in-built AI features are separately labelled as "Imagined with AI".

## Placing an AI label

Platforms must also decide on the location and appearance of labels. Typically, AI labels are overlaid onto content to remain constantly visible to viewers. Paradoxically, however, some of the experts we interviewed told us that these always-visible labels could be overlooked by users who may not fully process their significance, potentially due to cognitive overload or habitual exposure.[24]

A different approach involves platforms using a 'reveal' label, which becomes visible only *after* content has been consumed. This aims to prompt reflection and provoke a viewer to reflect more deeply on the content. An 'interstitial' label functions similarly but appears in the middle of a video or audio sequence, creating deliberate friction in the user experience. Due to their dynamic nature, reveal and interstitial labels may attract more user attention. However, there is also the risk that users disengage with content before the label appears.

**Figure 7: Different types of AI labels that users might see**



Altered or synthetic content
@ggowgigs
subs
re sounds #glasgowgigs
of the future
Like
Dislike
Share
Remix
Home
Shorts
Subscriptions
Library
The Unreal Earth
Made with AI
Description
My natural beauty regime #shorts
2K
2M
2w ago
How this content was made
Altered or synthetic content
Sound or visuals were significantly edited or digitally generated. Learn more

[24] This is similar to a 'pre-roll' label which is visible before a piece of content is played.

# What are the benefits of labels?

## AI labels are popular with users, who appreciate their simplicity

In our survey, 85% of respondents said that it is important for platforms to label AI content, with 57% citing this as 'very important'. In our following qualitative research, many of the users we spoke to welcomed the idea of labelling synthetic content with a prominent icon. When presented with three options for conveying information about content – via AI labels, provenance metadata or context annotations (see the next chapter for more detail on annotations) – users in our interviews told us that AI labels were the easiest to understand and engage with. Many appreciated that they did not need to take further action, whether that be clicking through to find out more, reading a lengthy text explanation or having to do their own research. At a time when people are consuming hundreds if not thousands of images, videos, text posts and audio streams every day, icons have the advantage of communicating information in a manner that minimises cognitive overload.

Many of the users in our interviews felt that AI labels could be particularly useful in circumstances where synthetic content poses a greater risk of harm, such as content relating to politics or children. Users also felt that AI labels could be helpful in increasing the reporting of harmful content on platforms. When users were asked specifically about labelling abusive or offensive content created using AI in our interviews, they agreed that such content should be labelled due to its harmful potential. They also stated that they would be more likely to report such content.

## AI labels may help users to critically engage with synthetic content

Some studies appear to justify users' enthusiasm towards AI labels, with evidence emerging that this method can help people to engage more critically with content. Researchers from the Massachusetts Institute of Technology found that users across several countries were able to understand the distinction between 'process labels' (e.g. those noting that content was AI-generated) and 'impact labels' (e.g. those flagging content as a 'deepfake' or 'manipulated'). In our own qualitative research, several users told us in interviews that the presence of an AI label would prompt them to be more cautious when viewing content.

In addition, the immediacy of AI labels may give them an edge over other types of attribution measures in influencing how users perceive content. Academics at the University of Western Australia argue that 'first impressions' are important, as people's initial views of a subject are difficult to shift even after they are presented with contrary evidence. As they put it, "once inaccurate beliefs are formed, they are remarkably difficult to eradicate." This suggests that visible, easy-to-encounter labels could play an important role in helping people critically engage with content from the outset.

### Box 5: New types of AI labels

New techniques for labelling content continue to emerge. For example, researchers are exploring how a new type of process-based label, known as an 'epistemic disclosure' label, could provide more context to users (e.g. the name of the model used to create the content), whilst still being visually simple and easy to digest.

In another development, researchers at New York and Carnegie Mellon Universities developed a tool called 'PITCH' which combines audio deepfake detection with a visible label to warn users that a call they are receiving could be a deepfake voice-clone. Initial test results appear to be promising, with the combined human-AI contribution enabling a detection score of 84.5% compared to a human-only detection score of 72.6%.

# What are the limitations of labels?

## Some users will choose not to label their own content

Labelling schemes that involve self-disclosure depend on users being willing to call out their own content as synthetic. Yet deepfakes are often created and posted by bad actors who are extremely unlikely to self-disclose in this way. Indeed, bad actors are more likely to try and remove labels from content, for example by using simple screenshotting or label scrubbing methods. Several free AI image-editing tools explicitly market their ability to remove visible icons.

Beyond malicious intent, everyday users might also be reluctant to add labels to their content. They could fear that doing so might result in their posts being downranked or otherwise made less visible by a platform.

## Labels are less suited to content that is only partially synthetic

The decision about when to apply an AI label is relatively straightforward in cases where content has been wholly generated by AI. However, this decision is less clear cut in situations where real content has been partially edited using AI tools, for example a landscape scene with buildings added or removed, or a photo where someone's physical features have been altered.

This is a live challenge for platforms, with some already facing criticism from users for applying labels to content that has only been lightly edited using AI. Last year, for example, Meta made the decision to change the wording of its label from 'Made with AI' to 'AI info' following complaints from photographers that the original version was creating confusion among their audiences. These concerns are likely to become more pronounced as we move into an age where most content is at least partially augmented by digital tools. We are also likely to see cultural norms evolve in relation to the level of AI augmentation that users believe to be acceptable.

## Despite their simplicity, labels could still be misinterpreted

As noted above, emerging evidence suggests that AI labels could help users to engage with content more critically. There is a risk, however, that these effects in fact become too strong, and that users view AI labels not just as indicating that content is synthetic (or partially synthetic) but that it is deliberatively deceptive and untrustworthy (remember that AI-generated content can of course be entirely benign). Our survey found that 59% of users said they may not trust content with an AI label attached. Conversely, there may be an 'implied truth effect' whereby users believe that all content without a label is authentic (when unlabelled content can be deceptive).[25]

Furthermore, platforms use different visual icons to signal whether content is synthetic, which may cause further confusion among users. Our survey found that 55% of users said they would treat or respond to content differently depending on what the AI label looks like. Research by UX design company Nielsen Norman Group found that the sparkles icon, often used to indicate AI generated

[25] After the 2016 US Election, Facebook started putting warning tags on news stories that fact-checkers judged to be false. In 2019, researchers at the University of Regina, Yale University and Massachusetts Institute of Technology found that tagging some stories as false had the effect of making readers more willing to believe and share other stories even where the untagged stories turned out to be false. See: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384

content, is also employed by other platforms to signal new features, functionalities, discounts or promotions. They argue that the ad hoc adoption of labels by platforms, rather than through official standards bodies or commercial agreements,[26] has created confusion about what labels refer to. Consequently, they recommend that labels should only represent a single characteristic (e.g. synthetic or secure) to prevent ongoing misinterpretation.

Some of the experts we interviewed felt that AI labels could be more effective if targeted at journalists, content creators and others who are likely to have a better sense of their meaning and limitations.

## What's next for AI labels?

Moving forwards, platforms could take several actions to manage these limitations and make the most of AI labels:

- **Avoid relying solely on user self-disclosure when determining which content to label** – Asking users to voluntarily disclose that their content is synthetic would be a valuable first step in establishing an AI labelling system. However, platforms could explore additional methods for identifying synthetic content, and ideally deepfake content, for example by using the watermarking and metadata tools discussed in earlier chapters.
- **Clearly explain to users why and when AI labels are applied to content** – To ensure users don't misinterpret the meaning of labels, platforms should clearly communicate to users why and when they are used. This includes explaining whether they intend to signal that content is partially synthetic and not just wholly AI-generated. Platforms could also use campaigns to boost user understanding of labels.
- **Make AI labels visually eye-catching and use them consistently across devices** – In our qualitative research, users told us that they would like platforms to display AI labels prominently on content, for example through the use of bold text and contrasting colours. AI labels also need to work across all device types, including on smaller mobile screens where they could otherwise be missed.
- **Test AI labels before deploying them** – Platforms could test several label designs before committing to a final version.[27] These tests could assess labels according to how easily they are seen, whether their meaning is understood, and whether they alter the way users engage with the content, for example the likelihood of them liking or sharing it. Platforms could also monitor how users engage with AI labels once a system is fully established. Some of this data could be made publicly available to support wider research efforts.
- **Consider standardising AI labels** – Platforms could discuss whether there would be merit in adopting the same style of label to signal that content is AI-generated or edited. Platforms could also develop a common threshold for when to apply an AI label. While there are valid reasons for platforms to use their own labels – for example, because they cater to different demographic groups or operate in countries where certain icons carry a particular meaning – a universal label could significantly minimise confusion among users.

[26] However, attempts to create common international labelling standards will likely need to consider cultural differences in label interpretation. One study for example found that participants from China interpreted the terms "artificial" very differently than participants from other countries.

[27] A consumer study of the perceived safety of Internet of Things (IoT) devices presented participants with four draft label designs and asked them to provide feedback on how well they convey key information.

# Context annotations

## What are context annotations?

Context annotations provide more detailed information about a given post or its author. They can also provide an alternative viewpoint or explain where a post might contain inaccurate or misleading information. For example, annotations might provide different sources of information to dispute a claim, link to an original image to show how content has been manipulated or point out an author's potential motivations for creating and sharing the content in question.

Annotations are most commonly attached to content that is controversial or sensitive, but which doesn't necessarily reach the threshold for full removal from a platform. This could include, for example, dubious information about vaccine safety, fake claims made about politicians, or misleading information about international conflicts.

Different actors can be involved in context annotations, from independent fact-checking organisations through to news organisations and platforms users. Third party fact-checking has traditionally been the most prevalent type of context annotation, with platforms collectively spending millions of dollars employing fact-checking organisations to support them in responding to misinformation and disinformation. However, several large platforms are shifting towards annotation approaches that involve everyday users, with Meta and TikTok now testing user-led systems. These schemes follow in the footsteps of X's Community Notes, which allow eligible users[28] to annotate one another's posts. Proponents of annotations frame them as a way to "allow more people with more perspectives to add context to more types of content". This makes annotations particularly popular among users who feel that online platforms are over-moderating certain types of content.

## How are annotations deployed?

There are two main types of annotation, one that relies on expert third parties (including fact checkers) and another that involves platform users.

### Expert annotation

Expert annotation is the most established approach and involves trusted organisations like media providers or independent fact-checker organisations providing critical commentary on content. Since 2016, Meta has spent over $100 million on independent fact-checkers, who have rated content as false or altered, including misleading information about the climate crisis and the COVID-19 pandemic.

### Crowdsourced annotation

Crowdsourced annotations involve the collective judgement of platform users, or a subset of eligible or expert users. This approach is based on the 'wisdom of the crowd' theory, which believes that the aggregate opinions of groups are more accurate than those of individuals.

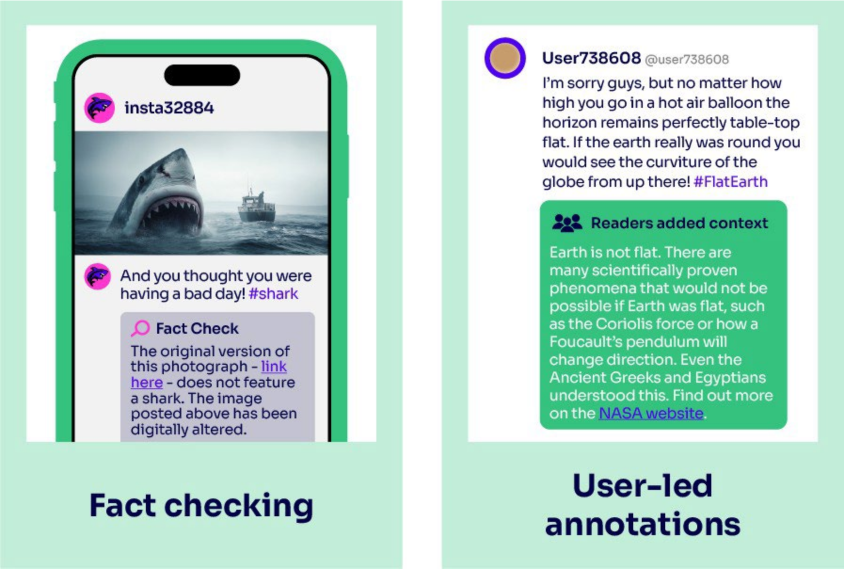
[28] Known as Community Notes contributors on X

Evidence indicates that users tend to focus on different issues than experts when annotating content. A Cornell University study, for instance, showed that crowdsourced annotations focused more heavily on social fraud content, while expert annotations concentrated more on political misinformation.[29] A separate study from the University of Giessen found that crowdsourced annotations on X tended to be placed on posts from influential accounts with a high number of followers.

Regarding crowdsourced annotations, some platforms allow any user to annotate content, while others limit annotation rights to a select group of 'super-users':

- **All users** – X's Community Notes function allows any user to sign up to become a 'contributor', enabling them to annotate any post.[30] These annotations, in turn, can be rated by other contributors based on their helpfulness. If enough contributors from 'diverse perspectives' rate the note as helpful, it is then attached to the post for all users of X to see. The diversity of perspectives is defined by how users have rated notes previously. X says a good indicator of a helpful note is when users who have typically disagreed on something in the past agree on its helpfulness.
- **Super-users** – Some platforms, such as Wikipedia, adopt a tiered approach to crowdsourced annotations, whereby 'super-users' – those deemed the most credible annotators – are allowed to overrule inaccurate Wikipedia entries written by others. An emerging and relatively untested method is the use of super-recognisers for deepfake detection. In this approach, a subset of individuals with demonstrated above-average ability to identify and explain potential deepfake content could be responsible for providing context annotations.

**Figure 8: Types of context annotations**



insta32884

And you thought you were having a bad day! #shark

Fact Check

The original version of this photograph - link here - does not feature a shark. The image posted above has been digitally altered.

Fact checking

User738608 @user738608

I'm sorry guys, but no matter how high you go in a hot air balloon the horizon remains perfectly table-top flat. If the earth really was round you would see the curviture of the globe from up there! #FlatEarth

Readers added context

Earth is not flat. There are many scientifically proven phenomena that would not be possible if Earth was flat, such as the Coriolis force or how a Foucault's pendulum will change direction. Even the Ancient Greeks and Egyptians understood this. Find out more on the NASA website.

User-led annotations

[29] The study looked at the Taiwanese platform Cofacts. Readers should be wary of overgeneralising the findings. Firstly, they are limited to the Taiwanese context, where digital civic participation is particularly strong. Secondly, the Cofacts platform is a dedicated community of engaged citizens.

[30] Any X user can sign up to become a Community Notes contributor, but to be eligible, accounts must have: (1) no recent notice of violations of X's Rules, (2) joined X for at least six months and (3) a verified phone number from a trusted phone carrier. Contributors start with the ability to rate notes, and over time, can earn the ability to write them.

# What are the benefits of annotations?

## Annotations can identify content as a deepfake, not just synthetic

Context annotations have the advantage of explicitly signalling to users when content is likely to be a deepfake (i.e. audio-visual content which is designed to misrepresent someone or something). This stands in contrast to some of the other measures discussed in this paper, such as provenance metadata and AI labels, which only describe whether content has been created or edited using AI. In this way, context annotations may have a more immediate impact on user behaviour, as users are not required to make further assessments about the content. In addition, annotations may be less likely to create the kind of confusion that is associated with some of the other measures we have discussed, for example the false belief that an AI label necessarily means content is untrustworthy.

Crowdsourced annotations, in particular, may have the advantage of drawing on a wide pool knowledge to identify deepfakes.[31] The largest platforms are likely to host users with a wide range of interest areas that could be valuable in pointing out inaccurate content relating to niche subjects, be those in the domain of politics, celebrities and entertainment, or natural disasters. In a possible sign of the quality of crowdsourced annotations, some platforms have chosen not just to display them to everyday users but to use them within their own content moderation systems. This includes Reddit, where annotations are used to provide human moderators with more context about content. Researchers at the Universities of Washington and Michigan that these annotations improved moderation decisions, with moderators finding it particularly helpful when annotations referred to alternative sources of that these annotations improved moderation decisions, with moderators finding it particularly helpful when annotations referred to alternative sources of information. Signals from the context annotation process – such as the volume of annotations attached to a post – could be used to inform whether the content is prioritised for human review.[32]

## Evidence indicates that annotations may reduce the virality of misleading content

Some of the claims made about the efficacy of context annotations appear to be borne out in research studies, with evidence suggesting that annotations have the potential to reduce the virality of deepfakes. Researchers at the Massachusetts Institute of Technology found that warning labels from professional fact-checkers considerably reduced user belief in false news posts and also reduced the likelihood of users sharing those posts by on average a quarter (24.7 percent). A separate study looking at a database of 285,000 context annotations on X found a similar pattern. In this case, the researchers found that the addition of an annotation below a post reduced the number of retweets by almost half (49 percent). Annotations also increased the probability of authors deleting their post by 80%.

The results of these studies are echoed in the findings of our own qualitative research. Several of the users we interviewed as part of this work told us that the descriptive nature of annotations and links to alternative sources of truth helped them to more critically engage with content. They felt this was because the annotations warned them of potential risks and signalled whether the content was

[31] Most crowdsourced annotations focus on diversity of *views*, rather than requiring annotators to have specific areas of expertise.

[32] The same study also found that a high volume of user notes or flags against content could indicate that content may need to be prioritised for human review.

likely to be accurate. Many users then suggested this information helped them decide what action to take, such as reporting the content, ignoring it, or conducting further research.

It is important to recognise, however, that not all studies of context annotation systems reveal positive effects. For example, analysts at the Brookings Institution who looked at a corpus of data about Community Notes on X found that, paradoxically, some users whose content was annotated experienced an *increase* in followers, possibly due to the annotations drawing attention to their account. However, it is difficult to be sure of causation effects with the available data.

### Box 6: Who do users say they trust to annotate content?

There is an ongoing debate about whether general users or experts are best placed to annotate content. Having relied for several years on fact-checkers to moderate its posts, Meta recently announced that it would end its third party fact-checking programme and establish in its place a crowdsourced annotation system, beginning in the US. In a press release announcing these changes, Meta claimed that crowdsourced annotations would be "less biased… because [they] allow more people with more perspectives to add context to posts". In response, Full Fact explained that "fact checkers are committed to promoting free speech based on good information" and are trained to respond to misinformation online.

Several studies have begun to shed light on how both users and content moderators perceive different types of annotators. A study looking at Reddit found that a quarter of the platform's moderators distrusted professional fact checkers. Similarly, our own survey of UK users found that a third (30 percent) were not confident that professional fact checkers would share accurate information about a post containing synthetic content. While this figure appears high, more than twice as many respondents (64 percent) said they would not trust other users of a platform to accurately annotate synthetic content.

In addition to looking at trust, researchers have considered whether some types of annotators are more effective than others in identifying fake content. Academics at Stanford University conducted a large online experiment to understand how well typical users could answer a series of factual questions. Looking at the responses provided by nearly 2,000 participants, they found that 'the crowd' performed better than the average individual respondent, lending some credence to the theory of the 'wisdom of the crowd'.

# What are the limitations of annotations?

## Annotations take time to appear on content

Annotations do not usually appear on content immediately, reflecting the fact that users and experts need time to review material and determine whether it is trustworthy. While this time lag may result in more considered responses, it also means that many users could be exposed to misleading material before it has been debunked. Research by the Center for Countering Digital Hate found that 74% of Community Notes that were submitted for attachment to inaccurate or misleading posts about the US election failed to reach X's publication threshold, which meant that these posts were left un-annotated. This includes a note that was submitted for attachment to a post of an AI-generated audio file impersonating Kamala Harris.

The efficacy of annotations also depends on people – be they users or experts – being able to correctly identify fake content. However, this challenge may become more pronounced as deepfakes become more realistic and sophisticated, with annotators needing more time to identify and debunk them. Researchers at the Brookings Institution believe that AI-generated content will not only rapidly increase in volume over the coming years but be significantly more difficult for untrained users to detect on their own.

Researchers and online platforms are exploring how partially automating the annotation process could help to alleviate these challenges. X has adopted a feature in its Community Notes system to speed up the annotation of content by automatically attaching an annotation to matching visual media posts (where the original visual media post is annotated), reporting that 3,500 original annotations now appear on 331,000 distinct posts. Academics at the University of Washington have developed a large language model called 'MUSE', which aims to identify and explain inaccuracies in posts, as well as provide references to support its statements. According to researchers who tested the tool's capabilities, MUSE is able "to write high-quality responses to potential misinformation—across modalities, tactics, domains, political leanings, and for information that has not previously been fact-checked online—within minutes of its appearance on social media."[33] While these early responses are promising, MUSE has only been tested on a single social media platform and it cannot yet process video inputs. There is also a risk that GenAI models may hallucinate. More recently, X launched a pilot that allows users to create their own AI Note Writing bots, which can propose community notes for posts that have been flagged for review. X believes this will "deliver increasingly accurate, less biased, and broadly helpful information", noting that humans will still be required to rate the notes produced by AI Note Writers.

### Box 7: Visual AI-based annotations to support explainability

Although AI tools have the potential to identify and flag fake content, users can be reluctant to believe these warnings. This has led some researchers to explore novel methods for communicating the decisions of AI annotation tools. For instance, academics at the Massachusetts Institute of Technology ran an experiment where they pointed out to users the aspects of content deemed to be suspicious, such as the naturalness of motion and the coherence of faces in a video. They found that people had higher confidence in the deepfake warnings of AI tools where these visual artefacts were highlighted, compared with traditional text-based prompts.

## Annotations could be gamed by bad actors or be seen to be biased in favour of certain viewpoints

Some have raised concerns that crowdsourced annotation approaches could be taken advantage of. For example, users could deliberately post annotations containing disinformation, coordinate with others to agree or disagree with certain draft annotations, or open multiple accounts to write and rate annotations to write and rate annotations.[34] Previous studies have also warned that crowdsourced mechanisms can be abused by bad actors to target specific user groups, often women and LGBTQ+ communities. X has since adapted their Community Notes feature to respond to coordinated action on rating proposed notes, by treating coordinated ratings as if they all came from a single user.

[33] Researchers at Adobe, the BBC and the University of Oxford have developed a similar system called 'LLM Consensus'. This uses two models that 'debate' the accuracy of content (e.g. an image), before arriving at a unified view, which is then presented as a plain English explanation. The system uses search engine APIs to ensure its responses correspond with the latest news and events. Researchers found that the content explanations generated by LLM Consensus improved the ability of both expert and non-experts to detect misinformation, with the performance of journalists increasing by 12% and ordinary users by 15%.

[34] A WIRED report from 2023 claimed that Russian embassies have engaged in coordinated activity to downvote Community Notes on X containing anti-Russian sentiment, thus reducing the likelihood of them being published to all users. See: https://www.wired.com/story/x-community-notes-disinformation/

There is also the possibility that crowdsourced annotations give disproportionate weight to the views of some user groups over others. Studies suggest that there is an imbalance in the appetite of different demographic groups to annotate posts. For example, a survey published in the Harvard Kennedy School's Misinformation Review suggests that younger and more educated adults are more likely to annotate content.[35] The survey also found that annotation rates are higher among those who use the internet to access the news.

## What's next for annotations?

Looking to the future, platforms that choose to deploy context annotations could take the following steps to maximise their impact:

- **Enlist annotators from a variety of backgrounds and demographic groups** – To reduce the risk of bias in context annotations, platforms could seek contributions from organisations and users reflecting a diversity of viewpoints and demographic groups. This could involve a platform advertising its annotation programme to particular types of users, collaborating with interest groups to promote annotation practices, and allowing annotators to declare information about themselves and their political leanings.
- **Avoid solely relying on users to identify deepfakes** – It is important that platforms don't only rely on users to make decisions about the authenticity of content. Platforms can also learn from one another about how to prevent annotation systems from being misused and manipulated.
- **Make annotations clearly visible to users** – Context annotations will only be effective in so far as users encounter them. Platforms could boost the prominence of annotations by using larger or more striking text, using contrasting colours, or positioning the annotation above content rather than below it. Platforms need to be wary, however, of overwhelming users with too much information that could result in annotation fatigue.
- **Consider including alternative sources of information within annotations** – Several participants in our qualitative research said they would prefer context annotations to include hyperlinks to trusted information sources, allowing them to check information for themselves. This sentiment aligns with the a study of X's Community Notes system, which found that the key determinant of a user finding a note helpful is whether it contains a link to an external source.[36]
- **Test context annotation approaches before deployment** – Platforms could test annotation approaches before they are deployed, to understand how users interact with them and to identify – and mitigate – potentially adverse or unintended consequences. It will be particularly important for platforms to test AI-based annotation initiatives, given they are still in their infancy.

[35] The researchers studied users in the UK and other western nations.

[36] Notes are perceived to be 2.7 times more helpful where the Note includes a link to an external source of evidence.

# 8 key takeaways

This paper has introduced four attribution measures: watermarking, provenance metadata, AI labels, and context annotations. We looked at what they involve, their strengths and weaknesses, and how they could be deployed more successfully to combat deepfakes. While each measure is unique, with its own promises and pitfalls, our analysis reveals several overarching findings that should guide future action by industry, government and researchers in this area.

Below, we expand on the 8 key takeaways that we outlined in our Overview.

1. **Evidence shows that attribution measures can help users to engage with content more critically.** An emerging body of research shows that, when deployed with care and proper testing, attribution measures can improve the capacity of users to spot misleading and inaccurate content. Moreover, studies have shown that some attribution measures like context annotations can reduce the likelihood of users sharing such content. These findings are supported by our own qualitative research, with platform users telling us that they felt attribution measures could help them to make better sense of content. Ofcom's best practice design principles for media literacy provides guidance for how platforms can design, deploy and iterate on platform interventions to promote media literacy.
2. **Users should not be left to identify deepfakes on their own.** While attribution measures like AI labels and context annotations can help to empower users, platforms should avoid placing the full burden on individuals to detect misleading content. Some of the information captured by these tools – such as content metadata – could be used by platforms to inform their own moderation efforts, including which content to prioritise for moderation review. Platforms should also consider whether, in some cases, attribution information would be better directed at expert audiences to make sense of – such as journalists, academics, and civil society groups – rather than users writ large.
3. **Striking the right balance between simplicity and detail is crucial when communicating information to users.** There is a tension between the desire to create concise messages that users can easily grasp, and the need for users to be shown sufficient detail to make informed judgements. Too much information can be overwhelming, but too little can result in confusion. With AI labels, for example, there is a danger that these icons are misread by users as meaning that content is necessarily untrustworthy. These effects can frustrate content creators who post legitimate material, and in at least one case a platform has rowed back on its labelling scheme following complaints from its users. It is also possible that if labelling becomes common, the *absence* of a label on a piece of content could lend that content undue legitimacy (though such claims need further investigation).
4. **Attribution measures need to accommodate content that is neither wholly real nor synthetic.** AI models are increasingly being used to augment existing content, not just to create new content from scratch. Examples include simple AI filters to brighten visual content, AI editing tools for superimposing people or objects into photos, and AI-powered audio translation software that alters soundtracks while preserving the original video footage. Attribution measures will be more effective where they can convey this nuance to users, communicating *how* AI has been used not just *whether* it has been used.
5. **Attribution measures can be susceptible to removal and manipulation.** Several studies have shown that the information that is attached to content via attribution measures can in some

cases be removed through accidental or deliberate means. Our own technical evaluation of publicly available watermark tools revealed that they could be removed from image content following certain types of edits, such as cropping. There is also a risk of bad actors manipulating context annotation schemes, for example by using them to spread disinformation. A further concern is that some attribution measures rely too heavily on the goodwill of model developers, for example to agree to watermark or add metadata to their content.

6. **Greater standardisation could boost the efficacy and adoption of attribution measures.** Model developers, platforms and others often use different approaches to apply attribution measures, which can increase costs and create confusion. For example, each model developer currently has its own method for embedding watermarks, which forces downstream platforms to deploy multiple detection algorithms to pick up on these signals. Similarly, the absence of a unifying approach for applying AI labels risks confusing users who encounter different symbols across various sites. One domain where different actors have successfully coordinated efforts is provenance metadata, with joint standards like the C2PA specification emerging to facilitate its take up.

7. **The pace of change means it would be unwise to make sweeping claims about attribution measures.** Many of the measures and related initiatives explored in this paper are relatively new. Mainstream platforms have only begun to use AI labels in the last two years, and model developers have only recently started watermarking their content. Moreover, we are continuing to see improvements and modifications, for example adjustments being made to context annotation schemes which aim to reduce the likelihood that notes are politically biased. This rapid evolution means it is premature to make sweeping judgements about particular measures or to dismiss their usefulness in the long run.

8. **Attribution measures should be used in combination to tackle the greatest range of deepfakes.** Rather than view measures like watermarking and metadata as substitutes, we should instead see them as complementary safeguards, whose collective deployment increases the probability of curtailing the spread of deepfakes. Model developers and platforms should also remember there are other actions they can take besides introducing attribution measures. This includes using AI classifiers to detect deepfakes and employing red teaming methods to stress test AI models. Indeed, alternative interventions like these may be more effective than attribution measures at combatting the spread of certain deepfakes like non-consensual sexual content.[37] Our first paper on deepfakes summarises this wider universe of measures.

[37] Attribution measures that indicate whether content is synthetic, on their own, won't address the harm posed by non-consensual deepfakes.

# Evaluating open-source watermarks

## Rationale and objectives

This Annex details our method for testing the durability of watermarked images, using publicly available watermarking tools. The purpose of this evaluation was to assess how three publicly available watermarks performed in response to a range of real-world image transformations (edits), so we could understand their robustness in the wild.

To achieve this, we put ~700 watermarked images through a series of test scenarios. These tests were grouped into two main categories:

1. Incidental alterations that are applied to images online, for example, many social media platforms automatically compress images that are shared on their platforms.
2. Deliberate attempts by a skilled user to actively try to remove the watermark from the image.

## Our method

### Generative model and core dataset

To create a controlled and reproducible testbed, we utilised the following open-source components:

- **Generative model:** The MirrorDiffusion Model (MDM), an open-source model from the Georgia Institute of Technology, was employed as the base for image generation. The model was used for image-to-image tasks.
- **Training dataset:** To fine-tune and enhance the generative capabilities of the model, we used the PASS (Pictures, Art, and Sketches with Semantics) dataset from the University of Oxford. This dataset comprises 1.4 million licensed images, intentionally excluding identifiable human subjects.

### Watermark implementation

Our evaluation was conducted on datasets derived from a single, consistent source: a base corpus of ~700 images produced by the diffusion model. This ensured all watermarking techniques were tested against the same set of images. We prepared sets of watermarked content to reflect different implementation scenarios:

- **In-model watermarking:** The diffusion model was fine-tuned, guided by a pre-trained watermark detector, to embed a watermark directly during the image synthesis process.
- **Post-hoc watermarking:** Starting again from the original, non-watermarked corpus of ~700 images, we then applied two distinct post-hoc watermarking techniques, applying the watermarked directly to the pixel data of the pre-generated images.

### Evaluation framework

The utility of any watermarking system is defined by its ability to withstand manipulation. Our framework centred on testing robustness against a suite of attack vectors (real-world image transformations).

### Primary evaluation metric: Watermark Detection Rate (WDR)

To quantify robustness, we employ the Watermark Detection Rate (WDR). The WDR is the percentage of images in which the watermark remains successfully detectable by the corresponding detector algorithm after a given transformation has been applied. A WDR of 100% indicates perfect robustness against a specific attack, while a WDR of 0% signifies complete failure of the watermark to survive that attack.

### Attack vectors: Simulating real-world threats

We applied both simpler and more complex transformations to the images, to mirror techniques than an average user and more competent agent could use to remove a watermark. We then deployed the watermark detectors on the altered images to assess whether the watermark could still be detected.

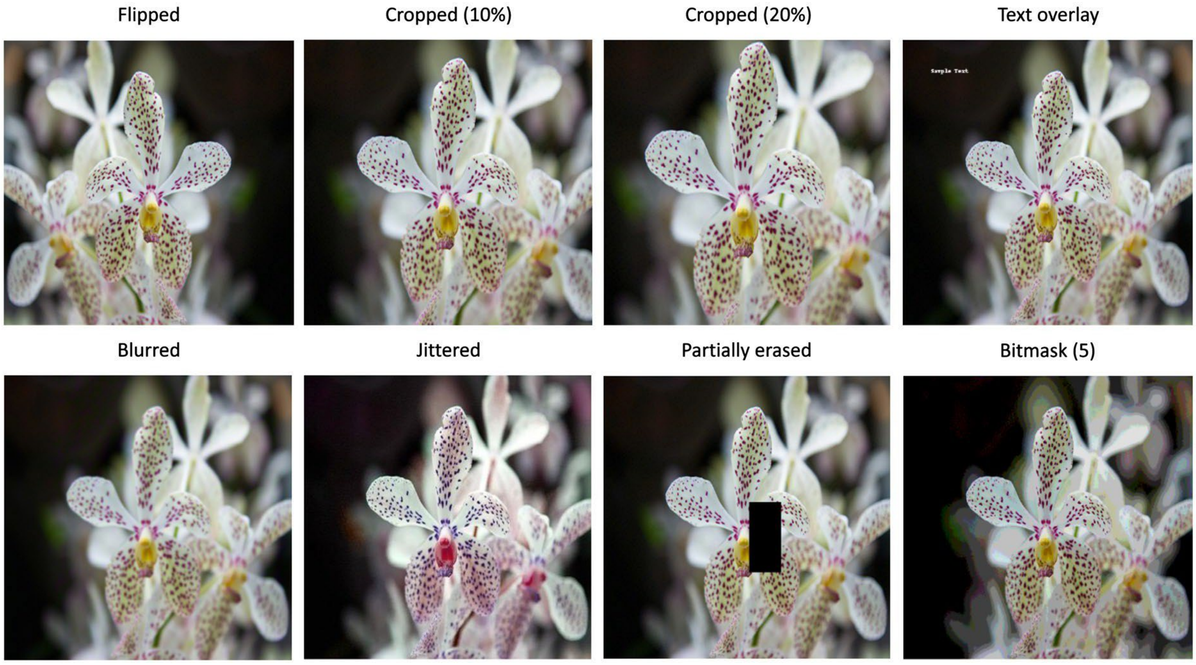
## Image transformations (edits)

We applied 26 distinct transformations to the ~700 watermarked images. Within several of the 26 transformation types, we undertook further tests which resulted in almost 100 image attacks (for example, we performed different intensities of blurring).

- **Blurred**– Adjusts individual pixels to be more similar to their neighbours, the effects of which are comparable to viewing an image through a translucent screen (technical term: gaussian blur).
- **Brightness adjusted** – Changes the luminosity of an image.
- **Compression** – Reduces the size of an image file by selectively discarding certain data, commonly applied to uploaded photos to save on the space required to store them.
- **Contrast adjusted** – Changes the tonal range of an image.
- **Cropped** – Cuts away parts of an image, usually the parts at the edge of the image, often used to fit an image into a certain space requirement.
- **Part erased** – Replaces a small section of the image by a black square.
- **Flipped** – Flips the image horizontally.
- **Gamma adjusted** – Brightens or darks shadows in the image (it roughly corresponds to brightness (<1) or darkening (>1)).
- **Greyscale adjusted** – Converts an image into greyscale, whereby all colours are changed to shades of grey.
- **Hue adjusted** – Adjusts the hue (i.e. colours) of an image.
- **Jittered** – Makes a small, random adjustment to the brightness, contrast, saturation and hue of an image.
- **Normalised** – Normalises an image per a mathematical function, usually causing extreme visual distortion to the human eye, for example, with bright and clashing colours.
- **Perspective adjusted** – Applies random perspective transformations to an image.
- **Resized** – Changes the size of an image's pixels (our evaluation resized images to 224x224 pixels).
- **Rotated** – Turns an image at various angles. Commonly used in photo editing applications.
- **Saturation adjusted** – Adjusts the saturation of an image: Low saturation leads to 'washed out' colours, with high saturation increasing the vividness and intensity of the colours.
- **Screenshotted** – Captures an on-screen image, causing distortions like resolution and colour shifts. Screenshots are often taken by users to capture content they see online.
- **Sharpness adjusted** – Adjusts the sharpness of an image, making the features in an image more or less well defined.

- **Text overlaid** – Adds text over an image.
- **Bitmask 0**[38] – Alters the least significant digit ('bit') in each pixel. For Bitmask 0, no pixels are altered.
- **Bitmask 1** – As above, but 1/8 pixels are altered.
- **Bitmask 2** - As above, but 2/8 pixels are altered.
- **Bitmask 3** - As above, but 3/8 pixels are altered.
- **Bitmask 4** - As above, but 4/8 pixels are altered.
- **Bitmask 5** - As above, but 5/8 pixels are altered.
- **Bitmask 6** - As above, but 6/8 pixels are altered.
- **Bitmask 7** - As above, but 7/8 pixels are altered.
- **Bitmask 8** - As above, but 8/8 pixels are altered, making the image completely black.

**Figure 9: Examples of transformations applied to the images**



# Our results

Our findings revealed vulnerabilities to multiple image transformations. These findings are supported by academic research and testing.

We have chosen not to include details of how the three watermarking tools performed against each type of content edit, to avoid providing information that could assist a user to remove an invisible watermark from an image.

[38] Pixels in the tested images are made up of three sets of 8 bits (digits which can only be 0 or 1), representing the amount of red, blue and green in the pixel. 'Bitmasking' is defined as setting the last bit of each set of the pixel to 0. The last bit of each set corresponds to only minor changes in the colour of the image, in a way which is not detectable to the human eye. Bitmask-3 involves changing the final three bits in a pixel to 0, so where a pixel's red value has a value of 10101110, applying Bitmask-3 changes the value to 10101000. This means that Bitmask-0 does not change the image, while Bitmask-8 renders the entire image black. See: https://www.mdpi.com/2076-3417/12/9/4202#:~:text=Least%20Significant%20Bit%20(LSB)%20Embedding,Embedding%20with%20message%20bytes

# Limitations

These experiments were run by a small team of experts at Ofcom using publicly available documentation, simulating the approach an organisation would take to test and implement a watermarking tool. There may be undocumented optimisations which could have improved our method. The vulnerabilities we identified can be understood in the following context:

- Our evaluations were limited to open-source, publicly available watermarking tools. The performance of proprietary, closed-source watermarking tools or different types of watermarking tools may yield different results.
- Our evaluations focussed primarily on the robustness of the watermark's presence (a 'zero-bit' detection). We did not assess the payload capacity (i.e. the number of bits that can be reliably extracted).
- While we confirmed that the watermarks were imperceptible via visual inspection, we did not perform a quantitative analysis of perceptual quality.
- The evaluations were conducted on a corpus of ~700 images and a defined set of 26 transformations. While this captured a diverse range of transformations, it does not represent the infinite spectrum of potential image edits. More complex attacks, or transformations not included in our suite could yield different results.