European Commission

Technical Solutions for Marking and Detecting AI-Generated Image and Video Content in the Context of Article 50(2) AI Act

Prof. Dr. Mario Fritz

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

EUROPEAN COMMISSION

Directorate-General for Communications Networks, Content and Technology (CNECT)
Directorate A — Artificial Intelligence Office
Unit A.2 — Regulation and Compliance

Contact: *Nandi Robijns*

E-mail: *nandi.robijns@ec.europa.eu*

European Commission
B-1049 Brussels

. *Technical Solutions for Marking and Detecting AI-Generated Image andf Video Content in the Context of Article 50(2) AI Act*

Directorate-General for Communications Networks, Content and Technology (CNECT)

Manuscript completed in October 2025

Luxembourg: Publications Office of the European Union, 2026
KK-01-25-158-EN-N
ISBN 978-92-68-34423-1
doi:10.2759/9763153

Abstract

This study provides a comprehensive review of technical solutions for ensuring compliance with the transparency obligations under Article 50 of the Artificial Intelligence Act, with a focus on synthetic image and video content. It maps the current state of the art in marking and detection technologies, including metadata-based approaches such as content credentials, cryptographic signatures, digital watermarks, and fingerprinting techniques. These are examined in the report through both academic research and industry applications, with illustrative examples of their effectiveness, reliability, and practical deployment.

The analysis assesses these solutions against the core requirements set out in the AI Act. It is required that outputs of the generative AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Additionally, technical solutions must be effective, interoperable, robust and reliable as far as this is technically feasible, taking into account vulnerable groups due to age or disability. The analysis finds that while several techniques demonstrate promise in terms of robustness and scalability, none provide a complete solution across all contexts. Trade-offs between effectiveness, interoperability, privacy preservation, accessibility, and implementation costs are identified as critical considerations. Solutions must also demonstrate applicability to multimodal systems and ensure interoperability with standards being developed for other content types such as text and audio.

Contents

Chapter 1

Introduction

The European Union's Artificial Intelligence Act (AI Act) places a strong emphasis on transparency for AI-generated and -manipulated content. In particular, Article 50 establishes obligations for both providers and users of certain AI systems to ensure that people are informed when they encounter AI-generated or AI-manipulated content. Article 50(2) requires that providers of AI systems capable of generating images, videos, audio or text (including general-purpose AI models) mark synthetic content with an appropriate identifier that is machine-readable and detectable. The aim is to enable automated tools and human users to recognize AI-generated content, thereby preventing deception and disinformation. Additionally, Article 50(4) mandates that deployers of AI systems that create "deepfakes" – artificial or manipulated audiovisual content that impersonates real people or events – clearly disclose the artificial origin of such content when it is presented to others. Article 50(4) paragraph 2 similarly requires disclosure for AI-generated content in the context of news or media intended to inform the public, with limited exceptions. To support these transparency requirements, Article 50(7) mandates the development of a Code of Practice on AI transparency, which will detail practical measures and technical solutions to implement content marking and disclosure at scale. This study has been prepared to support the European Commission in facilitating the next steps of implementing Article 50. It provides a review of the state-of-the-art technical solutions for marking and detecting AI-generated image and video content. The scope is limited to visual content (images and video). The study focuses on real-world solutions (including academic research) and emerging standards. By mapping existing techniques and evaluating them against the

AI Act's criteria of effectiveness, interoperability, robustness, reliability, accessibility, the study aims to inform best practices and identify gaps where further innovation or standardization is needed.

The remainder of this report is structured as follows. Chapter 2 surveys the landscape of technical solutions for marking AI-generated visual content and for detecting such content, grouping solutions into broad categories: cryptographic signatures and content credentials; metadata-based identifiers; digital watermarking; and AI-based content fingerprinting/detection. For each category, we explain the underlying principles, technical assumptions, typical implementation approaches, levels of deployment, and the responsibilities of various actors (AI model and system providers, content publishers (like social media), end-users, regulators) in making the solution work. It also provides real-world examples from research prototypes and industry deployments to illustrate how these solutions operate and their proven effectiveness where data is available. Chapter 3 reviews relevant technical standards and standardization efforts, highlighting how the industry is improving certain protocols and where gaps remain. In Chapter 4, a structured assessment evaluates each solution category against key requirements derived from Article 50 and general AI governance principles – including effectiveness in marking/detecting AI content, interoperability across platforms, robustness against tampering or circumvention, reliability (accuracy of detection and low error rates), accessibility and usability for creators, users and regulators, and privacy or security impacts. This assessment brings out the trade-offs inherent in different approaches. Chapter 5 discusses practical considerations and best practices, examining how combinations of solutions can be employed to achieve greater transparency, how to balance burdens between large and small AI providers, the role of users and regulators in the verification infrastructure, and international approaches that can inform the EU strategy. Finally, Chapter 6 and Chapter 7 concludes with recommendations and future work – including priority areas for research and standardization to address current limitations and prepare for emerging challenges in AI content transparency.

Chapter 2

Technical Solutions for Marking and Detecting AI-Generated Visual Content

In this chapter, we map and describe the main technical solutions presently available to mark AI-generated images and videos (so that they can be identified as such) and to detect AI-generated or AI-manipulated content. "Marking" typically refers to measures implemented at the time of content generation or publication to embed an identifier or provenance information into the content. "Detecting" refers to post-hoc analysis techniques that can indicate whether a given piece of content is AI-generated, even if no explicit mark was provided. Many of these solutions are complementary: a robust transparency regime may employ both proactive marking and reactive detection mechanisms. We categorize the solutions into four groups:

- Cryptographic Signatures and Content Credentials – attaching provenance metadata (e.g. information about how an image or video was created [1]) secured by cryptographic signatures.
- Metadata Identifiers (Non-Cryptographic) – simpler approaches to label content via metadata or labels that are not cryptographically secured.

[1] While in practice, such additional information might be relevant - the requirements in Article 50 are only concerned with the information if the content is AI generated or not

- Digital Watermarks – embedding hidden signals or patterns into the pixels/frames of images or video that can later be detected, typically in a way invisible to human viewers.
- AI Output Fingerprinting and Passive Detection – techniques that leverage the inherent characteristics or artifacts of AI-generated content (or train AI detectors to recognize them).

For each category, we explain how it works and give examples of real-world implementations from academia or industry. It should be noted that these categories can overlap – for instance, some solutions use a combination of metadata and watermarks.

2.1 Cryptographic Signatures and Content Credentials

One of the most secure ways to mark AI-generated content is to attach tamper-evident metadata to the content file, and to protect that metadata (and optionally the content itself) with cryptographic mechanisms such as hash codes and digital signatures. The basic principle is borrowed from the concept of digital document signing and content provenance tracking. A cryptographic hash of the image or video is computed and stored, along with relevant metadata, and then digitally signed by a trusted party. Any subsequent alteration of the content or the metadata would invalidate the signature, alerting verifiers to potential tampering. In this way, anyone with access to the public key of the signer can verify that the content is authentic (i.e., it was produced by a known source and has not been altered since). If the metadata includes an indication that the content was AI-generated, this provides in principle a reliable transparency mechanism: the content carries a cryptographically verifiable signature describing its creation.

A leading example of this approach is the work of the Content Authenticity Initiative (CAI) and the Coalition for Content Provenance and Authenticity (C2PA), a cross-industry effort to establish a standard for content provenance metadata. The C2PA standard [19] defines a data format (often called "content credentials") that can be embedded in common image or video file formats. This metadata can record information such as who or what AI system created the content, when it was created, and what edits have

been made. Crucially, C2PA metadata includes hashes of the media content and is sealed with digital signatures to prevent undetected modification.

For example, if an image is generated by an AI model, the generation software could embed metadata indicating the model name, a label that it is AI-generated, and a cryptographic signature from the provider. If someone later tries to alter the image or strip out the metadata, the signature verification will fail. Users or platforms can then be alerted that the provenance is broken. If the metadata remains intact and the signature verifies, one can trust the information about how that image was generated. This provides a high level of integrity assurance. While beyond the requirements of Article 50, each piece of content can carry a chain of signed "manifests" describing its history, from initial creation to subsequent modifications, enabling full transparency of origin.

Real-world implementation of cryptographic content credentials is underway. Adobe, one of the founders of the CAI, has integrated a Content Credentials feature into tools like Photoshop [4] to allow creators to attach provenance data to images upon export. Truepic, a startup specializing in image authentication, has worked with Qualcomm to implement secure capture of photos on smartphones [61], wherein the camera hardware immediately signs images with authenticated metadata (e.g., time, location, device) at the moment of capture. These projects demonstrate the feasibility of creating an end-to-end provenance chain: from the AI generator or camera capturing the content, through any editing software, to the platform where it's published, the content can retain a verifiable log of its creation and modifications.

In 2023–2024, major companies including Adobe, Microsoft, Intel, Arm, BBC, and Truepic collaborated in the C2PA to publish an open standard for these content credentials. Tech companies are beginning to adopt it: for instance, Google joined the C2PA and in 2024 announced that it is incorporating Content Credentials into Google Search and its advertising products. In Google's "About this image" feature, users will be able to see if an image has C2PA metadata indicating it was created or edited by AI [27]. Cloudflare, a large internet infrastructure provider, also launched a service to integrate Content Credentials at the network level, allowing web images to be verified; the system attaches a digital metadata tag tracking who created the image and whether generative AI tools were used [18].

The theoretical strengths of this approach lie in its security and clarity. The use of standard public-key cryptography means that, assuming private keys are well managed, it is practically infeasible for an adversary to forge

the provenance metadata or alter content without detection. The approach is also very general: it can be applied to any content type (images, video frames, audio files, etc.) as long as a container format supports attaching the metadata. It does not degrade the visual quality of the content in any way (since all information is in metadata sidecar). The assumptions are that there exists a chain of trust – typically a certificate authority or trust list of approved signers – so that verifiers know which signatures to trust.

Indeed, the C2PA ecosystem is developing a trust list of reputable issuers so that a platform can validate, for example, that an "image taken by CameraModel X" claim is signed by the manufacturer of CameraModel X and thus likely genuine. The practical implementation involves updates to software: AI content creation tools must be equipped to generate and attach the signatures; image hosting sites and social media platforms need to not strip out the metadata (as many currently do for size or privacy reasons), and instead preserve it; and user devices or apps may need capability to read and display the provenance info to users.

The "level of implementation" can vary: in an ideal scenario it is end-to-end (from creation to consumption), but it could also be implemented just at certain stages (for instance, a news organization might add a signature when publishing an image even if previous steps did not, in order to vouch for authenticity from that point onward).

The responsibilities of actors in this scheme are as follows:
AI content providers (generators) should incorporate provenance signing in their generation pipeline (e.g., the company providing a generative image API signs each output, or the software used by an artist to create AI art signs the file). Note that there is some tension between signing the content in a trusted manner and not conflicting with the privacy of the content creator. In particular, the EU AI Act does not require data of the author/creator.
Distributors and platforms (like social media, news websites) should support passing through this metadata and ideally also verify it and act on it (for instance, flag content with broken or missing credentials as potentially unverified). They might also add their own signature when curating or publishing content, extending the chain of provenance.
Users are not burdened heavily here except to use tools that display the information; however, media literacy is needed so that users understand what a verified content credential means. Other information systems need an appropriate integration, too.
Regulators or auditors might maintain or endorse the trust lists of reliable

signers and could build reference verification tools to assist in enforcement (e.g., a regulatory body could sample content and check for proper credentials as part of compliance audits).

Illustrative Example: Suppose an image is generated by a text-to-image AI model. The service providing the model (say, an online platform) automatically embeds Content Credentials metadata stating "Generated by AI – Model X, version Y, at time Z," and signs this metadata with its private key. The user downloads the image and later posts it on a social network. The social network's system detects the C2PA metadata; it verifies the signature against a list of known providers (recognizing the model provider's public key) and confirms the data integrity. The platform then could, for example, display an icon or a note "AI-generated image (verified)" on the post. If someone tries to modify the image in an editor which does not preserve the metadata, or crops it and re-saves without credentials, then the signature no longer verifies or is lost – the platform could then either show "source unverifiable" or treat it without the authenticity badge. This example shows how cryptographic content credentials can facilitate user disclosure while providing a high level of trust in the accuracy of the label.

Despite these strengths, the C2PA approach has several notable limitations identified also by recent analyses [42]. Key issues include:

- **Unverified provenance data:** C2PA explicitly does not require verification of the embedded metadata. In other words, a strong signature may simply seal whatever information the author provides, even if that information is false. Thus, cryptographic signing alone does not ensure the truth of the provenance fields.

- **Complexity and lack of transparency:** The C2PA standard is complicated and was developed largely behind closed doors. It uses several layered data formats (like JUMBF, CBOR, XMP, and JSON) together with a digital signature. The main open-source tool implementing it (c2patool [20]) is large and depends on many external libraries. Experts note that some of these layers seem to duplicate functions without clear justification. Because the design process was not open to public review, it is hard to understand the reasons for many choices, and this may create problems for interoperability.

- **Certificate and trust issues:** C2PA uses standard X.509 signing certificates, but these introduce practical problems. Acquiring a signing

certificate can be expensive, and recent policy changes (e.g., Adobe's "known certificate" permit list) make the system vendor-centric and opaque. Moreover, the C2PA specification currently ignores critical certificate validity checks: it permits using expired or even revoked certificates without invalidating a signature. This undermines the intended non-repudiation guarantees.

- **Privacy and dependence on vendors:** In C2PA, the metadata can be stored separately from the content in so-called "sidecar" files. To verify such files, users often need to connect to a server run by a company or certificate authority. This means the host can see who is checking which content, raising privacy concerns and creating dependence on specific vendors. By contrast, newer approaches such as the Secure Evidence Attribution Label (SEAL) [41] use a decentralized system. SEAL stores public keys in the internet's domain name system (DNS), allowing verification without contacting a central server—protecting privacy and reducing vendor lock-in.

- **Metadata size and scalability:** C2PA credentials tend to be large. A typical metadata block contains nested structures and at least one X.509 certificate. In practice, C2PA metadata often exceeds 10 KB, sometimes larger than the media content itself. This overhead imposes storage and bandwidth costs; many image hosting services strip out large metadata blocks to save space. By contrast, SEAL signatures are very compact and leverage existing metadata fields.

In summary, while cryptographic content credentials (exemplified by C2PA) can in principle provide strong integrity and provenance guarantees, they currently can have several conceptual and implementation related issues as outlined above. Research and standards bodies emphasize that addressing these issues—for example, by simplifying the architecture or decentralizing trust—is critical for scalability.

Emerging alternatives such as the Secure Evidence Attribution Label (SEAL) aim to make content attribution simpler and more transparent. Instead of introducing complex new file structures, SEAL works by digitally signing the metadata that already exists within the file. It uses a decentralized approach for verification, publishing public keys through the Domain Name System (DNS). This means anyone can verify the signature directly,

without relying on a central authority or proprietary service—reducing complexity, improving transparency, and supporting open, privacy-preserving verification.

2.2 Metadata Identifiers without Cryptography

A simpler but less secure method of marking AI-generated content is to include identifying information in the metadata of the image or video file without necessarily using signatures. Most digital image formats (JPEG, PNG, etc.) and video containers (MP4, MOV) allow embedding of metadata fields such as descriptions, tags, or custom data blocks. An AI content generator or editor could populate a specific metadata field with a value indicating the content is AI-generated. For instance, the software might set the "Software" field to "AI Generator v1.0" or add an XMP tag like `<SyntheticMedia>true</SyntheticMedia>`.

Some industry groups have discussed standard metadata flags for synthetic media – for example, the International Press Telecommunications Council (IPTC), which maintains widely used photo metadata standards, introduced a property to indicate if an image is original, edited, or entirely computer-generated. A provider could simply set "DigitalSourceType = ComputerGenerated" in the IPTC metadata of an AI image to label it as such. The advantage of this approach is its simplicity and low implementation cost. It does not require any complex cryptographic infrastructure or coordination. Even a small startup or an individual developer of an AI image tool can implement a metadata tag in the output files with minimal effort. It is also human-readable in many cases – for example, a user checking the file properties might see a comment "This image was AI-generated using X tool."

Because Article 50(2) specifically mentions a machine-readable format, these metadata tags can be designed to be both human and machine-readable. They can be easily detected by software scanning content (for example, a content management system could automatically parse metadata for the "AI-generated" flag). However, without cryptographic protection, the trustworthiness of metadata identifiers is limited. Anyone with basic tools can edit or remove metadata from a media file. Malicious actors could strip out the "AI-

generated" tag or even falsify it (for example, adding "AI-generated:false" to an image that is actually fake to disguise it).

Thus, as a standalone solution, simple metadata labels rely entirely on voluntary compliance by content creators and on the assumption that downstream platforms do not tamper with the metadata. They are therefore best suited to cooperative scenarios – e.g., a reputable AI content provider labeling their outputs to help users, with no adversarial intent – but they are not robust against intentional misuse. In terms of Article 50's criteria: such labels are interoperable to the extent that metadata standards are universal (any system that can read JPEG metadata can retrieve the tag), but not reliably effective or robust if an adversary decides to remove or alter them (which is trivially easy).

Nonetheless, there are real-world instances of this approach. Some stock photo websites that allow AI-generated images (such as Shutterstock) have started including an indication in the metadata or filename that the image is AI-generated, as part of their content policies. In the context of video, some creators of deepfake videos (for research or satire, posted on platforms like YouTube) will voluntarily add disclaimers in the video description or even in the video frames (like a subtitle "This video is synthetic"). These are essentially metadata or perceptible labels applied as good practice. The responsibility for implementing this lies primarily with content providers/generators – it requires them to be transparent. Platforms can assist by preserving these metadata fields on upload (rather than scrubbing metadata, which is sometimes done for privacy), and by perhaps making the information visible or searchable. Regulators could consider standardizing a particular metadata schema for AI content to encourage uniform adoption. This is indeed being discussed in standardization circles adjacent to C2PA for cases where full cryptographic signing is not feasible.

In summary, non-cryptographic metadata labels are a low-barrier, immediate solution to at least tag content. They serve as a basic form of transparency but should be seen as a supplement to more robust methods.

2.3 Digital Watermarking of AI-Generated Content

Watermarking offers a more proactive approach [2, 40, 73, 16, 62, 30, 15, 57]. Watermarking embeds a hidden, resilient signal directly into the output during generation. This signal—while imperceptible to humans—remains detectable later, offering a consistent mechanism for verifying the AI origin of content [65, 7, 35].

Unlike passive detection methods, watermarking does not rely on identifying statistical anomalies, which become increasingly elusive as models evolve. Instead, it provides an intentional and model-aware method of content attribution, supporting greater transparency and accountability in the deployment of GenAI systems.

The process typically involves three main components: an embedder, a key, and a detector (or decoder). The embedder inserts a hidden signal (the watermark) into the content, often during or after generation, using a secret key that defines how and where the watermark is placed. The watermark is typically designed to be imperceptible to humans but detectable by the detector, which uses the same or a related key to verify whether the content carries the watermark or to extract the hidden message. A well-designed watermark should remain detectable even after common modifications like compression, resizing, or format changes, while preserving the visual quality of the content.

The following analysis is aligned with recent surveys on the topic [74, 14]. Watermarking schemes can exhibit a variety of important properties, with the relevance of each depending on the intended application. They are detailed as follows:

Quality Considerations in Watermarking A key aim of any watermarking method is to keep the quality of the generated content as high as possible. Some methods only test this through experiments or user feedback, while others can prove that the watermark barely affects quality for any input.

Quality can be checked for single images (to see if each looks good) or across many images (to make sure there's no hidden bias or pattern). For instance, a watermarking method might accidentally make a model produce more pictures of dogs than cats—each picture looks fine on its own, but

the overall set shows bias. That's why it's important to test not just single examples, but also collections of outputs, to ensure fairness and consistency.

Undetectable Watermarks A watermarking method is called undetectable if it's practically impossible to tell whether content comes from the original model or from a watermarked one—unless you have the secret key. In simple terms, no one can spot or remove the watermark just by looking at the output.

This idea first came from text models but now also applies to images. When a watermark is undetectable, it means the content's quality and performance stay the same as usual, and the watermark doesn't affect how the model works. It also makes it very hard for anyone to guess or copy the secret keys used for adding or detecting the watermark, which protects it from tampering or forgery.

False Negative Rate and Robustness An essential property of any watermark detection algorithm is maintaining a low false positive rate. This ensures that content created independently of the watermarking scheme is rarely, if ever, incorrectly identified as watermarked.

Importantly, this definition applies uniformly across all content and does not rely on assumptions about the distribution of natural or human-generated data. This distribution-agnostic formulation enhances the detector's robustness and trustworthiness—ensuring that no particular type of content is unfairly targeted or misclassified.

The *false negative rate* measures how consistently a watermark can be detected in unaltered content generated by a watermarking scheme. This depends not only on the watermarking method itself but also on the variability of the model's outputs.

Robustness refers to a watermark's ability to remain detectable even after content has been modified—intentionally or unintentionally. Real-world content may undergo transformations such as paraphrasing in text or compression in images. A robust watermark should survive these changes.

Defining robustness formally is more involved than defining false positives or false negatives, due to several complexities:

- Robustness must be considered with respect to a specific *channel*, which models possible alterations or attacks.

- The strength of the attacker (or channel) may vary depending on whether they have access to keys or detection mechanisms [34, 36, 51].
- Robustness depends on the content having enough entropy, meaning natural variation or randomness that allows small, hidden changes to be added without altering quality. If a model always produces the exact same output (is deterministic), there's no variation to hide a watermark in—so embedding one reliably without distortion becomes impossible.

Unforgeability Some watermarking methods add marks using a secret key, but a stronger guarantee—called unforgeability—means that a watermark can only be created if someone actually has that key. This prevents others from faking or copying the watermark. However, making a watermark both unforgeable and very robust can be difficult: if it's too tolerant of small edits, even slightly changed content might still look "valid." To balance this, systems can use a strict version of the watermark for secure verification and a more flexible one for general detection.

Message Embedding Capability Watermarks that can encode specific messages within generated content are known as *multi-bit watermarks*. This functionality is particularly valuable in scenarios that require fine-grained traceability—for example, embedding metadata such as model version or user identifiers. These features are essential for managing accountability in large-scale deployments involving multiple models and users.

The amount of information that can be embedded depends on the *entropy* of the generated content. High-entropy outputs—such as images, audio, or video—allow more flexibility and can support longer embedded messages. Text, by contrast, generally offers lower entropy, limiting current text watermarking schemes to only a few bits of embedded data.

Computational Efficiency For watermarking to be practical, the generation process must remain efficient. Ideally, generating watermarked content should not be significantly more resource-intensive than standard generation in terms of computation, memory, or speed. Excessive overhead can hinder deployment, especially in production-scale systems.

Similarly, the procedures for detection, decoding, and attribution should be lightweight to support efficient, real-time verification in operational environments.

Threat Model Understanding the resilience of watermarking schemes requires analyzing potential threats posed by adversaries. Following [74], we categorize threats based on the attacker's goals and their level of access or capabilities. We identify three primary goals that an attacker might pursue:

- **Watermark Removal.** Here, the attacker aims to alter AI-generated content such that it evades detection or leads to incorrect decoding, all while preserving content quality. Typical removal techniques include light modifications such as noise injection or resizing in images, e.g. [7, 75].
- **Watermark Forgery.** In forgery attacks, the adversary seeks to produce content that appears watermarked, despite bypassing the official watermarking process. This can lead to the wrongful attribution of content, undermining system credibility. Forgery methods may involve mimicking known watermark patterns or exploiting knowledge of the watermarking algorithm to embed counterfeit markers.
- **Secret Key Extraction.** A more advanced attack aims to recover the secret keys used in watermarking. While not required for removal or forgery, success would allow an attacker to fully compromise the scheme. This type of attack targets the underlying cryptographic integrity of the watermark.

Adversary Capabilities The threat posed by an adversary depends on the level of access and computational power they possess. In addition, several key factors include: access to different data types (generators, watermarked/non-watermarked content, knowledge of the watermarking algorithm, control over key for watermarking, access to verification, access to shadow model (similar watermarking schemes). These factors help define the strength and feasibility of potential attacks, guiding the design of more robust watermarking schemes. As a lesson from cybersecurity, methods that are still secure despite additional knowledge of the adversary (e.g. watermarking algorithm) are preferable.

Evasion Attacks Evasion attacks attempt to remove a watermark from generated content while preserving its perceived quality. We identify three main types:

- **Edit Attacks:** These involve localized, small-scale modifications—such as deleting characters in text, applying noise to images, or replacing words with synonyms. Such transformations simulate noisy communication channels. In more advanced settings, attackers with access to the watermarking model or detector may use optimization techniques to construct adversarial examples [34, 36, 64].

- **Regeneration Attacks:** In this strategy, watermarked content is passed through a secondary generative model—such as a paraphraser, summarizer, or denoising system—to regenerate similar content without the watermark [69, 52, 50, 75, 55]. These transformations are again modeled as noisy channels.

- **Downsampling Attacks:** Here, a subset of the watermarked output is extracted, effectively removing the watermark. For example, the *Emoji Attack* inserts special tokens during generation, which are removed afterward to disrupt watermark patterns. A similar concept could apply to image watermarking, such as cropping watermarked regions.

Forgery Attacks Forgery attacks aim to produce content that falsely appears watermarked, even though it was never passed through the official watermarking process. These attacks often require no access to watermark keys. Some methods reverse the logic of evasion: instead of degrading watermark signals, they construct outputs to mimic watermark characteristics. For example, [29, 36] show that adversaries can approximate the behavior of the red-green watermarking scheme [40] using labeled examples and then generate content with high watermark scores. In a similar vein, [55] train a surrogate model on both watermarked and clean images to produce adversarial examples—crafted using techniques like Projected Gradient Descent (PGD) [46]—that trick the watermark detector into false positives. These forged examples often generalize well across detectors, raising concerns about spoofing in practical deployments.

2.3.1 Image Watermarking Schemes

Similar to text watermarking, image watermarking methods can be classified into **post-processing** and **in-processing** techniques. Post-processing

methods, traditionally preferred for their broad applicability, embed watermarks into images after generation. These techniques often operate in the frequency domain, leveraging transformations such as Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT) [5], and DWT-DCT-SVD [48]. Furthermore, deep learning-based encoder-decoder architectures, including HiDDeN [77], RivaGAN [70], StegaStamp [60], and SSL Watermarking [22], employ neural networks to efficiently embed and extract watermarks. Conversely, in-processing methods integrate watermarking directly into the generative model or the sampling process, inherently embedding watermarks within the generated images. Here, we focus on the more sophisticated in-processing watermarking techniques for image generation. Below, we discuss notable approaches in this category.

Stable Signature. Stable Signature [23] introduces watermarking into the generation process by fine-tuning model parameters without modifying the model's architecture. It adapts a pre-trained Latent Diffusion Model (LDM) [54] to ensure that all generated images encode a specific binary signature. A pre-trained watermark extractor retrieves the signature, followed by a statistical test to verify its authenticity. The approach consists of two main steps: (1) pre-training to extract binary messages, and (2) fine-tuning the LDM decoder to embed a fixed signature into all outputs. Similar strategies, such as DiffusionDM [76], also modify diffusion models to incorporate watermarks. However, these techniques often suffer from degradation in image quality, reduced robustness against regeneration attacks, and a lack of theoretical guarantees [7].

Tree-Ring Watermark. The Tree-Ring watermark, proposed by Wen et al. [62], modifies the latent sampling distribution in LDMs and employs an inverse diffusion process for detection. This method constrains concentric rings in the Fourier domain of the latent space to zero. Detection is performed via DDIM inversion [59] to estimate the initial latent, with the watermark deemed present if the estimated latent exhibits notably small values in the watermarked rings. Subsequent refinements of this heuristic latent pattern have been explored [72, 17]. However, the Tree-Ring watermark significantly distorts the Gaussian latent distribution, resulting in lower image quality and reduced variability. While robust against certain attacks, it remains susceptible to adversarial surrogate attacks since the latent pattern can be

easily learned by neural networks [55].

Gaussian Shading Watermark. The Gaussian Shading watermark [66] embeds a watermark by restricting latent space sampling to a predefined quadrant determined by a watermarking key. During detection, the latent is recovered, and its proximity to the watermarked quadrant is assessed. Yang et al. [66] claim "lossless performance," supported by a proof that the distribution of an individual watermarked image remains indistinguishable from that of an un-watermarked image. However, this proof does not account for dependencies across multiple generated images, which are essential for evaluating diversity metrics such as FID [32], CLIP Score [53], and Inception Score [56]. In practical applications, the same random watermarking key is used to generate multiple images for quality assessment. Since all images originate from the same quadrant in latent space, this method inherently reduces variability, adversely affecting diversity metrics.

PRC Watermark. The PRC watermark [30] introduces an undetectable watermarking approach tailored for latent diffusion models. Similar to the Gaussian Shading watermark, it employs structured sampling in latent space, but instead utilizes a pseudorandom error-correcting code (PRC) to dynamically select a fresh quadrant for each image generation. A key advantage of this approach is that its undetectability ensures that image quality remains uncompromised, even when evaluated across multiple generations using metrics such as FID, CLIP Score, and Inception Score. Furthermore, the robustness of the pseudorandom code directly translates into the robustness of the embedded watermark. Additionally, this technique enables message encoding within the watermark itself.

Steganography-based approach. Steganography-based approaches [68] employ a novel watermarking mechanism (which in [68] is referred to as "fingerprinting") that draws conceptual inspiration from steganography. Unlike traditional image steganography, which embeds hidden information directly into pixel-level modifications of carrier images, the paper embeds unique watermarks into the parameters of deep generative models themselves. This approach ensures that every image generated by a fingerprinted model subtly contains a unique, imperceptible watermark that can be later decoded to trace its origin. The technique relies on modulating the convolutional

filters of the generator network using a fingerprint embedding, allowing the generation process itself to encode identifying information in a non-intrusive and robust manner. This shift from direct image manipulation to generator-level fingerprinting marks a fundamental departure from traditional steganographic methods, offering enhanced scalability, secrecy, and resilience against common perturbations.

2.3.2 Video Watermarking Schemes

Video watermarking aims to embed imperceptible information into video content for purposes such as copyright protection, content authentication, and tamper detection [10]. Traditional methods have focused on spatial, frequency, and hybrid domains. Recently, deep learning has shown great promise in addressing challenges like robustness and adaptability, especially in dynamic video environments.

Traditional Video Watermarking Techniques Traditional video watermarking techniques are typically categorized by their embedding domain. They are post processing techniques, as they have not been developed with or integrated in a generator. In the spatial domain, information is embedded directly into the luminance or chrominance values of the video frames. Methods such as Least Significant Bit (LSB) replacement have been widely used, despite their vulnerability to attacks [8]. To improve robustness, techniques like Middle Significant Bit (MIDSB) [9] and Intermediate Significant Bit (ISB) modifications have been introduced. Frequency domain approaches, on the other hand, rely on transformations such as the Discrete Cosine Transform (DCT)[33], Discrete Wavelet Transform (DWT)[11], and Singular Value Decomposition (SVD)[6] to embed watermarks into transform coefficients. Hybrid approaches combine spatial and frequency domains to harness the benefits of both [3]. With the rise of compressed video formats, watermarking techniques have also been developed that embed information directly into compressed video streams, such as those conforming to MPEG, H.264, and H.265 standards [21, 26].

Deep Learning-Based Video Watermarking Deep learning has introduced new paradigms in video watermarking by enabling models to learn complex embedding and extraction patterns. A typical deep learning water-

marking system comprises an encoder that embeds the watermark, a distortion simulation module, and a decoder that recovers the watermark [10]. The involved simulation hardens the system against typical perturbations and changes. Architectures based on Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) are most prevalent. CNN-based techniques such as DVMark [45] and methods introduced by Bistron and Piotrowski [12] employ multi-scale feature extraction and entropy-based embedding to enhance robustness. GAN-based models, like RivaGAN [71], incorporate adversarial training to increase watermark imperceptibility and resistance to attacks. These attacks can include cropping, color adjustments, re-encoding, or other editing operations intended to make the watermark undetectable. RivaGAN proposed an architecture that employs a critic and adversary network to ensure the quality and security of the embedded watermark. The model includes an attention mechanism that identifies optimal regions for embedding. DVMark by Luo et al.[45] uses a multi-scale encoder-decoder framework and includes a GAN-based discriminator to differentiate between marked and unmarked frames. Bistron and Piotrowski [12] developed a CNN model that uses entropy-driven mapping for embedding, enhancing robustness without significantly affecting perceptual quality. Gao et al.[25] proposed a zero-watermarking scheme that utilizes CNNs in conjunction with a self-organizing map (SOM) and Polar Complex Exponential Transform (PCET) to avoid modifying the original video content. In the compressed domain, Kaczyński and Piotrowski [37] introduced a method that uses deep neural networks and an adjustable subsquares property algorithm to embed watermarks during H.265/HEVC encoding. The decoder then retrieves the watermark post-decompression, preserving high video quality. A novel approach using mosaic images derived from video frames was introduced by Mansour et al. [47]. This technique applies image watermarking strategies to a synthesized mosaic representation of the video, thereby enhancing robustness to attacks like collusion and compression. Similarly, Ke et al. [38] incorporated curriculum learning and attention mechanisms within a GAN framework to dynamically adapt the embedding process based on training difficulty.

Unfortunately, for the most part, video watermarking has been developed for copyright detection and has not undergone testing for the suitability for genAI transparency. VideoSeal [24] is an open-source neural watermarking framework designed for efficient and robust video watermarking – and is one of the few method that are dedicated to watermark video genAI con-

tent. It introduces a novel technique called *temporal watermark propagation*, which enables watermark embedding only every k frames and propagates the watermark distortion to the remaining frames.The method jointly trains an efficient U-Net-based embedder and a transformer-based extractor using a multistage pipeline that includes image pre-training, hybrid video training, and extractor fine-tuning. This makes it especially suitable for watermarking AI-generated videos, where imperceptibility and resilience to editing and compression (e.g., H.264) are critical. Extensive experiments show that VideoSeal outperforms existing image watermarking models repurposed for video in terms of robustness under geometric transformations, compression, and their combinations. While the watermark is mostly imperceptible, minor artifacts may appear in flat regions or under rapid motion. Limitations include potential visual artifacts at high step-sizes in temporal propagation and the need for further refinement in temporal perceptual metrics. Nevertheless, its high performance, open-source availability, and practical efficiency make VideoSeal a strong candidate for watermarking AI-generated video content.

Google's SynthID has been presented to also be applicable for marking generated AI content, but up to now more detailed information is missing. This makes the technology difficult to judge and assess.

Conclusion Deep learning has significantly advanced the field of video watermarking, offering new capabilities in robustness, invisibility, and adaptability. While most deep learning efforts have concentrated on image watermarking, only recent developments have been dedicated to video watermarking. Researchers are also encouraged to develop techniques robust against video-specific attacks such as collusion, frame swapping, and re-encoding.

Key challenges remain in balancing imperceptibility, robustness, and efficiency. Most methods are evaluated on short clips, but real-world deployment demands fast embedding/extraction for high-resolution, long videos. Combined distortions (e.g. heavy compression plus geometric transforms) still break many systems – e.g. even VideoSeal reports degradation under extreme warping. Open-source benchmarks and tools are scarce, though efforts like VideoSeal's public code are addressing this. Finally, as digital media proliferates, watermarking must keep pace with new threat models (e.g. deepfakes) and legal standards. In summary, recent video watermarking research blends traditional signal-processing insights with end-to-end deep learning, yet future work must further improve attack resilience and practical deployment.

2.3.3 Taxonomy and Tradeoffs

As an intermediate summary, we can conclude that while there are promising candidates for image and video watermarking, these technologies are in development and have not yet seen large deployment. At this stage, providing specific recommendations is difficult. Therefore, in the following section, we present a taxonomy – aligned with previous surveys [13] – that categorizes different classes of watermarking schemes and highlights those with more desirable properties. We conclude by outlining key metrics that will be essential for evaluating current and future watermarking approaches, as these can inform best-practice recommendations.

Direct Replacement The simplest method of embedding a covert watermark involves directly modifying specific elements of the media content. In image and video files, this typically means altering selected pixels or frequency coefficients. One widely known example is least significant bit (LSB) manipulation, where the least significant bits of pixel values are overwritten with the watermark payload. This approach is computationally efficient and easy to implement, requiring minimal overhead.

However, the simplicity of direct replacement makes it especially vulnerable to modification. Common operations such as compression, cropping, filtering, or even minor resizing can destroy the watermark. Moreover, its predictability allows adversaries to easily target and overwrite the modified regions. While suitable for some controlled environments, this method lacks the robustness and security required for more adversarial or public-facing applications.

Hashing and Encryption To introduce greater resilience and secrecy, many watermarking systems incorporate hashing or encryption into the process. In this context, cryptographic or perceptual hash functions generate pseudo-random values that determine how and where watermark data should be embedded. For example, a watermark encoder might compute a hash of the content—or a portion thereof—and use the resulting value to select specific frequency coefficients for modification.

Encryption-based watermarking methods further enhance security by using secret keys to encrypt the embedding pattern or the watermark payload itself. This ensures that only entities with the appropriate decryption key can detect or validate the watermark. In practice, these methods are highly

effective at resisting unauthorized tampering or spoofing, provided the keys are managed securely.

Nonetheless, these approaches often increase computational cost and complexity. Additionally, the robustness of hash-based schemes can be brittle if the watermark relies on fragile content-dependent features, which may change during benign transformations.

Randomness Randomness-based perturbation techniques employ stochastic processes to determine how the watermark is embedded into the content. Typically, a private random seed or cryptographic key is used to drive pseudo-random number generators that select embedding locations or define acceptable output ranges. In diffusion-based image or video generation models, randomness may be introduced at the noise initialization stage, ensuring that only model outputs conditioned on the specific seed will contain the watermark.

This strategy significantly complicates watermark removal by unauthorized actors, as detection depends on access to the original random seed. The stochastic nature of the embedding also makes it harder for adversaries to infer consistent patterns across watermarked outputs.

However, randomness-based methods generally require private operation. Without access to the seed or the generator, detection is impossible, which restricts public verification. Furthermore, managing and securing randomness inputs—especially across distributed systems—can introduce logistical challenges.

Machine Learning-Based Perturbation The most advanced watermarking strategies leverage machine learning to learn how to perturb content in ways that are both effective and resilient. These systems typically involve training a neural network to encode watermark signals directly into media representations. A corresponding detector—often another learned model—is trained to recognize these signals even after the content has been modified through compression, editing, or transformation.

Examples of such systems include encoder-decoder frameworks that embed robust, noise-tolerant signatures into image or video frames. Google's SynthID [28] is one such tool that seems to show strong performance in embedding persistent, invisible watermarks into AI-generated images. As mentioned above, at this point in time, insufficient information is available

for the excact methodology of SynthID on images and videos.

Machine-learned watermarking techniques offer high robustness and flexibility. Because they can learn complex mappings between watermark signals and perceptual space, they outperform traditional hand-crafted perturbations in resisting removal and spoofing. However, these systems are not without drawbacks. They require large training datasets, considerable computational resources, and may struggle to generalize across media types or domains not seen during training. Additionally, the lack of transparency in learned embeddings can make it difficult to validate or audit the watermarking process.

Conclusion As the domain of watermarking is taking shape, there is no clear winner. As it seems unlikely the formal garuantees can be given, it looks like this domain will remain in a state of arms race / cat and mouse game. Hence, continuous testing and monitoring are important along the outlined dimension of importance. The situation is complicated as the current literature (a) has no unified testing benchmarks (b) most of the work is academic and lacks large scale, real-world test beds (c) there is little to no experience with methods in deployment and what actual attack might be faced by these methods. Nevertheless, the adaptability of machine learning based methods seems a strong plus. Furthermore, it can be easily imagined that it can achieve a super set of the other methods during the learning/adaptation/optimization procedure. In this way, future adversaries can be accounted for or adapted against once known.

2.3.4 Evaluating Watermarking Techniques for AI-Generated Content

Evaluating watermarking techniques for AI-generated content requires a comprehensive framework grounded in practical, measurable metrics. While conceptual properties such as robustness or fidelity are important, rigorous assessment must translate these into quantifiable benchmarks that reflect real-world performance and constraints. In this section, we present a methodology for evaluating watermarking schemes based on five core categories: fidelity, detectability, robustness, capacity, and security. Each category is accompanied by metrics that can be objectively computed using established evaluation protocols.

These evaluation criteria can be linked to the key requirements of Article 50 of the EU AI Act, which emphasizes effectiveness, reliability, robustness, interoperability, and accessibility in marking AI-generated or AI-manipulated content:

- **Fidelity → Reliability and Accessibility.** Fidelity measures how well watermarking preserves the original quality of the content. Watermarking methods should not degrade usability or perceptual quality as best as possible. Such high fidelity ensures that markings do not distort or reduce the accessibility of AI-generated content, maintaining user trust and experience.

- **Detectability → Effectiveness and Reliability.** Detectability corresponds to how consistently a watermark can be recognized in AI-generated content. This directly supports Article 50's requirement that markings be *machine-readable and detectable*. High detectability ensures that tools can reliably identify AI-generated or manipulated media.

- **Robustness → Robustness (Article 50 term).** Robustness assesses whether a watermark remains detectable after processing operations (e.g., compression, cropping, or format conversion or even adversarial manipulation). This aligns precisely with Article 50's requirement for robustness, ensuring that markings persist through normal use and resist tampering or removal.

- **Capacity → Interoperability and Scalability.** Capacity refers to the amount of information that can be embedded within a watermark, such as model identifiers or timestamps. While Article 50 only requires a clear "AI-generated" label, higher capacity enables interoperability with other systems and standards, supporting future extensions of transparency frameworks.

- **Security → Reliability and Integrity.** Security evaluates protection against forgery or manipulation of watermarks. This supports Article 50's objective of ensuring reliable and trustworthy marking by preventing false or misleading labels and preserving the integrity of the verification process.

Fidelity Fidelity refers to the extent to which watermarking alters the quality or intended semantics of the generated content. This must be evaluated quantitatively to ensure the watermarking process does not degrade the utility of the content. For images, common metrics include the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which measure pixel-level distortion and perceptual similarity, respectively. In generative contexts, Fréchet Inception Distance (FID) is particularly relevant; it quantifies the difference in feature distributions between generated and real images. Lower FID scores indicate better visual quality. Human evaluation may be included but should be supported by standardized Likert-scale annotation with inter-annotator agreement statistics to ensure consistency. In order to put the metrics roughly in perspective: Fidelity should be considered acceptable when PSNR exceeds 30 dB for images, SSIM approaches 0.95 or higher, and FID remains within 5-10 points of the unwatermarked baseline. But exact numbers may vary between application domains.

Detectability Detectability captures the reliability with which a watermark can be detected when present, and conversely, the ability to reject non-watermarked content. Detection performance should be measured using standard binary classification metrics. True Positive Rate (TPR) and False Positive Rate (FPR) quantify the system's sensitivity and specificity. Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) provide a robust summary of the detection trade-offs. However, the definition of the test distribution is critical to how meaningful the score is.

Robustness Robustness refers to the watermark's resilience to common post-processing transformations and adversarial modifications. To evaluate robustness, watermarked content should be subjected to perturbations including JPEG compression (with quality factors of e.g. 75 and 50), Gaussian blur, resizing, cropping (e.g., 80% central crop). After each perturbation, detection performance should be re-measured using TPR and FPR. Robust watermarking schemes should maintain high detection AUC on all tested transformations. The average degradation in detection accuracy across perturbation types can be computed as a robustness score.

Security Security evaluates the watermark's resilience to intentional tampering, forgery, or removal by adversaries. It is key that for a given watermarking scheme a precise threat model is formulated. Potential attacks include white-box and black-box attack settings. In a white-box setting, the adversary has access to full information of the watermarking scheme. It is preferable and recommmended by best practices in the security domain, that the watermarking scheme remains secure and effective even if the algorithm and methodology is known to the attacker. In a black-box setting, the adversary does not know the internals, but may e.g. try to train surrogate models or generate counterfactuals that evade watermark detection. Additionally, the system should be immune to false attribution attacks, where content is manipulated to appear as though it carries a specific watermark. This is measured by a false positive rate under attack. It is debatable what attacker strength is reasonable to assume. It is conceivable that more sophisticated attacks are possible in the future and broadly available even for low-skill users. An advanced removal attack would constitute applying another genAI to resynthesize the content and thereby remove the watermark. This should also be tested and monitored how broadly available such tools become. But unfortunately it is difficult to conceive a method that is completely robust to such resynthesis attacks.

2.4 AI-Generated Content Fingerprinting and Passive Detection

Even in the absence of any explicit mark or metadata, AI-generated content can sometimes be identified by analyzing the content. This is often termed *passive detection* or *forensic analysis*. It relies on the premise that generative AI models leave statistical fingerprints or anomalies in the outputs that differ from real content. For example, early deepfake videos often had noticeable irregularities in face blinks or lighting inconsistencies. AI-generated images might have characteristic noise patterns or spectral signatures (e.g., how pixels correlate) that differ from those produced by natural photographs. Researchers have developed numerous AI detection algorithms – typically machine learning classifiers – trained to distinguish real vs AI-generated content. These detectors take an image or video and output a prediction of whether it is AI-created [67, 31, 49, 44, 39, 63]. One approach is to exploit known

"fingerprints" of specific generation methods. Academic studies have shown it is possible to identify not only if an image is AI-generated but even which model (DALL-E vs. Stable Diffusion vs. Midjourney, etc.) likely produced it, by training classifiers on images from those models – essentially learning each model's fingerprint. Another approach is more generic: training a neural network on a large dataset of real images and a large dataset of AI images (from various sources) to learn generalizable differences. For deepfake videos, many detectors use deep neural networks to analyze facial movements or pixel-level artifacts frame by frame. The research community and global efforts have invested in such detection – for example, the Facebook Deepfake Detection Challenge in 2020 spurred development of video deepfake detectors. Agencies like the U.S. DARPA (Media Forensics program) and companies have produced deepfake detection tools (e.g., Microsoft's Video Authenticator). These are continually evolving as new types of fakes emerge. However, the consensus is that passive detection is inherently challenging and can be brittle. The U.S. Government Accountability Office (GAO)'s assessment in 2024 [1] highlighted the limitations of current approaches and that results not necessarily translate to real-world settings and that changing the conditions or the generation method significantly degrades detector accuracy. For example, a detector trained on deepfakes with a certain resolution might fail on higher-resolution fakes; a detector looking for a specific artifact might be fooled if the next-gen model doesn't produce that artifact. Attackers can also take an AI-generated image and post-process it (slightly blur or add random noise) to confuse detectors – a kind of adversarial attack. Thus, while detection algorithms improve, AI generation techniques also improve to produce more photorealistic and detector-evasive output. Nevertheless, passive detection remains a critical piece for enforcement and oversight because it seems unlikely that all AI content will be marked – in particular when bad actors and adveraries are involved.

For the AI Act's transparency requirement to be enforceable, authorities will need tools to catch instances where someone used AI to, e.g. generate disinformation and deliberately removed any watermark/metadata. In such cases, an AI forensic detector might be the only line of defense. It can flag suspicious content for further human review or even trigger an investigation. Detectors can also be used in a monitoring capacity by platforms: for instance, a social media site might automatically scan uploaded videos with a deepfake detector; if the score is high, and the video is purportedly of a political figure, they might temporarily label or limit it pending verification.

From an implementation perspective, deploying passive detection involves machine learning models which can be computationally intensive (especially for video, which might require analyzing many frames). Large platforms with resources can integrate these into their content moderation pipelines. Smaller platforms or organizations might rely on external services or open-source models. There are already companies offering deepfake detection services to businesses, which indicates a market for this technology. Malicious deployers of AI content are obviously not incentivized to use detectors on themselves, but legitimate users could use detectors for self-regulation (e.g., a news outlet verifying if a user-submitted photo might be AI-generated). Platforms and intermediaries are a key actor: they can integrate detectors to enforce their policies (for example, not allowing undisclosed deepfakes). Regulators and law enforcement will likely need their own advanced detection capabilities or access to those of others, in order to audit compliance with Article 50(4) (which bans undisclosed deepfakes except narrow exceptions).

In conclusion, AI output fingerprinting and passive detection are reactive solutions that complement proactive marking. They are essential for identifying unmarked AI content and for verifying suspicious media. Their effectiveness can be high in controlled tests but tends to drop in adversarial settings, making them less reliable as a sole measure. They also raise the issue of false positives – incorrectly flagging genuine content as AI-generated can be harmful (e.g., accusing a real video of being fake could undermine truth). Therefore, careful calibration and likely a combination of multiple detection signals (including checking for absence of credentials or watermark) should be used. In addition, passive detection can have an important role in monitoring and enforcing of regulations.

Chapter 3

Standards and Emerging Standardisation Efforts:

A number of technical standards and industry efforts are emerging around the transparency of AI-generated content. Standardization is crucial for interoperability – one of the key requirements in Article 50(2) is that the technical solutions used for marking synthetic content be interoperable, meaning different tools and platforms can recognize and act on the marks. Here we outline some of the most relevant standards and initiatives:

Coalition for Content Provenance and Authenticity (C2PA) Standard: As introduced earlier, the C2PA has published a comprehensive technical specification for content provenance metadata (Content Credentials). This standard covers how to embed manifest data into various file formats (images: JPEG, PNG; video: MPEG, etc.) and how to sign and verify that data. The C2PA specification is open and royalty-free, intended as a global standard. Since its initial release in 2021, it has undergone revisions to improve security and cover more use cases. The standard is being trialed in products (e.g. [4]). Continued work in C2PA focuses on ease of implementation and alignment with other standards like IPTC.

IPTC Photo Metadata Standards: The IPTC, which maintains widely-used standards for image metadata (such as EXIF and XMP schemas used by cameras and news media), has been working on definitions to indicate AI involvement. In 2022, IPTC introduced an extension to identify "synthetic media" in metadata. This is not a full provenance solution, but a

standardized label that could be included in image metadata to signify AI generation. Using such a standard field would ensure that different software all write/read the label consistently. IPTC's standard is complementary to C2PA, as IPTC metadata fields can be part of the signed content credentials.

Content Authenticity Initiative (CAI): While not a "standard" itself, the CAI (led by Adobe and partners) functions as an advocacy and coordination body to drive adoption of content provenance standards. It also provides user experience guidelines – for example, how to display content credentials info to users in a meaningful way. The CAI's work has influenced standardization in that it funnels requirements from creative industries and news organizations into C2PA's technical spec.

W3C and Web Standards: The World Wide Web Consortium (W3C) has communities looking at verifiable media and data authenticity. One relevant effort is the development of Data Integrity and Verifiable Credential standards (JSON-LD and Linked Data Proofs) which could, in theory, be applied to media files. For instance, a W3C verifiable credential could be issued for a piece of content asserting it is AI-generated. While not specific to AI, these standards might be leveraged for additional provenance frameworks, especially in web contexts (like a HTML tag indicating content authenticity).

Media Formats and MPEG/JPEG: ISO/IEC Moving Picture Experts Group (MPEG) and JPEG committees have begun considering standards around media authenticity and deepfake detection. The JPEG committee launched an initiative called "JPEG Fake Media" to address creation and detection of manipulated images. This could result in standards for embedding signals in images (watermarks or sidecar data) and for evaluating detection methods. MPEG, responsible for video coding standards, has discussed standards for forensic metadata in video streams (for instance, an MPEG-4 standard might one day allow an authentication track within the video). While these are early-stage, the fact that traditional media standard bodies are involved means future codecs and formats might natively support AI transparency features.

Emerging Standard for AI Watermarking: Unlike metadata, watermarking is less mature and does not yet have standards, partly because

companies have proprietary methods. It could however be important to standardize the interface: for example, defining a standard way to express "this content contains a watermark of type X" so that detectors know which algorithm to apply. In the future, industry might agree on a baseline watermarking method that could be open-sourced or licensed openly, ensuring interoperability of detection. Google's SynthID approach and other academic proposals are promising approaches, but there is significant work to be done for standardization – if possible at all. As of now, this is an open area for standardization.

Voluntary Codes and Guidelines: Apart from technical standards, there have been multi-stakeholder guidelines such as the Partnership on AI's "Responsible Practices for Synthetic Media" which encourage labeling AI content. These are not technical specs, but they often mention the types of techniques discussed (watermarks, metadata) and set expectations. The US Government's voluntary commitments extracted from AI companies in 2023, which included watermarking, can be seen as a political standard – it sets a norm that could translate to technical implementation. Similarly, China's regulatory requirement for deepfake watermarks effectively sets a de facto standard within that jurisdiction (and Chinese companies have to implement some form of watermark, which could become uniform if mandated by their regulators). Tracking these implementations can inform EU standardization: e.g., if a particular watermark scheme becomes prevalent in Chinese apps, the EU might consider if it's compatible with what we encourage here, or at least ensure detectors can handle content from abroad.

In summary, the standards landscape is evolving. The most concrete standard available now is the C2PA for provenance metadata, but it also has several short comings as outlined above. Meanwhile, there is a gap in standardized watermarking – currently it is more proprietary, though commonalities exist as outlined above.

Chapter 4

Assessment of Solutions Against AI Act Requirements

Relevance of Privacy and Security. While Article 50 of the EU AI Act focuses on transparency, effectiveness, robustness, reliability, interoperability, and accessibility, privacy and security remain essential underlying principles for any content marking system. First, privacy is a cross-cutting obligation throughout the AI Act and EU law, notably under the GDPR. Marking or provenance systems that process personal data—such as information about creators, tools, or platforms—must therefore ensure data minimization, user consent where needed, and protection against tracking or profiling. Security is equally critical because the trustworthiness of transparency mechanisms depends on safeguarding keys, metadata, and verification processes against tampering or misuse. Weak security could allow forged or removed markings, undermining reliability and transparency—the very goals of Article 50.

In practice, privacy and security support the spirit of Article 50: they ensure that content identification systems are not only effective and interoperable, but also trustworthy, safe, and respectful of users' fundamental rights. Accordingly, privacy and security are included in this chapter alongside the Article 50 criteria in the assessment of solutions."

4.1 Cryptographic Signatures and Content Credentials

Effectiveness When properly implemented, cryptographic content credentials reliably convey authenticity and AI-origin information. A valid credential stating "AI-generated by X" can be trusted with near certainty, since any change to the content or metadata breaks the signature. This approach directly fulfills the transparency requirement. However, overall effectiveness hinges on widespread adoption. Unsigned content cannot be flagged by this method. Critically, because C2PA does not mandate independent verification of metadata, malicious actors could sign content with false claims; the system will trust the signature without questioning the underlying data.

Interoperability In principle, C2PA's open standard enhances interoperability: multiple vendors can read and write the same credential format. Signed images remain valid JPEG/PNG files and can be processed by ordinary tools, and the same framework can span images, video, audio, and text. In practice, however, interoperability could be hindered by vendor-specific controls. If credential verification requires consulting a proprietary permit list, then the ecosystem risks fragmentation. By contrast, SEAL's DNS-based model requires no special formats or vendor APIs, ensuring broad compatibility.

Robustness Cryptographic signatures are unforgeable under current cryptography, providing robustness against direct forgery. However, they are fragile against content transformations: even minor edits (cropping, compression, or format conversion) can break the signature. A determined adversary can easily strip the credential by modifying the media slightly. Thus, the mark is tamper-evident rather than tamper-proof: it detects tampering but does not prevent removal.

Reliability When a C2PA signature verifies, it provides a clear binary indicator with essentially no false positives or negatives. This makes it extremely reliable for confirming a declared provenance. However, the long-term reliability of C2PA depends on how well its cryptographic keys are managed and revoked. Each digital signature relies on private keys that must remain secure and valid over time. If these keys are lost, stolen, or not properly

revoked after compromise or expiration, malicious actors could create or verify false provenance data. Currently, C2PA provides limited mechanisms for transparent key renewal, expiration tracking, or revocation checking. This weakens trust in the system over the long run, since users may not be able to tell whether a signature remains valid or whether the signing authority is still trustworthy.

Accessibility and Usability. For content authenticity systems to be effective, the information they provide must be understandable and accessible to all users, including persons with disabilities. This means that provenance data—such as whether an image is AI-generated—should be displayed through clear, consistent, and perceivable user interfaces (e.g., with text alternatives, screen reader compatibility, and high-contrast visual indicators). While the C2PA standard defines how to embed and verify content credentials, it does not specify how these should be presented to users. As a result, accessibility depends heavily on how platforms and applications implement the interface. Larger technology providers may have the resources to build accessible, integrated displays of provenance data, but smaller developers or open-source projects may struggle due to C2PA's technical complexity and resource demands. In contrast, simpler and more lightweight schemes—such as SEAL—could be easier to integrate across a wider range of tools and devices, supporting greater inclusivity and ensuring that transparency features reach all users, regardless of technical ability or disability.

Privacy and Security C2PA credentials raise privacy concerns, as detached metadata validation often requires contacting a third-party server. Embedding creator identities via certificates could enable unwanted content linkage. In contrast, SEAL's decentralized design offers better privacy guarantees. Additionally, C2PA metadata size increases storage and bandwidth burdens, which could disincentivize full adoption, especially on platforms sensitive to performance and cost.

In summary, cryptographic content credentials offer very strong authenticity guarantees when used, but their real-world efficacy depends on broad ecosystem support and careful trust management. Emerging alternatives like SEAL may help address critical shortcomings, particularly regarding privacy, simplicity, and resilience against vendor control.

4.2 Metadata Labels (Non-Cryptographic)

Effectiveness: Simple metadata labels are only as effective as the honesty and consistency of those who apply them. In cooperative environments, they achieve the basic goal of informing that content is AI-generated. However, because they can be removed or altered unnoticed, their effectiveness in an adversarial context is low. If someone wants to deceive, they will not leave the label intact. Thus, these labels might help for casual transparency (a good actor informing users), but they fail as a robust mechanism to guarantee transparency when faced with bad actors or just the entropy of the internet (metadata often gets lost through conversions). For instance, if an image with an "AI=true" tag is uploaded to a platform that strips metadata (which many social networks do to save bandwidth), the label is gone when other users see it.

Interoperability: On the positive side, using standard metadata fields can be highly interoperable. Anyone who can parse that field can get the info. If IPTC or another standard body defines a common field for "AI-generated", and this is adopted, then cameras, editors, and platforms globally could use it. Even without formal standards, a de facto approach (like using the EXIF Comment field with a certain keyword) could work across many systems. So interoperability is potentially high if agreed upon. But because there's no cryptographic lock, different providers might use different wording or fields. One might put it in EXIF:ImageDescription, another in XMP under a custom namespace, leading to fragmentation. A particular recommendation of metadata field for basic labeling could help to remedy this, so that at least industry uses it uniformly until a more secure method is in place.

Robustness: As mentioned, it is not robust at all against deliberate removal. It's also not robust through transformations – even innocent ones. E.g., if you take a screenshot of an AI-generated image, the screenshot is a new file with no metadata, so the label doesn't carry over. Converting file formats (JPEG to PNG) might drop metadata too. Thus, any change in the content packaging can drop the label. This lack of robustness is a major drawback for using metadata labels as the main solution.

Reliability: The reliability in terms of correctness of the label depends on who writes it. There's no verification, so someone could lie in the metadata. An automated detector may incorrectly flag content as AI-generated. While metadata is usually provided by the creator, and is reliable when creators are truthful, you cannot depend on it universally, since the overall coverage is unknown. For users, if they see a label "AI-generated" in metadata, that's clear, but often users won't see metadata without special action. And if a label is not present, it doesn't reliably mean "not AI" – it might just have been removed or never added. So absence of label tells little. That's a reliability problem: it's a one-way indicator (presence might mean AI, but absence is inconclusive).

Accessibility and Usability: The big issue is that metadata is typically hidden from end-users in normal viewing contexts. Unless the platform explicitly displays it (which, if they were going to, they might prefer the stronger content credentials approach anyway), an average person will not inspect image metadata. So for user disclosure, relying on a metadata tag that is not exposed in the UI is insufficient. Platforms could read the tag and then add their own visible notice ("This image was labeled by the creator as AI-generated"), which would improve accessibility. Without that, these labels mostly benefit power users or automated tools scanning content. It might be more useful in B2B or internal workflows (e.g., a news org receiving images could have a script that flags any with "AI-generated" in metadata).

Privacy and Security: There are minimal privacy concerns since it is just a flag – unless additional identifying info is stored. In terms of security, someone could maliciously tag a real image as "AI-generated" to discredit it. With cryptographic credentials, that scenario is less likely because the signer identity is attached, but with open metadata, misinformation could also be done by tampering with tags. Trade-offs: The main trade-off of metadata labels is simplicity vs. strength. They are extremely easy and cheap to implement (a plus for small players and quick deployment), but they offer weak guarantees. Over-reliance on this in the long term would not meet the spirit of "robust and reliable" solutions. However, it might still have a place as a recommended practice in addition to stronger methods, for example, "Always include an 'AI-generated' metadata tag for human readability even if you also apply a watermark or signature."

4.3 Digital Watermarking of AI-Generated Content

Effectiveness: Watermarks can be very effective in enabling detection of AI-generated content provided that the content indeed has a watermark embedded. When present, a well-designed watermark like SynthID allows automated systems to detect the AI origin with high confidence. In tests, such watermarks can be detected even after common edits, meaning they effectively "travel" with the content. Therefore, if major AI generators watermark their outputs, a large portion of AI images/videos in circulation could become identifiable. The effectiveness is limited primarily by coverage (not all generators may implement the same watermark or any watermark) and by the arms-race issue (some watermarks might be defeated by adversarial transformation). As of state-of-art, robust watermarks can survive simple manipulations but more extreme manipulations (cropping out significant parts, heavy recompression, adding noise, or even resynthesis) might break or obscure them. Still, even in such cases, a detector might at least flag uncertainty. One measure of effectiveness is: how often do we get false negatives (missing a watermarked image)? If a watermark is robust, false negatives will be low for normal transformations – e.g., Google reported SynthID had strong detection rates post mild processing. False positives (detecting watermark in an image that is actually not watermarked) should be near zero if the watermark is well-chosen (the pattern space is huge, so accidental presence is extremely unlikely, and detectors are tuned to avoid misidentification). Thus watermarks can be a highly precise way to tag AI content. However, they don't directly inform the user unless detection is performed – unlike a visible label, the watermark has to be checked by software. So their effectiveness in user transparency depends on integration of detectors into user-facing apps or sites. Unlike cryptographic signatures which fail when modified, a removed watermark yields an image that looks fine and just doesn't trigger detection. That could fool an enforcement system unless they also have a generic AI detector as backup. In practice, an enforcement might have to say: if you remove the mandated watermark, you're in violation – but catching that requires either noticing the watermark is gone (hard if you don't know it should have been there) or just catching them through other means. This highlights that watermark mandates likely need to be combined with legal deterrents for tampering.

Interoperability: Currently, watermarking approaches are not standardized, so interoperability is an issue. If Company A uses Watermark A and Company B uses Watermark B, their detectors are different. A platform would need to support multiple detection algorithms to cover all marks. This is doable (e.g., run a suite of detectors), but could quickly become impractical. An interoperable solution would be one watermarking scheme adopted widely, or at least an agreement that each watermark scheme's detector will be made available to others. In the voluntary commitments in the US, companies did not specify using one method, only that each would watermark. One could foresee a clearinghouse or a framework where model providers register their watermark method with a trusted body so that law enforcement or platforms can obtain the detection method. Over time, maybe a few dominant schemes will emerge. Another interoperability aspect is multimodal: can the same watermarking approach extend to both images and video? Images are easier (single frame). Video watermarking can leverage image techniques on each frame but also possibly embed an ID across frames. There's no fundamental barrier to doing it in each domain but one scheme might not directly transfer. Google's SynthID is currently image-specific, while for text they developed a different method (SynthID-Text). So we might have to accept different standards per content type (one for images, one for audio, etc.). However, from a user perspective or compliance perspective, these differences can be abstracted away – as long as each content type has some watermark. Interoperability is more a concern for the detection ecosystem – ensuring everyone who needs to detect can do so easily.

Robustness: This is a key criterion for watermarks. A robust watermark should survive common transformations and attempts to remove it without excessive effort. E.g. SynthID claims robustness against resizing, cropping, re-encoding. That addresses many casual edits. But consider more aggressive attacks: if someone knows the image is watermarked, they could apply an adversarial filter specifically to disrupt the pattern. For example, adding a certain kind of noise that they guess might confuse the detector, or warping the image geometrically. Robust watermarks anticipate some of these (maybe the detector searches patches or uses error-correcting codes so partial distortion still reveals it). Research in this direction is still ongoing. In summary, current watermarks are moderately robust. They raise the bar significantly for someone trying to pass AI content as human-made. The

adversary must take additional steps to scrub the watermark, which likely degrades the content or requires an additional process. This situation can reduce casual misuse. Also, if multiple layers are used (e.g., watermark + metadata), one has to remove both to fully hide origin.

Reliability: If a watermark is present and the detector is run, the result is usually reliable (with confidence scores). Good watermarks have low false positive rates, which means that one will generally not mistakenly flag normal images as watermarked. There might be rare coincidences or if an image is very distorted, a detector might be unsure and give a false negative (not detecting when watermark was there but got weakened). For instance, heavy cropping might remove so much of the pattern that detection fails. But in those cases, one could consider it effectively removed. In general, watermark detectors can be tuned to be conservative – e.g., have a threshold to only flag when quite sure, to avoid mislabeling non-AI generated content. This is important if user disclosure is automatic. It should be avoided that someone's real photo or a digitial image human generated is labeled as "AI-generated" incorrectly. With cryptographic methods, that doesn't happen (it either verifies or not). With watermarks, a slight chance exists of mis-detection. In practice, robust scheme design and thresholding can bring error rates down to very minimal – if there is no adversarial intend. We should still include in guidelines that detection results are probabilistic and ideally combined with other info if available.

Accessibility and Usability: Watermarks themselves are typically invisible, so accessibility to users depends on detection being integrated into user-facing systems. A user won't "see" a watermark unless an app tells them. So similar to content credentials, we need software to expose the info. One approach is building detection into popular image viewers or social media. For example, a messaging app could scan images as they are uploaded and if an AI watermark is found, show a small tag "AI-generated image" on it in the chat. This requires computational resources and could have privacy implications if scanning personal media. For public or web content, it is easier to justify scanning. Given that watermarks are meant to be machine-detectable, "machine-readable and detectable" is exactly what Article 50(2) calls for, so this meets that criteria in spirit. Companies that deploy watermarks should be encouraged to also provide means to detect them openly,

or at least to trusted parties. For users, an advantage of watermarks is that they do not need to trust a central authority. Any instance of the content contains the mark, so even decentralized sharing retains it. Users might not personally run detectors often, but the presence means any stakeholder can check content. That fosters an environment where it is harder to conceil the origin if detection tools are widely available.

Privacy and Security: Watermark detection usually requires analyzing the content but not referencing external data unless secret keys are needed. In some cases, detection might involve a secret key if the watermark is keyed. E.g. SynthID detection presumably uses a secret model. If detection keys are widely shared, that could leak into attacker hands. If they are not shared, then not everyone can detect. A possible compromise is to give detection ability to platforms and maybe upon request to law enforcement, but not to the general public. However, that reduces transparency, openess, and possibility to external scrutiny. From a security standpoint, methods that are open and secure are preferable in order to avoid a "security by obscurity" paradigm. Regarding privacy: scanning images for watermarks is similar to antivirus scanning files – it is typically looking for a specific pattern, not extracting personal data. It can likely be done locally or on-device to avoid sending content out. So privacy impact can be low, as long as done responsibly. If done server-side, it's part of content moderation which users usually consent to on platforms. Trade-offs: Watermarking provides a solution that is less heavy infrastructure-wise than cryptographic signatures (no public key infrastructure (PKI) required for detection), and can persist in content across transfers. The main trade-off is the possibility of removal and the need to maintain secrecy or update schemes if broken. It also does not provide any additional info beyond "AI-made" unless you encode more data into it - if the payload allows for it. Most current proposals keep it to a minimal identification. Combining watermarks with cryptographic credentials yields a more powerful combination: credentials give detailed provenance when available; if someone strips them, hopefully the watermark still allows detection. If they also remove the watermark (harder), then you rely on passive detection.

4.4 AI Content Fingerprinting and Passive Detection

Effectiveness: Passive detection algorithms have shown varying effectiveness. Under controlled conditions, when the detector is trained on similar data to what it is tested on, many detectors can achieve impressively high accuracy (80-99%) at distinguishing AI vs real. For example, a research model might correctly identify most deepfakes in a test set. However, as soon as conditions change – a new AI generation method, or a slight domain shift – effectiveness drops. The GAO report [1] notes that detection methods may not work well if the fake was created by a different method than those the model was trained on. This then is a generalization problem, and in general, the perfect representation of a cat-and-mouse game. In practical terms, detectors are useful as an additional line of defense but cannot be fully relied on. The effectiveness also degrades for lower quality content (blurry or small images are hard to analyze, yet fakes might be intentionally low-quality to hide flaws). As AI generation becomes nearly photorealistic and consistent, passive detection might approach the inherent limits of detecting statistical differences. Some experts even fear a future where AI content is practically indistinguishable without auxiliary info – increasing the importance of watermarks and provenance techniques. Nevertheless, it raises the effort required for high-quality deception since not everyone will use the latest models or know how to avoid artifacts. Also, ensemble approaches using multiple detectors and human expertise increase success. As an enforcement tool, detectors can scan large volumes and highlight potential violators for review.

Interoperability: Each detection model is essentially a proprietary or open algorithm – not a standardized output. But they usually output a score or binary decision, which could be standardized (e.g., a common confidence metric). EU, NIST and other bodies could standardize evaluation criteria (they do for facial recognition, etc.). If many actors create detectors, sharing information about performance and test datasets is key. There might eventually be standard APIs – for example, a standard for an "AI-content-detection" service interface – to allow interoperability. But currently, interoperability is not applicable in the same sense as marking. It doesn't need to interoperate across platforms, except in the sense that a detection result needs to be explainable or comparable.

Robustness: The robustness here refers to the detector's robustness to evasion. Many detectors are not robust: changing an image slightly can drop their confidence drastically. Adversarial attacks that add a carefully calculated perturbation can often fool a neural network detector while keeping the image looking the same to humans. This is a known vulnerability in ML. It is conceivable that someone generating a deepfake can also generate an adversarial noise to append that will confuse known detectors. While robust detection methods are an active research area, this aforementioned weakness can be exploited by anyone who knows or reverse-engineers the inner workings of the detector. This contrasts with cryptographic or watermark methods where if you remove them, at least you've altered the content. With adversarial attacks, one can alter content imperceptibly to humans but ruin the detector's judgment – a problematic scenario. This again emphasizes that detectors should not be solely relied on for final judgments without other evidence.

Reliability: The reliability in terms of false positive/negative can vary widely by detector and the application scenario / test distribution. The cost of a false positive – which predicts a real image as AI – is serious because it could undermine trust in real media or wrongfully accuse someone of using AI. So any deployment for user-facing labeling would have to minimize false positives, likely by having a threshold such that only very certain detections are labeled, communicate confidences/uncertainty, or require multiple methods to agree. One way to improve reliability is to use multi-factor detection: check for watermark, check for content credentials, and then use passive detection as a complement. If watermark/credentials say "AI", then this is a strong signal. If they say "not present", then a passive detector can serve as another line of defense. That layered approach can yield high reliability for decisions.

Accessibility and Usability: Passive detection results are not typically exposed to end-users except possibly as "this content is suspected to be AI-generated" labels if a platform chooses to do so. For example, some social media might label state-altered media or use detection to tag deepfakes. However, as of now, most detection is internal – like content moderation – or by third-party fact-checkers. For general users to use detection, they would need access to tools or websites where they can upload an image/video

to test. But these are not mainstream, and they require user initiative. It should be encouraged to make detection tools more available to the public e.g., a browser extension that can warn you if an image on a webpage is likely AI. But caution is needed so as not to create false alarms or require technical literacy to interpret. From a transparency standpoint, detection is reactive and does not directly tell a user "this is AI" unless someone has integrated it into the viewing experience. So similar to other solutions, integration is key for it to actually inform users. One beneficial use-case is media outlets using detectors to vet user-submitted content, thereby only publishing verified real images or clearly labeling any questionable ones. This process is invisible to the public but improves the final transparency.

Privacy and Security: Running detection on content does not itself impact privacy beyond analyzing the content which itself might be private if not public. If done on user-uploaded content, platforms should have it covered in their terms of services. For client-side tools, the analysis could be local, which is privacy-preserving. If a user sends an image to a cloud detector service, then obviously that service is exposed to the image, which has privacy implications. So for sensitive contexts, local or on-device detection might be preferable, which requires efficient models or lightweight heuristic detectors. In terms of security, detectors themselves can be targets of adversarial attacks. As mentioned above, an adversary could attempt to trick a detector into labeling real content as fake. It underscores that detection should ideally be paired with provenance so that real content can be authenticated as real too. Some advocate not just marking AI as AI, but also signing camera media as authentic – e.g., the truepic system – so detectors can know something is definitely real if signed by a camera. That two-sided approach (authenticate real and mark fake) covers both false positive and false negative issues but is more complex.

Trade-offs: Passive detection requires continuous adaptation as AI models evolve – it's an ongoing expense and research effort. It potentially puts more burden on moderators/regulators rather than on AI creators. In contrast, watermarks shift burden to creators to mark. There is some anology to virus scanners vs. software signing: you prefer software to be signed by trusted publishers (like content credentials), but you still need antivirus to catch malware from those who do not sign or try to spoof. Similarly, we require

AI content to be labeled (via metadata or watermark etc), but detection will catch the unlabeled (e.g. legacy data) as a safety net. In terms of cost, developing good detectors might be beyond small companies, but governments and large firms can provide these tools and perhaps open-source models made available to all. There should be a commitment of collaborative improvement of detection tools.

4.5 Trade-off Analysis and Summary

Each solution has strengths and weaknesses w.r.t. the AI Act requirements which are summarized in the following.

Effectiveness: Cryptographic credentials and robust watermarks are effective at positively identifying AI content when present. They embed an unambiguous signal. Passive detection is not guaranteed but fills the gaps where watermarks or cryptographic credentials are absent. Metadata labels are effective only in cooperative scenarios.

Interoperability: Content credentials have made some progress towards standardization and have seen broad industry engagement. Watermarks currently less so, but could improve with standards. Passive detection methods need sharing of knowledge to be interoperable. Metadata labels can be interoperable if standardized, otherwise fragmentation can occur.

Robustness: Watermarks and passive detection are in an arms race where countermeasures also significantly improve over time. Cryptographic methods are robust to forgery but not to removal. Watermarks add robustness against casual removal, whereas cryptographic methods fail if content is not intact. A combination could give better robustness: even if metadata is stripped, a watermark might remain.

Reliability: Cryptographic credentials are most reliable (binary, provable). Watermarks and passive detection have empirical reliability to a certain degree. Cryptographic credentials have essentially no false positives/negatives except due to due to trust issues related to the cryptographic keys. Thus for legal or formal verification, cryptographic evidence is strongest. For quick screening, watermarks/passive detectors are useful.

Accessibility and Usability: All methods require integration to be accessible. A visible label (like metadata turned into an on-screen caption) is easiest for users but has to be added by UI. Content credentials can be shown in UIs and are being integrated. Watermarks need detection integration to show something to user. Passive detection might be hidden in moderation unless chosen to be displayed.

Privacy: None of the methods inherently violate user privacy if implemented correctly. Cryptographic credentials could even enhance privacy because they reduce need for invasive detection. Yet, credentials could include too much detail and should avoid personal information unless needed. Watermarks and passive detection operate on the content itself and do not require personal data unless are run on a server.

Feasibility and Cost: Simpler metadata is relatively simple to implement, cryptographic is moderately complex (PKI infrastructure, certificate issuance, etc.), watermarking requires ML expertise to implement and possibly proprietary technology. Yet, an open and transparent system is preferable that still works in a white box setting. Passive detection requires ongoing investment in AI R&D and computing power to run at scale (cost for platforms).

In conclusion, cryptographic content credentials best fulfill the criteria of reliability and interoperability, watermarks improve robustness and coverage in the wild, and passive detection covers non-compliance and unknown cases. The trade-off essentially comes down to proactive vs reactive, and secure vs easy. The optimal strategy is to utilize the strengths of each while mitigating their weaknesses through a layered approach. The next chapter will delve into how in practice we can combine these and set good practices so that both large and small actors can contribute to an effective transparency ecosystem without undue burden.

Chapter 5

Practical Considerations and Best Practices

Having examined the tools available, we now turn to how they can be deployed in practice, considering different stakeholders and real-world constraints. The future goal should be to translate these options into actionable, balanced measures that industry players can commit to. Here we outline best practices and recommendations, keeping in mind the needs and capacities of large vs. small providers, the role of users, and enforcement by regulators. We also look at international approaches and how businesses globally are tackling AI content transparency, as these can provide examples.

5.1 Layered Transparency and Multi-Solution Approach

A recurring theme in our analysis is that a combination of solutions is more powerful than any single measure. Therefore, a best practice is to implement multiple layers of marking and detection for AI-generated content. Concretely:

AI model and system providers (creators of generative models or services) should, wherever feasible, embed both a cryptographic content credential and an invisible watermark into image/video outputs. For example, an image generator can attach a C2PA-signed manifest declaring the image AI-generated, and also encode a hidden watermark in the pixels. The

cryptographic manifest serves as the first line of transparent disclosure and is immediately verifiable by any compliant platform, while the watermark serves as a backup identifier that remains even if the metadata is lost or stripped. This dual approach means that casual distribution where metadata often remains intact, e.g., in professional workflows, carries the explicit label, and even if someone tries to tamper e.g., by screenshotting or re-saving, the watermark still tags the content. The two marks are different in nature. One is overt but removable, the other covert and robust, which makes it harder for an adversary to remove all traces.

Platforms and content distributors should implement detection and verification for both layers. That means reading and honoring content credentials, or using it internally to decide how to treat the content and also scanning for watermarks in case credentials are missing. Platforms should never strip provenance metadata unless absolutely necessary. If they must – say, to protect user privacy by removing GPS EXIF data – they could still retain the AI flag portion. If a platform finds an AI watermark in content that lacks a metadata label, it should treat it as suspect – perhaps flag it for review or automatically label it as AI-generated if confident. By doing so, platforms act as enforcers of the transparency: even if a user tried to remove the credential, the watermark reveals the truth. Many big tech platforms have the resources to implement these scanning features, and indeed some are moving in this direction (e.g. Cloudflare's integration [18], Google's plans [27]). Smaller platforms might rely on libraries or services provided by larger ones to do this (e.g., an open API where an image can be sent to get back "watermark detected: yes/no").

Regulators and authorities or independent auditors should also adopt this layered approach in oversight. For instance, if investigating a piece of content or auditing compliance of a provider, they can check for attached credentials and run forensic detectors. Regulators could maintain their own instances of detection tools potentially pooling knowledge from industry. This is important because relying solely on platforms to self-regulate might miss cross-platform issues e.g., content that spreads from one platform to another or is on private channels. In particular, authorities can use passive detection as a safety net to catch uncooperative actors. As a good practice, regulators might periodically test widely circulated media for hidden AI watermarks

or artifacts – similar to how some agencies scan for deepfakes in election misinformation monitoring. The results can inform if further action – like requiring a platform to take down or label a piece of content – is needed. This multi-layer approach aligns with the idea of "defense in depth." Even if one mechanism fails or is circumvented, others are in place. It acknowledges that the landscape includes both honest participants and adversaries, and prepares for both.

5.2 Catering to Different-Sized Providers and Models

One concern is that what's feasible for large providers may not be feasible for a small startup or an open-source model community. Broad participation should be encouraged by scaling recommendations to capacity: Large AI providers (big companies, or anyone releasing widely used generative models) should lead by example and adopt the most rigorous transparency measures. This includes implementing content credentials with official signatures, robust watermarking, and publishing information about their methods. They should also contribute to standards development and perhaps offer support to others, for example, open-sourcing parts of their watermark detection code or assisting in integrating credentials into common file formats. Their adoption can also set industry norms. If all major players watermark and sign their outputs, it becomes an expected default, and content lacking such markings might be naturally viewed with suspicion – which incentivizes even smaller players to follow suit to be trusted. Small and medium providers, including startups or individual model developers, might not have in-house cryptography or watermark specialists. For them, available tools should be used: for example, using the open-source C2PA toolkit to attach content credentials (which the CAI has made available), or using simpler metadata tags as an interim measure. Perhaps an alliance of companies or an industry association can provide a shared signing service – e.g., a centralized "AI content signing authority" where a small developer can send a hash of their output and get back a signature attesting it. Similarly, watermarking might be offered as a cloud API by larger providers. If a small company uses a big cloud's generative model, the cloud might handle watermarking automatically. Another practice is for the open-source community: those releasing open models could

provide optional add-ons that watermark outputs. Article 50(2) also explicitly includes general-purpose AI. That means models that are not specifically built to produce disinformation but can be used for many things (like image generation models available to public). The providers of such models should ensure the model or its default utilities include transparency features. For example, if someone downloads a generative model from a repository, it could come with instructions or code to watermark any image it generates. If the user fine-tunes it or uses it offline, it's harder to enforce, but at least out-of-the-box it should have transparency on. Model cards/documentation should mention if outputs are watermarked or not. Ideally, these foundation model providers voluntarily include these features by default; if not, it could be encouraged through provided measures or required as part of certification. Media and entertainment companies who may deploy AI in content production have a responsibility too. A movie studio using AI for special effects should mark AI-generated parts, a video game with AI-generated graphics likewise. Even if the content is benign, following the practice normalizes transparency. Also, these industries often have their own standards (like watermarking for copyright). They could integrate AI identification into those. Edge cases and non-commercial developers: What about someone who trains a small model at home and posts images? They are unlikely to adopt complex measures. This is where at least encouraging a culture of tagging and maybe OS-level support can help. For example, operating systems or editing apps might detect "this looks AI-generated, do you want to add a disclosure tag before sharing?" as a prompt. User-side tools could assist individuals to comply in a lightweight way. Ultimately enforcement on individuals will likely be light except in malicious cases, but providing easy tools to "label your AI creation" akin to how we encourage alt-text for images could foster voluntary uptake.

5.3 User-Facing Transparency and Communication

Even the best technical marking is useless if users are not made aware or do not understand it. So part of the implementation must be about conveying the information clearly to end users:

The AI Act will require deployers to ensure disclosure in an appropriate

manner, so platforms can operationalize that by such visual cues. Consistency in these cues across platforms will help users recognize them (like the blue check mark convention, etc.). UX (user experience) guidelines should be created for labeling AI content, informed by research on what users notice and comprehend. There should also be an "About this content" detail available. This is analogous to checking a document's properties or a webpage's certificate. Educating users that they can inspect this for suspect media is part of media literacy efforts.

User education is a best practice: stakeholders (platforms, educators, media) should educate the public that (a) AI can generate realistic fakes, and (b) there are now tools/indicators to help spot them. It is also important to clarify what an indicator means. If something is labeled as AI-generated, that does not inherently tell you malicious intent – it might be an artistic creation or an AI-assisted edit. The label is value-neutral, just informational. The context can elaborate (some platforms might allow the creator to add context, e.g., "Generated with XYZ, for art purposes"). If something is identified via detection as AI (but wasn't labeled by the user), a platform might additionally warn "The origin of this content could not be verified. It may have been generated or altered by AI." This kind of phrasing warns but doesn't assert 100% if they aren't sure. Transparency in communication means also not overstating what we know – maintain honesty about confidence levels if relevant.

On the user side, there should be low barrier tools for them to verify content themselves if they choose to. For instance, the Content Credentials concept envisions a user can drag an image into a verification website to see its credential chain. Or for watermarks, maybe a simple app or plugin that tests an image. The more this is accessible, users or journalists can independently vet content. This is part of a healthy ecosystem of transparency similar to how anyone can verify a digital signature if they have the public key; here they should be able to verify a content signature with a public tool. For privacy reasons, it is preferable if such checks can be locally performed by the user.

5.4 Ensuring Minimal Burden and Protecting Innovation

While implementing transparency, it's critical to consider the cost and avoid over-burdening innovation, particularly for startups or open research. Some best practices and principles to consider:

Non-intrusive implementation: The solutions should not drastically degrade performance or quality of AI services. So far, content credentials add small file size and negligible compute to sign. Watermarks add minimal compute and often perceptually little loss in quality. It is important to emphasize these constraints. Testing is needed to ensure the marks are invisible and small. A solution that does not comply with these requirements might hamper adoption and incentivize actors to circumvent it. Current approaches seem to offer lightweight solutions – although the actual overhead of a method that right now only exists as a research publication is hard to judge.

Exceptions: The AI Act already provides some exceptions (e.g., no need to label if the AI is just doing minor editing, or for law enforcement use). This needs further clarification. For example, if an AI is used for red-eye removal in a photo (assistive editing), it does not have to tag the photo as AI-generated because that could confuse. Over-labeling could reduce clarity. So best practice is to label content that is substantially AI-generated or altered. Minor AI usage might not require formal marking, though arguably a content credential could still log the edit in a provenance chain for full transparency. However, this requires careful analysis and disclosure what type of changes an AI technique (e.g. enhancement, inpainting, stitching in a phone app) can cause in order to judge its impact. These important nuances should be captured.

Gradual implementation and testing: It is recommended to phase in requirements. Initially, encourage adoption and run pilots. There could be an initial period where companies implement and refine these methods, share data on how often things are labeled, issues encountered, etc. This collaborative environment can refine the practices before any hard enforcement is taking place. If for example a particular watermark turns out to be too weak or a standard field is being misused, adjustments can be made. This can also

account for the evolution of technology in general. On the generation side – as well as the regulation/enforcement side.

Balancing transparency with creative freedom: Some artists or creators might not want a mark on their work if it is for artistic expression. For the future implementation of Article 50, it should be clarified that genuine art, satire, etc. should still disclose AI usage but perhaps in a way that does not break the experience/expression (like in credits or description rather than a watermark on the video itself). Technical measures can accommodate this, e.g., a deepfake in a movie could have a credential in the file or a declaration prior to presentation rather than a visible text/symbol on screen, so that the utility is preserved, but transparency is still ensured. This indicates that disclosure tailored to context is important.

Privacy of creators/users: Ensure that transparency measures do not backfire by revealing sensitive information. For example, if an investigative journalist uses AI to protect a source's identity in a photo (blurring face by AI), labeling that photo as AI-altered is fine, but we would not want the metadata to reveal the original image or the identity. Content credentials can be configured to not expose original data unless needed. It is crucial to set defaults that respect privacy – e.g., do not automatically include user's account name in the credential, just a generic provider name, unless the user wants attribution. These nuance points should be emphasized so that adoption of transparency technology does not inadvertently cause harm or excessive burden.

5.5 International Approaches and Business Practices

We briefly compare how other jurisdictions and companies are handling AI content transparency:

China's mandatory watermarking: As noted, since January 2023 China requires "deep synthesis" content to carry a "clear identifier" like a watermark. Chinese tech companies have developed their own watermarking tools to comply. E.g., Tencent and Alibaba reportedly have internal systems to

stamp AI-generated videos. The enforcement in China is strict, with penalties for non-compliance. One lesson is that government mandate did mobilize companies to implement solutions quickly. However, it is unclear how sophisticated or interoperable those watermarks are (likely proprietary per platform). The focus is more on obvious visible labels in some cases (for instance, some Chinese apps put a label on AI-generated avatars). The EU might prefer less intrusive, more standard approaches. But the success is that now Chinese users often see an "AI-generated" label on content, raising awareness. We can take from this the importance of regulation to ensure compliance, but also note that an overly isolated approach can lead to isolated solutions. If each region uses different watermarks, detection globally is harder.

United States Voluntary Commitments: The July 2023 White House commitments by leading AI firms to develop watermarking show a best practice of proactive industry self-regulation. Following that, some large providers have publicly acknowledged work on watermarking. Large companies in the space perform research on watermarking GPT outputs and companies are participating in standards like C2PA. A best practice from the US side is the multi-stakeholder collaboration: government did not mandate directly but got commitments and also involved NIST. A notable business-led example is Cloudflare's integration of Content Credentials across its network, allowing any image hosted via Cloudflare to retain and display provenance info including AI manipulation. Such industry-driven adoption, spanning international networks, underscores the feasibility and benefit of these solutions at scale. Similar integrations and cross-industry partnerships in Europe (for instance, between cloud providers, social media, and news organizations) should be encourated to create a unified transparency infrastructure.

Chapter 6

Further Input via Workshop, Interactions, and other Consultations

Based on an initial draft, workshop, and presentation of this report, additional input was received. Also in discussions, further points were developed and noted. This part is representing some of this input that was categorized as most related, not fully covered, or otherwise noteworthy.

Role of Generative AI for training AI and its implications. As more and more AI generated data is becoming available and is being diseminated, there is an increased chance that large parts of training corpora for AI will include synthetic data. This can pose a significant risk to the quality of trained models. Some studies even predict a catastophic mode collapse in extreme cases of such recursive training [58]. This also mandates a need for marking AI generated content in order to prevent such effects as biases or collapse.

In addition, synthetic and AI-generated data is likely to play a key role in the future development of AI. Therefore, access and use of synthetic and AI-generated data for the purpose of training advanced and competitive AI models should be facilitated in order to ensure the capacity to lead innovation in AI in Europe. Only having access to marked generated AI content could severely limit Europe's capacity to innovate and lead AI research, development, and innovation.

Accounting for Human Input and Human-AI-Collaboration. The area and field of human-AI-co-created content is difficult to navigate. In many areas, content is created with a human in the loop and potentially is edited by a human after the generation by AI. Attribution and delineation is not clear. Challenging questions arise e.g. at what point the content still has to be marked.

Induced bias via marking. There is a problem scope that results from an information ecosystem where only part of the content is marked or the mark takes on a unwanted semantics. This is also sometimes referred to as the "implied truth effect". This can go both ways, as the presence or absence of the "mark" can carry an unwanted semantic of "correct", "incorrect", "truthful", or "untruthful".

Fingerprinting and Logging techniques. There is a range of logging techniques that attempt to keep track of all content that was ever created by an AI, so that it can be referenced and detected later on via matching. While there is also recent work [43], that proposes this type of approach to be effective for domains like text, there are several limitations to extend this to other domains. In the absence of large scale studies (as far as we are aware), here is the reasoning why applications to image and video does not seem feasible and/or sustainable: First, as mentioned above, the delineation between generated and not generated content can be arbitrarily small. This implies that the capacity of the logging/fingerprinting space needs to be very high – probably close to the orginal space as little to no compression is possible. In addition, this space needs to be defined beforehand when designing and deploying the logging system, which is quite challenging. The only space that comes to mind, would be a space that is designed to separate generated from non-generated data. If we would have such a space, we would also have high performing deepfake detectors that are future proof. But we do not have those right now – which does not give a path how to implement a strong fingerprinting/logging technique. However, there is some debate on this line of argumentation.

Ecosystem of Watermarking Technology Providers. It was brought up that there is an existing industry on providing watermarking technologies that has worked on the broader problem for over 25 years. However, to the

current understanding, these lack the required robustness. Larger studies that include such techniques in evaluations are missing. Therefore, their utility cannot be exactly determined.

Security against Removal of Watermarks. It is widely recognized that watermarking techniques are not robust against advanced transformations. This will remain a challenge. Several points have been brought up in this context.

If other AI algorithms, like denoising, would be the cause of removal, they might be subject themselves to watermarking requirements. However, the line is not clear here - and it is likely that also non-AI based transformations will continue to remove watermarks.

It was also mentioned that another form of protection could be a legal one, that prohibits the removal of watermarks. It was pointed out, that this could lead to complications, as neural compression and video codecs could accidentally remove watermarks – or would need to be designed to keep watermarks.

Relevance of the JPEG Trust Framework It has been highlighted that *JPEG Trust* framework (ISO/IEC JTC 21617) provides a modular, standards-based approach for establishing authenticity, provenance, and integrity in digital media, complementing the objectives of Article 50 of the EU AI Act. The group supports watermarking as a core component of content authenticity but stresses the need for interoperable, flexible, and extensible solutions rather than a single universal method. They highlight that effective implementation requires alignment between Article 50(1) and 50(2), ensuring both machine-readability and clear communication to users. JPEG Trust calls for coordinated EU-level governance, shared evaluation practices, and adoption of interoperable standards to guarantee reliability and accessibility across platforms. The feedback also notes that human interpretation of AI-generated content remains a challenge and that trust indicators must be adaptable to diverse cultural and contextual factors.

While the JPEG Trust framework aligns with many of Article 50's goals, some limitations seem to remain. The framework does not explicitly address accessibility for end users, especially persons with disabilities, as it focuses on metadata structures rather than user interface requirements. Moreover, while it supports authenticity and provenance through metadata binding

and watermarking, its current design lacks clear mechanisms for transparent, machine-readable labeling of AI-generated content that is easily perceivable to natural persons, as required by Article 50(1). Finally, governance and key management processes seem to be under development, meaning long-term reliability and accountability are not yet fully clear. Unfortunately, the framework does not seem part of recent evaluations which makes it difficult to compare. It would be important and interesting to include in further quantitative studies.

Chapter 7

Conclusions and Recommendations for Future Work

In conclusion, the toolkit for marking and detecting AI-generated visual content is progressing and, if employed in combination, can substantially support the transparency obligations of the AI Act. Cryptographic content credentials provide a verifiable record of provenance that aligns strongly with requirements for reliability and user disclosure. Invisible watermarks offer a practical means to embed machine-detectable signals in content without impacting user experience, addressing the need for an "effective and robust" marking method resilient to typical content modifications. AI-based detection algorithms serve as an indispensable safety net to identify AI content that escapes other measures, ensuring that deployers who attempt to evade marking can still be held accountable. Relevant technical standards like C2PA are in place or emerging, facilitating interoperability across a broad range of stakeholders from camera manufacturers to web platform.

While this recommendation lays out a general landscape and methodological advice, it is difficult to commit to a specific approach. Watermarking has still several research challenges until it becomes a more mature technology – despite promising results. Also many approaches are still at a research level and only few instances (like synthID) are transitioning to production. Also, as these methods do not provide an absolute level of assurance, realistic evaluation prior to deployment is highly challenging. Equally, C2PA is a promising methodology, yet several short comings have been pointed out.

Therefore, the current field is in a tension between possible technologies, realistic evaluations, and maturity for deployment.

Realizing the full potential of these solutions will require coordinated effort and further development. The following key recommendations emerge from this study:

Embrace a Multi-Layered Transparency Strategy: A layered approach should be promoted wherein AI content is simultaneously marked with cryptographic provenance metadata and, whenever feasible, with an invisible watermark. Each method compensates for the other's weaknesses, creating defense-in-depth. Concretely, providers of generative AI tools should commit to adopting the C2PA Content Credentials standard or preferably an open standard to attach signed origin metadata to all AI-generated images and videos at the time of creation. In parallel, they should implement or integrate robust invisible watermarking in their generation pipelines leveraging available technologies like Google's SynthID or similar. Open technologies are preferable. Platforms and social media should commit to preserve and display provenance metadata, and to deploy detection for known watermarks. This dual commitment could be made explicit. The outcome would be that, by default, AI-generated content carries multiple reinforcing signals – making it significantly easier for both automated systems and human observers to identify it.

Develop Supporting Infrastructure and Tools: To ease adoption for smaller players and end-users, shared infrastructure should be created. This might include a trusted certificate authority or trust framework for content credentials, so that even small content creators can get their AI content signed under a recognized umbrella, without complex setup. Additionally, open-source toolkits and services should be promoted: for instance, a free library for watermarking images (and detecting those watermarks) that any developer can plug into their AI software. On the user side, the development of browser extensions or mobile apps that allow users to easily check an image or video for authenticity info (be it reading content credentials or scanning for watermarks) should be encouraged.

Encourage Standardization and Interoperability: The EU should champion the formal standardization of these techniques in international bodies

and incorporate them into European standards. While it may not be advisable to force all providers to use the exact same watermark, standardizing the detection format or at least agreeing on a small set of robust watermark algorithms that meet an evaluation protocol would benefit interoperability. Collaborative trials could be conducted to evaluate different watermark approaches for resilience and minimal impact.

Implement Monitoring, Testing, and Iteration Mechanisms: It is vital to continuously assess how well these solutions perform in real-world conditions and to update them as needed. The Commission, perhaps via the European AI Office or another entity, should set up a monitoring program to sample content "in the wild" and see if AI-generated items are correctly being marked and caught. They could periodically publish transparency reports. Such metrics will show progress and pinpoint gaps. Furthermore, stress-testing of the measures should be done: security researchers and journalists should be encouraged to try to circumvent the transparency measures – for instance, attempt to remove watermarks or fool detection – and then report back weaknesses – basically a red teaming approach. Incorporating the findings, the industry can iterate on improving robustness (much like cybersecurity practices). There should be a commitment to supporting independent testing and to reviewing guidance or measures periodically (say annually) based evolving threats and technological advances. Treating solutions as living measures that will be refined helps ensure their longevity and adaptability in the face of rapidly advancing AI capabilities.

Strengthen International and Cross-Sector Collaboration As a final recommendation, there should be openness to global cooperation on AI transparency. This could involve working with the United States (which is pursuing voluntary commitments) to harmonize approaches – for example, mutual recognition of each other's standard markings or jointly developing watermark technology. It also means engaging sectors beyond the typical AI companies: news media, for instance, have a stake in content authenticity and should be part of developing guidelines for how they will treat AI-generated visuals in reporting.

Recommendations for Further Research and Standardisation: In the course of this study, we identified several areas where further RD would

significantly improve and support AI transparency efforts:

Robust Watermarking Algorithms: Continued research is needed to devise watermarks that are even more resistant to removal or distortion, including against AI-driven adversarial attacks. This may involve new techniques (e.g., dynamic watermarks that adapt to content, or watermarks leveraging aspects of human perception to hide signals more deeply).

Multi-Modal and Real-Time Transparency: While images and prerecorded videos are the focus of this study, AI is moving into real-time generated content (e.g., deepfake live video streams or augmented reality) or even other modalities. Watermarking or credentialing real-time streams (perhaps via embedded signals in each frame or metadata in streaming protocols) is an emerging challenge. Research into methods for live watermarking that don't add noticeable latency or real-time deepfake detection for video calls could become important, especially for future provisions or voluntary practices. Investing in cross-modal solutions will future-proof the transparency ecosystem.

Authentication of non-AI Generated Content: Complimentary to marking AI-generated content, another approach is to authenticate content that is known to be real (for instance, photos taken by journalists or images from a certified camera). This is the flip side – if more authentic content is signed at capture as with some projects by camera maker, then anything without such authentication could be scrutinized more. While not an Article 50 requirement, it complements it by reducing false positives and providing a clear "truth baseline." The EU could encourage camera and device manufacturers to incorporate signing of media. Over time, if most real content is signed, the lack of a signature itself becomes a signal that something might be AI. This synergy between provenance of real and marks on AI builds a robust media provenance framework. Standardisation of secure camera signatures and ensuring they are recognized in content credential systems is a concrete step to pursue.

Legal-Technical Interface: More research is also needed at the intersection of technical measures and legal/policy effectiveness. For example, studying how users perceive labels, what phrasing or icons best convey "AI-generated" without causing undue alarm, or conversely without being overlooked. The human factors research can guide standardization of user inter-

face elements for transparency. Additionally, examining scenarios like malicious actors deliberately adding fake "AI" labels to discredit genuine content – this requires a mix of legal deterrence (penalize such acts) and technical verification.

In summary, the work and effort towards transparency comes at a critical moment. The technical solutions exist in early stages but functional forms, and real-world examples have demonstrated their viability and effectiveness. The recommendations above aim to strike a balance between strong transparency and practicality. It has been highlighted that recommendation needs to be seen in the context of the rapid evolution of generative AI. Best practices need to accommodate adaptation and adjustments over time, as e.g. generative methods become more sophisticated or new use cases open up. While the outlined recommendations are suitable to support the transparency requirements, limitations have to be clearly communicated. A false sense of security could lead to a situation where we have generated more harm than good.

Bibliography

[1] Science & tech spotlight: Combating deepfakes. GAO Report GAO-24-107292, U.S. Government Accountability Office, Mar. 2024. URL https://www.gao.gov/products/gao-24-107292. Accessed: 2025-05-22.

[2] S. Aaronson. Simons institute talk on watermarking of large language models, 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17.

[3] A. K. Abdulrahman and S. Ozturk. A novel hybrid dct and dwt based robust watermarking algorithm for color images. *Multimedia Tools and Applications*, 78(12):17027–17049, 2019.

[4] Adobe. Content credentials overview. https://helpx.adobe.com/creative-cloud/apps/adobe-content-authenticity/content-credentials/overview.html, 2025. [Online; accessed 02-November-2025].

[5] A. Al-Haj. Combined dwt-dct digital image watermarking. *Journal of computer science*, 3(9):740–746, 2007.

[6] M. Ali, C. Ahn, M. Pant, and P. Siarry. A reliable image watermarking scheme based on redistributed image normalization and svd. *Discrete Dynamics in Nature and Society*, 2016:1–12, 2016.

[7] B. An, M. Ding, T. Rabbani, A. Agrawal, Y. Xu, C. Deng, S. Zhu, A. Mohamed, Y. Wen, T. Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Z. Bahrami and F. Akhlaghian Tab. A new robust video watermarking algorithm based on surf features and block classification. *Multimedia Tools and Applications*, 77(1):327–345, 2018.

[9] I. Bayoudh, S. Ben Jabra, and E. Zagrouba. Online multi-sprites based video watermarking robust to collusion and transcoding attacks for emerging applications. *Multimedia Tools and Applications*, 77(11): 14361–14379, 2018.

[10] S. Ben Jabra and M. Ben Farah. Deep learning-based watermarking techniques challenges: A review of current and future trends. *Circuits, Systems, and Signal Processing*, 43(8):4339–4368, 2024.

[11] A. Bhardwaj, V. S. Verma, and R. K. Jha. Robust video watermarking using significant frame selection based on coefficient difference of lifting wavelet transform. *Multimedia Tools and Applications*, 77(15):19659–19678, 2018.

[12] M. Bistron and Z. Piotrowski. Efficient video watermarking algorithm based on convolutional neural networks with entropy-based information mapper. *Entropy*, 25(2):284, 2023. doi: 10.3390/e25020284.

[13] B. Chandra, J. Dunietz, and K. Roberts. Reducing risks posed by synthetic content an overview of technical approaches to digital content transparency. 2024.

[14] B. Chandra, J. Dunietz, and K. Roberts. Reducing risks posed by synthetic content an overview of technical approaches to digital content transparency. *NIST Trustworthy and Responsible AI - 100-4*, 2024.

[15] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

[16] M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR, 2024.

[17] H. Ci, P. Yang, Y. Song, and M. Z. Shou. RingID: Rethinking Tree-Ring watermarking for enhanced multi-key identification. *arXiv preprint arXiv:2404.14055*, 2024.

[18] Cloudflare. Cloudflare launches one-click content credentials to track image authenticity and preserve creator attribution. https://www.cloudflare.com/press/press-releases/2025/cloudflare-launches-one-click-content-credentials-to-track-image-authenticity [Online; accessed 02-November-2025].

[19] Coalition for Content Provenance and Authenticity. C2pa technical specification 1.4. https://c2pa.org, 2024. Accessed: 2025-05-22.

[20] contentauthenticity.org. Rust sdk for the core c2pa (coalition for content provenance and authenticity) specification. https://github.com/contentauth/c2pa-rs/. [Online; accessed 02-November-2025].

[21] H. Ding, R. Tao, J. Sun, J. Liu, F. Zhang, X. Jiang, and J. Li. A compressed-domain robust video watermarking against recompression attack. *IEEE Access*, 9:35324–35337, 2021.

[22] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3054–3058. IEEE, 2022.

[23] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[24] P. Fernandez, H. Elsahar, Z. Yalniz, and A. Mourachko. Videoseal: Open and efficient neural video watermarking. *arXiv preprint arXiv:2412.09492*, 2024.

[25] Y. Gao, X. Kang, and Y. Chen. A robust video zero-watermarking based on deep convolutional neural network and self-organizing map in polar complex exponential transform domain. *Multimedia Tools and Applications*, 80(4):1–25, 2021. doi: 10.1007/s11042-020-09904-4.

[26] M. Ghasempour and M. Ghanbari. A low complexity system for multiple data embedding into h. 264 coded video bit-stream. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4009–4019, 2019.

[27] Google. How we're increasing transparency for gen ai content with the c2pa. https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/. [Online; accessed 02-November-2025].

[28] Google AI. SynthID: Tools for watermarking and detecting LLM-generated Text, 2024. URL https://ai.google.dev/responsible/docs/safeguards/synthid.

[29] C. Gu, X. L. Li, P. Liang, and T. Hashimoto. On the learnability of watermarks for language models. *arXiv preprint arXiv:2312.04469*, 2023.

[30] S. Gunn, X. Zhao, and D. Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024.

[31] Y. He, N. Yu, M. Keuper, and M. Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[33] J.-U. Hou. Mpeg and da-ad resilient dct–based video watermarking using adaptive frame selection. *Electronics*, 10(20):2467, 2021.

[34] Z. Jiang, J. Zhang, and N. Z. Gong. Evading watermark based detection of ai-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1168–1181, 2023.

[35] Z. Jiang, M. Guo, Y. Hu, and N. Z. Gong. Watermark-based detection and attribution of ai-generated content. *arXiv preprint arXiv:2404.04254*, 2024.

[36] N. Jovanovic, R. Staab, and M. T. Vechev. Watermark stealing in large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

[37] M. Kaczyński and Z. Piotrowski. High-quality video watermarking based on deep neural networks and adjustable subsquares properties algorithm. *Sensors*, 22(14):5376, 2022.

[38] Z. Ke, H. Huang, Y. Liang, Y. Ding, X. Cheng, and Q. Wu. Robust video watermarking based on deep neural network and curriculum learning. In *2022 IEEE International Conference on e-Business Engineering (ICEBE)*, pages 80–85, 2022. doi: 10.1109/ICEBE55470.2022.00023.

[39] S. A. Khan and D.-T. Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24, page 1006–1015, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706196. doi: 10.1145/3652583.3658035.

[40] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[41] N. Krawetz. Secure evidence attribution label (seal). `https://github.com/hackerfactor/SEAL`. [Online; accessed 02-November-2025].

[42] N. Krawetz. Secure evidence attribution label (seal): Comparison. `https://github.com/hackerfactor/SEAL/blob/master/COMPARISON.md`, 2024. Accessed: 2025-05-22.

[43] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[44] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[45] X. Luo, Y. Li, H. Chang, C. Liu, P. Milanfar, and F. Yang. Dvmark: a deep multiscale framework for video watermarking. *IEEE Transactions on Image Processing*, 2023.

[46] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017.

[47] S. Mansour, S. B. Jabra, and E. Zagrouba. A robust deep learning-based video watermarking using mosaic generation. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 668–675. INSTICC, SciTePress, 2023.

[48] K. Navas, M. C. Ajay, M. Lekshmi, T. S. Archana, and M. Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd international conference on communication systems software and middleware and workshops (COMSWARE'08)*, pages 271–274. IEEE, 2008.

[49] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

[50] L. Pan, A. Liu, Z. He, Z. Gao, X. Zhao, Y. Lu, B. Zhou, S. Liu, X. Hu, L. Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024.

[51] Q. Pang, S. Hu, W. Zheng, and V. Smith. No free lunch in llm watermarking: Trade-offs in watermarking design choices. *arXiv preprint arXiv:2402.16187*, 2024.

[52] J. Piet, C. Sitawarin, V. Fang, N. Mu, and D. Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[55] M. Saberi, V. S. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

[56] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *Advances in neural information processing systems*, 29, 2016.

[57] R. San Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon, and T. Tran. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, volume 235, 2024.

[58] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[59] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[60] M. Tancik, B. Mildenhall, and R. Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020.

[61] TruePic. Truepic breakthrough charts a path for restoring trust in photos and videos at internet scale. https://www.truepic.com/blog/truepic-breakthrough-charts-a-path-for-restoring-trust-in-photos-and-videos-a [Online; accessed 02-November-2025].

[62] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

[63] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, 2024.

[64] P. Yang, H. Ci, Y. Song, and M. Z. Shou. Steganalysis on digital watermarking: Is your defense truly impervious? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[65] X. Yang, L. Pan, X. Zhao, H. Chen, L. Petzold, W. Y. Wang, and W. Cheng. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*, 2023.

[66] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024.

[67] N. Yu, L. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *International Conference on Computer Vision (ICCV)*, 2019.

[68] N. Yu, V. Skripniuk, D. Chen, L. S. Davi, and M. Fritz. Responsible disclosure of generative models using scalable fingerprinting. In *International Conference on Representation Learning (ICLR)*, 2022.

[69] H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

[70] K. A. Zhang, L. Xu, A. Cuesta-Infante, and K. Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019.

[71] K. A. Zhang, L. Xu, A. Cuesta-Infante, and K. Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019.

[72] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun, and H. Guan. Robust image watermarking using stable diffusion. *arXiv preprint arXiv:2401.04247*, 2024.

[73] X. Zhao, P. V. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for ai-generated text. In *The Twelfth International Conference on Learning Representations*, 2024.

[74] X. Zhao, S. Gunn, M. Christ, J. Fairoze, A. Fabrega, N. Carlini, S. Garg, S. Hong, M. Nasr, F. Tramer, S. Jha, L. Li, Y.-X. Wang, and D. Song. Sok: Watermarking for ai-generated content, 2024.

[75] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 2024.

[76] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung, and M. Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[77] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision*, 2018.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696,
– via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Publications Office
of the European Union