Solutions to the Content Authenticity Problem:

A Technical Overview of Creating, Managing, and Analyzing Digital Provenance

August 2024

trufo.ai

Abstract: The proliferation and increasing sophistication of AI-generated media necessitates reliable and scalable provenance. Numerous efforts are underway across a full range of developer and regulatory stakeholders, with the common goal of establishing trust in digital media without impeding the adoption of generative AI technologies. This white paper provides a high-level survey of potential solutions, with the primary conclusion that a system of labeling and detecting provenance should be adopted, where the system is secure and the provenance is durable.

Contents

1 Introduction

Generative AI represents a foundational transformation of content creation: the material previously created with human input can now be freely generated using AI. Furthermore, rapid improvements are quickly moving generated outputs beyond the ability of human detection. As societal and economic results unfold, one clear takeaway is that the labeling and detecting of provenance in digital content will become increasingly valuable. This demand for provenance will be driven by:

- Content consumers and content platforms who want to mitigate misinformation.[1]
- Individuals and businesses who want to protect their identities from deepfake fraud.[2]
- AI dataset curators who want to differentiate generated and non-generated content.[3]
- Artists and photographers who want to protect the creative copyrights of their works.[4]
- Journalists and attorneys who want to prove that their images and videos are real.[5]
- Celebrities and politicians who want to defend against personal deepfake attacks.[6]

These diverse use cases share two things in common: (1) the progression of generative AI will drive continued demand for solutions to these use cases, and (2) a necessary component of these solutions is robust provenance information for digital content.

Section §2 will discuss the urgency of the problem in relation to progression in generative AI and impending legal requirements. Section §3 will explore the benefits of using content labels to assist in content provenance analysis. Sections §4 and §5 will respectively cover technical methods of creating annotations and associations for digital content provenance. Section §6 will analyze the requirements of a content authenticity infrastructure and present how Trufo's platform architecture solves for these demands. The scope of this paper does not include concrete applications of solutions in the aforementioned use cases nor deeper dives into key technical components; future publications will address these.

2 Urgency

Interest in both the problem and solution have surged as AI-generated content explodes in popularity. Properly handling this drastic increase in AI-generated content has become a priority for governments and authorities, with the U.S. government introducing and soon passing a number of bills related to synthetic content.

For businesses, there is significant risk that customers may demand provenance or legislation may be enacted before the requisite solutions are implemented. For policymakers, understanding solutions to digital content provenance is critical to avoid ending up with a complex web of ineffective laws and standards.

2.1 Quality & Quantity

Both the quality and quantity of AI-generated content are quickly increasing.

The capabilities of AI-generated content have leapfrogged over the past two years following the release of ChatGPT. Capabilities have quickly expanded from generated text to audiovisual content and beyond.[7] Outputs are becoming more and more realistic:[8] quick tells like the number of fingers on a hand are no longer effective ways to identify AI. Even if fully-generated content can be identified, partially-altered content is far more clandestine. Advanced AI editing tools such as Google's Magic Eraser are being rapidly developed and represent a new level of difficulty for identification.[9]

High-quality synthetic content is a double-edged sword. For example, studios in Hollywood are using generative lip-syncing technology to improve the dubbed content experience.[10] On the flip side, a UPenn student who posted a video online later found that her visage was used in thousands of propaganda videos.[11] As quoted in a Europol report, "We are entering an era in which our enemies can make it look like anyone is saying anything, at any point in time. Even if they would never say those things."[12]

AI-generated content is also proliferating online. The percentage of top 20 Google search results that are AI-generated has gone from 4% in 2021 to 7% in 2023 to 10% in March to 13% in June.[13] An often-cited estimate predicts that over 90% of online content will be AI-generated by 2026.[14] The digital landscape is reaching an inflection point.

Already, social media is being inundated with AI-generated content, evidenced by AI-created images on Facebook garnering millions of engagements.[15] The massive volume means two things: (1) in order to handle the flood of images and videos generated daily, the solution must be scalable; (2) as generated content becomes the norm, we must establish robust methods to prove what is real, not just guess what is fake.

2.2 Legislation & Regulation

Legislators and regulators have been moving quickly to address generative AI. Proposals have shared the common theme of establishing reliable provenance in digital media:

- In June 2023, the EU passed the Artificial Intelligence Act, the first effort to comprehensively regulate AI.[16] The Act requires providers of AI systems that generate or manipulate audio, image, video, or text to mark outputs in a machine-readable format as artificially generated or manipulated using, for example, watermarks.[17] Many details about specific implementation, including technical standards, still need to be developed through secondary legislation and guidelines in the coming months and years.[18]
- In November 2022, China introduced regulations on the labeling of AI-generated content.[19] These guidelines were finalized in August 2023 and includes both invisible and visible watermarks; notably, spatial and transform domains were recommended.[20]

- In October 2023, Biden issued an Executive Order on promoting safe, secure, and trustworthy AI.[21] Subsequently, the Commerce Department established a public-private consortium that includes universities like Princeton, companies like Google, and startups like Trufo.[22]
- The U.S. Congress has seen numerous bills introduced in the current and previous sessions, including the: Advisory for AI-Generated Content Act (S.2765), Protecting Consumers from Deceptive AI Act (H.R.7766), REAL Political Advertisements Act (H.R.3044), Protect Elections from Deceptive AI Act (H.R.8384), AI Labeling Act of 2023 (S.2691), and the Artificial Intelligence Research, Innovation, and Accountability Act of 2023 (S.3312).
- In 2021, the DEEPFAKES Accountability Act was introduced as an early draft.[23] More recently, the DEFIANCE Act passed the U.S. Senate last month and is pending a House vote.[24] These target malicious uses of digital forgeries.
- In December 2023, the New York Times sued OpenAI and Microsoft for copyright infringement in a high-profile case.[25] Last month, the COPIED Act was introduced in the Senate, which prohibits the removal or tampering of provenance information in digital content.[26]

Ongoing legislative efforts that use provenance to manage generative AI suggest that robust content provenance systems will soon be required. Together, these solutions would constitute a digital trust infrastructure.

2.3 Ongoing Efforts

Multiple collaborative efforts that are ongoing. Two of the more prominent ones are:

AISIC.WG2 In response to Biden's executive order on AI Safety, the NIST has created an AI Safety Consortium (AISIC). The second working group, which Trufo is an active contributing member of, focuses on the labeling and detection of synthetic content.[27]

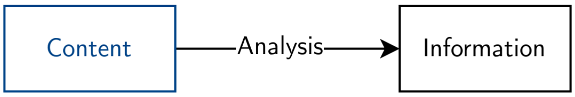
C2PA C2PA is the technical component of the CAI and provides the prevailing standard of signed metadata for digital content. It is led by Adobe and joined by companies such as Microsoft and OpenAI.[28] Trufo is also a member and participates in multiple working groups.

Other efforts (not exhaustive) include: Media-affiliated groups, such as the IPTC, the SMPTE, and the SAG-AFTRA, who are exploring potential industry standards and policies. Academic conferences, such as the NeurIPS, the ICML, and the ICCV, now increasingly accept papers and posters on labeling and detection.

The space is quickly evolving: technologies are being developed, standards are being adopted, and systems are being implemented. There are an incredible number of stakeholders, each with different problems to solve, each with the goal of selecting the right solution at the right time. The next few sections will summarize the various approaches to solving this urgent problem.

3 Assisted Analysis

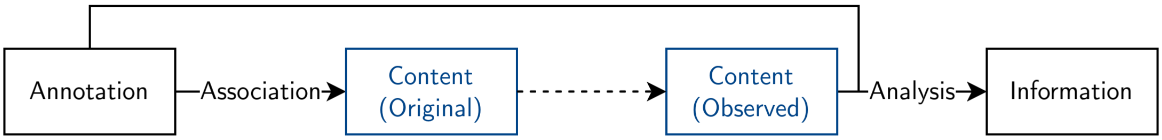
When content is consumed, the content can be analyzed to obtain information about the content.



This information can vary from one-bit classifications, such as whether the content is AI-generated, to more substantial profiles, such as the attribution and copyright details of the content.

Effective analysis needs to be both reliable and scalable. If the accuracy of the information is low, then the analysis cannot be safely trusted. If the information cannot be produced at scale, then the analysis cannot be widely adopted.

One important concept in analysis is assistance. Oftentimes, certain annotations (which contain provenance information) have been associated to the content. For example: adding a timestamp to the file metadata, embedding an 'AI-generated' watermark, and storing a forensic copy in blockchain.

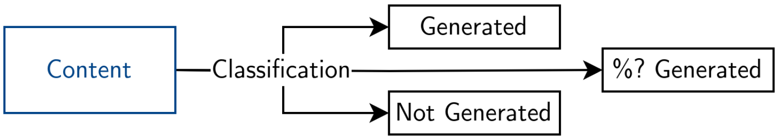


If the associated annotation can be retrieved, then the additional data can be used to assist in the analysis process.

3.1 The Need for Assistance

The main drawback of an assisted analysis model is that it is dependent on upstream annotation and association processes. This imposes an upfront cost of investment. In contrast, unassisted analysis models are easier to implement and therefore more readily adoptable.

AI detection models are the main example of unassisted analysis models. The idea is simple: use machine learning to train a model that can accurately detect whether content is AI-generated or not. Some examples include GPTZero for text and Illuminarty for images.



This is unfortunately much easier said than done. The main problem, as many users have quickly noticed, is that these detection models are highly inaccurate.[29] In a open deepfake detection challenge hosted by Meta with over 2000 participants and $1 million in prizes, the winning accuracy was 65%.[30] OpenAI had released a classifier for AI-generated text, only to recall it a few months later due to low accuracy.[31]

To further complicate things, generative AI are constantly evolving: detection models that work now may not work in the future. A generative model can even be trained to bypass specific detection models via adversarial training, resulting in an endless cat-and-mouse game. This dynamic perpetually limits the reliability of detection models, rendering them insufficient as a standalone solution.

Assisted models are capable of much higher accuracy. As an example, a verified digital signature can prove both the identity of the signer and the integrity of the content. Is this sort of assistance really necessary though? After all, provenance is not new—doctored photos and staged videos are an age-old issue.

In fact, fakes, whether doctored or staged or generated, are frequently identified by people, either noticing an error in the content or discovering circumstances that render the content impossible. For example, the viral image of the Pentagon that temporarily wiped out a hundred billion dollars off the U.S. stock market was confirmed not with sophisticated AI detection tools but through simple human detective work.[32]

In this human forensics approach, outside of obvious tells like six fingers attached to a foot, most evidence boils down to finding and analyzing provenance: a copy of the original content or a statement by the original publisher is what typically serves as proof. This entails retrieving an associated annotation. And while this method has been reliable so far, it is not scalable: at the rate at which we are consuming content, taking a few hours to track down whether or not something is believable ranges from impractical to impossible.

The conclusion is that assisted analysis is the only viable approach.

3.2 Investing in Content Labels

In order for assisted analysis to work, an extra step is required: content labeling. This requires the creation of an annotation and its subsequent association to the content. The methods of doing so will be discussed in detail in §4 and §5.

The implementation of this content labeling step requires a coordinated investment. Stakeholders, from individuals to corporations, would need to begin labeling their content, and do so in a coordinated fashion: if each party implements their own in-house label, then the fragmented nature of the annotations and associations would degrade the overall efficacy of assisted analysis. This sort of coordinated investment can be expensive but worthwhile.

For most content producers, the appeal of labeling content is not to improve assisted analysis, as there is little economic incentive. However, there is a tangible benefit: provenance management. By investing in content labels, creators can assert control over their content. For digital artists, this is an assertion of IP ownership through flexible copyright licenses. For AI companies, this is a way to comply with regulations and avoid recursive training. For journalists and individuals, this is a method to prove whether something is real or fake.

Perhaps most importantly, labeled content carries value. With the proper labels added, content can be more easily monetized, verified, traced, and more.

4 Annotations

Annotations come in all sorts of shapes and sizes, and information of any sort can be included. This section discusses several important types of information.

Importantly, ensuring that the information is trustworthy is more difficult than including information. This trust is dependent both on the entity that enters the information as well as how secure the information is after being entered. For example, traditional metadata can be easily modified. As annotations are being used to carry increasingly impactful information, annotations must be able to assert trust.

4.1 Attribution and Copyright

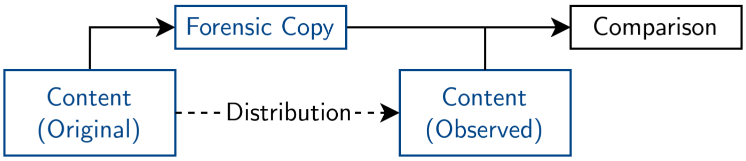
In addition to the standard metadata fields (timestamp, geo-location, device model), information identifying the publishing entity and the copyright license can be included. For example, a photographer may opt to include their name and a CC-BY-SA license, such that anyone who comes across their photos will be able to (1) easily give credit and (2) use the photo as intended.

A tag that explicitly specifies "do-not-train" or "AI-generated" can be included. There will soon be legal protection against the deliberate removal of these tags.[33] The latter will soon be legally required for all generated content.[34]

More substantial licenses have also been crafted for content in an age of AI, with Story and Trufo giving users more fine-grained control over derivative usage.

4.2 Forensic Copies

By storing an original, forensic copy of the content, later versions of the content can be compared against the original. There are two main benefits: (1) the full original can be accessed, (2) a thorough comparison can be made. This is particularly useful for digital asset management where the original can be quite large, in both raw pixel data and in contextual metadata.



The forensic copies can also be stored in a decentralized public ledger.

> Technology Focus—Blockchain: Originally intended for decentralized finance, blockchains can find potential applications in digital provenance. Centralized providers may not be trustworthy, and even if proper security precautions are taken, a simple mistake could result in critical data loss.[35] For those who value decentralized security over centralized efficiency, or for those who simply like cryptocurrencies and non-fungible tokens, public ledgers are attractive.

A number of startups, such as Numbers, provide blockchain-based solutions. When the forensic copy is coupled with an attribution identity (§4.1) and a secure timestamp (§4.5), an NFT is created.

4.3 Digital Signatures

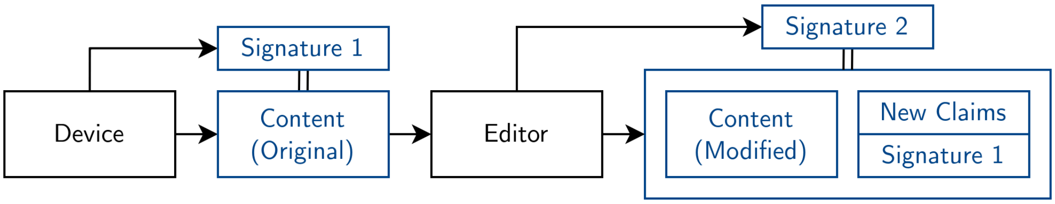
Digital signatures are incredibly useful because they can prove that a digital message (1) was sent from a particular entity and (2) has not been tampered with. They form the basis of HTTPS. In the context of content authenticity, they can prove that the content (or the associated annotation) has not been modified since it was signed by the content publisher. If the signer is a reputable entity, such as BBC, the signature establishes verifiable trust.[36]

Technology Focus—Public Key Cryptography: While traditional symmetric encryption uses a single secret key to encrypt and decrypt messages, asymmetric encryption uses a private and public key pair (s, p) to encrypt a message with the private key $m \to c := \mathsf{Enc}(m, s)$ so that the original message can be decrypted with the public key $\mathsf{Dec}(c, p) \to m$. This, along with cryptographic hashes H, makes digital signatures possible: the sender creates a digital signature $\sigma := \mathsf{Enc}(\mathsf{H}(m), s)$ and sends (m, σ, p) to the receiver; the receiver then validates the public identity of p and verifies that $\mathsf{H}(m) = \mathsf{Dec}(\sigma, p)$.

The main vulnerabilities to consider are (1) attribution of the public key to the wrong identity and (2) leaking of the private key. Building a PKI addresses the former vulnerability.

Technology Focus—Public Key Infrastructure: A public key infrastructure is an ecosystem that manages public-key encryption, in particular so that entities can register and thus bind their public keys to their identity. This requires a trusted root authority who can review the validity of these registrations (or delegate this power to intermediate authorities). The cryptographic tool used to prove valid registration is called a certificate, and the current standard is X.509.

C2PA is working to incorporate digital signatures into metadata annotations called 'manifests.' Derivative usage is supported: entities can not only sign their own claims (creating a manifest), but can also sign pre-existing manifests. For example, an editor may take a photo signed directly by the camera, make some modifications, write claims about these modifications, and then create an updated manifest that captures the entire provenance chain.



4.4 Fuzzy Authentication

While standard digital signatures can prove that a message m is completely unchanged, they fail once the message m has been altered in any form, such as when content is compressed or its file format changed. The goal with fuzzy authentication is to find a way to validate content that is (moderately) compressed or otherwise modified.

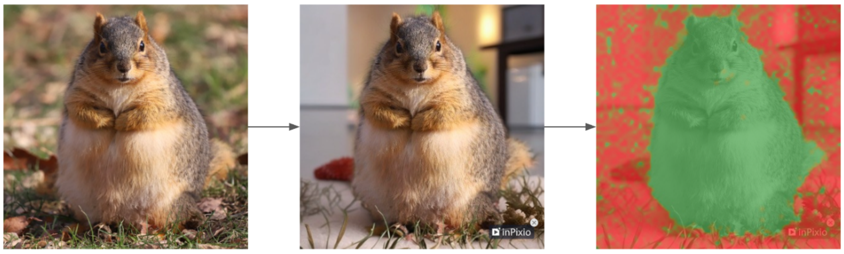
Doing so will require a new type of hash function H, along with a corresponding cryptographic scheme. There are two primary candidates: zero-knowledge proofs and differential analysis.

Technology Focus—Zero-Knowledge Proofs: Zero-knowledge proofs allow one party to prove certain statements about an object to another party without revealing any information beyond the statement itself. For example, one can prove knowing how to factor large integers without revealing the factors themselves. More advanced proofs, such as ones used in VerITAS, can prove that images have been resized, blurred, or cropped.[37] They do so by defining a specific set of circuits are defined and generating a special type of hash.[38]

Some commercial providers, such as MayaLabs, use zero-knowledge proofs. The main weaknesses are that (1) the verification process is computationally heavy, usually taking minutes on a commercial GPU, and (2) the set of edits that can be proved is limited by the specific circuits that are defined.

Technology Focus—Differential Analysis: Differential analysis relies on a new object called a differential hash: while perceptive hashes (e.g. reverse image search) are used to find matches, differential hashes are designed to identify differences between matches—more similar to the secure sketches used in biometric identification. Through computing the differential hash of an observed content item and comparing it to the original hash, modifications to the content can be detected and localized. No forensic copy is required, so privacy is preserved.

Currently, Trufo employs differential analysis in their watermark-based system. This method only takes milliseconds to run and works on a broad range of modifications.



From left to right: *Trufo* watermarked, *inPixio* modified, *Trufo* raw differential analysis.[39]

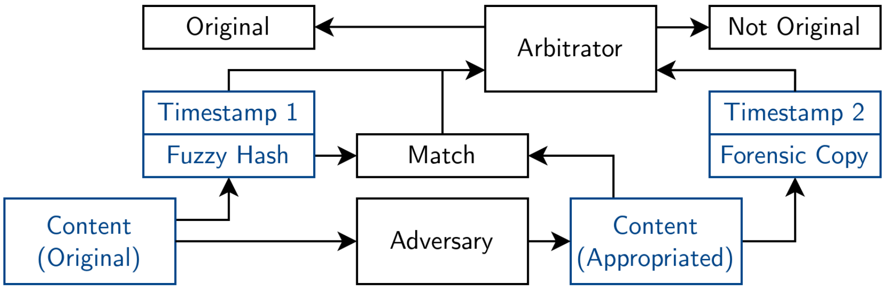
The main weaknesses are that (1) proving theoretical security of the differential hash is difficult, so adversarial exploits could exist, and (2) in order for the hash comparison to be effective, the content needs to be properly aligned.

Both of these are relatively new technologies, which means that they are somewhat untested but do have significant potential in improvement.

4.5 Timestamps & Ownership

Oftentimes, forensic copies include a secure timestamp to prove original ownership: if someone makes a forensic copy of content they did not produce, then the original producer of the content can use their older forensic copy to show that they are the owner. This does require that the original producer make a forensic copy, and this does require that the secure timestamp is recorded faithfully. As such, the secure timestamp should only be issued by a trusted provider.

A forensic copy is not needed for a secure timestamp to be effective. In particular, even if an adversary modifies the content before attempting to claim ownership, as long as the producer can find a approximate match via fuzzy authentication, ownership can be proven. This does require either an unsecured original copy or a lightly modified copy (e.g. the adversary's version).



The inclusion of secure timestamps, along with the other types of annotations discussions in this section, will help automate both assisted analysis and provenance management.

5 Associations

In order for the annotation to be useful, it must be associated to the content, so that the annotation can be retrieved and analyzed when the content is consumed.

5.1 Metadata

When digital audiovisual content is stored in a file, the file typically contains the encoded audiovisual data along with some attached metadata. The metadata is formatted according to existing standards: for images, these standards are XMP, EXIF, IPTC, and soon C2PA. Although there are standard fields set aside for specific information, any type of annotation can be included as long as whoever receives the file knows how to read the data.

The main strength of metadata is that it is local: everything can be done offline, without needing to pass through a central provider or even any degree of Internet access. Even digital signatures, which require validation of the public key identity, can be verified: all that is needed is a list of root certificate authorities (comes with the OS) and a valid X.509 certificate (in the metadata).

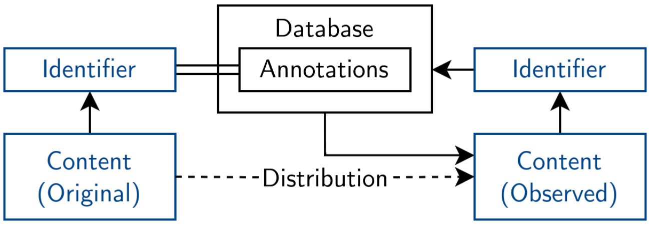
The main weakness of metadata is that it is both easy to remove and easy to modify. With traditional metadata, the fields can be changed without consequence. By incorporating digital signatures, as C2PA does, a degree of protection can be added; however, in the absence of a trusted provider, exploits are still possible.[40]

Removal is difficult to address. For example, when an iPhone photo is exported, the metadata is lost. For example, when an image is uploaded to Facebook, the metadata is lost. Making sure that every single operation preserves metadata is near impossible, and even if the metadata is preserved, the content is modified ever so slightly (e.g. iPhone changes the content format from HEIC to JPEG, Facebook aggressively compresses large images) so any digital signatures would break.

5.2 Secure Content Records

Rather than solely appending annotations to the content via metadata, embedding annotations could be useful in addressing the weaknesses of metadata. Embedding a lot of data in a robust manner is difficult, so the solution is to embed an ID and store the rest of the data elsewhere.

The records typically live in a secure database, so that they cannot be modified or removed without proper authorization. The database could be a public ledger, but a standard database with standard security precautions is more effective.



5.3 Perceptive Hashes

One method of extracting a content identifier is through a perceptive hash. These hashes effectively project the content into a much smaller space, taking 4K images or 1-hour videos down to a single vector a few hundred bits long. This means that the hashes can be efficiently stored and effectively searched. For example, Meta's PDQ model, an upgraded version of pHash, is used to identify harmful content by looking for matches against the hashes of known harmful content.[41]

Technology Focus—Latent Representations: Advances in machine learning, and in particular image classification, have opened up a new class of perceptual hashes, based on latent space embedding models rather than on low-resolution DCT coefficients. For example, in OpenAI's CLIP model, the embedded vector has dimension 512 and the similarity check is a simple cosine dot-product.[42]

The main weakness of perceptive hashes is that they are not secure. There is no protection against incorrectly-extracted identifiers, either due to inaccuracies in the hash similarity function or due to adversarial exploitations of the hash. Thus, they are best suited for applications like reverse image search and copyright detection where additional (often manual) review is conducted afterwards.

While perceptual hashes are acknowledged to be insecure, the fuzzy extractors used in biometric identification are viewed as secure.

Technology Focus—Fuzzy Extractors: In biometric identification, fuzzy extractors are successfully used. They work by storing a secure sketch p for each original identity w; if the input w' is very similar to the original $||w' - w|| < \delta$, then the secure sketch allows the key $\mathsf{H}(w)$ to be obtained.[43]

Why not use fuzzy extractors then? The key is that in biometric identification, the identity is already known. But this can still work in the context of digital content: a watermark can be used to determine the identity (by extracting the identifier) and a differential hash can be used for the secure sketch verification.

5.4 Watermarks

The content identifier (or one-bit "do-not-train" and "AI-generated" labels) can be embedded directly into the raw audiovisual data as a watermark. Unlike the watermarks commonly found online, these watermarks can be made invisible, so that the quality of the content is not affected, and durable, so that the watermark can be detected even when the content is modified.

Watermark technology is rapidly improving. For example, Trufo's watermarks perform an order of magnitude better than popular open-source libraries (used by Stable Diffusion[44]) in both invisibility and durability.[45] The topic of watermarks is an extensive one, and will be addressed more thoroughly in a future white paper. For now, here are some types of (image) watermarks.

Technology Focus—Image Watermarking Methods: Four broad watermarking methods are:

Transform The signal is stored in a transform space, commonly DCT, DFT, or DWT, of the image. For example, Eluv.io uses a DTCWT transform[46] in video.

Dense The signal is stored redundantly at high frequencies, for high cropping and warping resistance. For example, Digimarc uses this in physical product labels.

Feature An AI model, typically with a CNN and a GAN, is trained on a curated dataset of images and modifications. For example, Adobe's Trustmark uses this.

Diffusion A transform or feature pattern is added during the AI content generation process. For example, Meta's Stable Signature and Google's SynthID use this.

6 Building a Robust System

As discussed in §3, an effective solution must be both reliable and scalable; a system that cannot be trusted or adopted at large will not work except as a stopgap. To build such a solution, the proactive labeling and subsequent analysis of digital content provenance is needed.

Ensuring that such a provenance system is indeed reliable and scalable is not a given. Reliability cannot be achieved without secure provenance information that can then be consistently retrieved and validated. This requires a coherent and cohesive architecture that leverages cryptography and infrastructure. Scalability is limited by the durability of the association and the easiness of adoption. If the provenance is lost, then it cannot be retrieved let alone validated. If the the solution is too cumbersome to adopt, then it will not be used.

Thus, for a solution to be robust, three properties are needed: security, durability, and functionality. This section will explore these properties, with a focus on how Trufo's platform satisfies all three.

6.1 Levels of Durability

Recall that a provenance-based solution consists of two steps:

1. The labeling process creates an annotation and associates it to the content.
2. The detection process recovers the associated annotation and uses the information within.

These solutions fall under three categories based on the durability of the association:

Fingerprint Detection. When an accredited producer (e.g. Netflix, Disney) creates content, the content platform (e.g. YouTube, TikTok) stores a fingerprint of the content in an internal database. When new content is submitted to the platform, it is reviewed against the database of fingerprints and any matches are flagged.

- Annotation: the identity of the producer and the original content item.
- Association: the fingerprint, in particular a perceptual hash, located in a database.
- Analysis: algorithmic substring-style search over stored fingerprints.
- Durability: as long as the content is similar.

Variants: systems can specifically fingerprint known harmful content so that future potentially harmful content can be monitored and identified (e.g. suspected CSAM).

Cryptographic Metadata. When a content producer (e.g. camera, editor) creates or modifies a content item, they add a claim and a corresponding digital signature into the metadata. When the content is later consumed, for viewing (e.g. on a content platform) or for further editing, the chain of claims can be retrieved and verified from the metadata.

- Annotation: the identity of the producer along with added claims, plus digital signatures.
- Association: the metadata, located in the content file.
- Analysis: processing of metadata and the digital signatures/certificates within.
- Durability: as long as the metadata is fully intact.

Variants: a similar provenance chain can be built in Web3 by placing the content and associated claims into a blockchain infrastructure.

Watermark Provenance: When a content producer (e.g. media company, camera device) publishes content, they simultaneously create an annotation and watermark the content. The annotation is stored in a database. When the content is later consumed, the watermark is decoded, the corresponding annotation is retrieved, and the provenance information is processed.

- Annotation: the identity of the producer along with added claims, plus digital signatures.
- Association: the watermark, embedded into the content itself.

- Analysis: detection of watermark, extraction of identifier, retrieval of annotation, and subsequent authentication.
- Durability: as long as the watermark is intact.

Variants: by setting the annotation to be a single "AI-generated" bit, regulatory requirements can be met with minimal infrastructure investment.

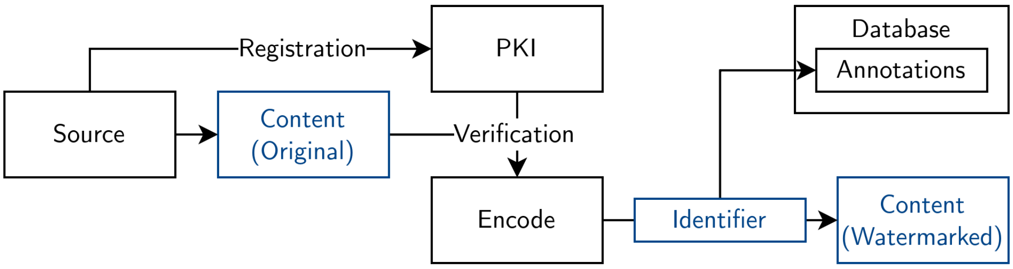
There are trade-offs at each level of durability. Fingerprint detection suffers from a high false positive rate. Cryptographic metadata fails if the metadata is removed or modified. Watermark provenance slightly reduces the quality of the content. The solution of choice may depend on the use case.

Where Trufo excels is that (1) it is the first secure watermark provenance provider and (2) Trufo's platform supports products across the full range of durability levels. The following two sub-sections will discuss these aspects of Trufo's platform in more detail.

6.2 System Security

In order to ensure that the digital content provenance is trustable, the system as a whole must be secure, from the creation to the analysis of the content labels.

In the content labeling process, the identity of the content producer must be registered with the trusted authority that creates the secure content record. The different content producer types have different security needs; Trufo supports tailored registration processes for: user accounts, physical cameras, verified organizations, and so on. Prior to labeling the content, Trufo will validate both the identity of the producer as well as the integrity of the content with digital signatures.



This ensures that the content labeling process is secure.

In order for the assisted analysis to be secure, the extracted provenance information needs to be accurate. The annotation can be secured with a digital signature and secure storage. However, the association is vulnerable to attacks.

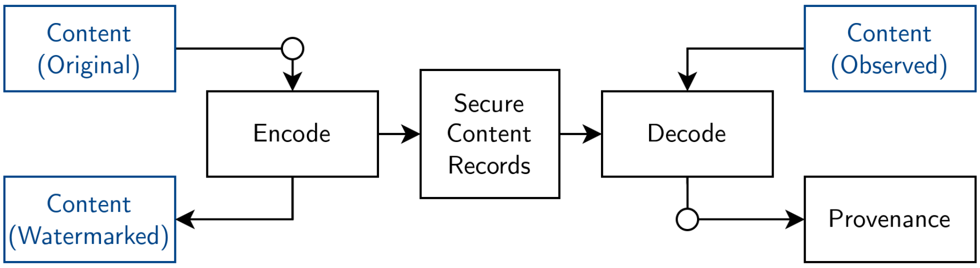
Traditional watermark-based systems do not provide adequate security in this regard. Most modern watermarks, for example those used by Google and Meta to comply with Biden's executive order, are vulnerable to both removal and forgery, limiting their practical effectiveness: if the watermark is removed, then AI-generated content is no longer labeled; if the watermark is forged, then real content can be claimed as fake.[47]

The architecture of Trufo's system allows these vulnerabilities to be patched. If the watermark is removed, no identifier can be extracted and thus no annotations can be retrieved from the secure content record. If the watermark is forged, then when the incorrect annotation is retrieved, both the digital signature(s) and the fuzzy authentication will fail. In both cases, the assisted analysis process will be able to avoid displaying incorrect provenance and potentially even infer that removal or forgery may have occurred.

This does not mean that the watermark itself is easy to remove or easy to forge—resistance against these attacks is still important, and Trufo is collaborating with the watermark removal challenge at NeurIPS 2024 to research watermark robustness. But, this does mean that even if Trufo's watermark is successfully attacked, Trufo's system can still be used with confidence.

6.3 Functionality

Trufo's patent-pending trust system consists of both secure labeling (encoding) and secure detecting (decoding), as illustrated below. For the user—the content producer or the content consumer—there is only one simple interaction to worry about: encoding (for producers) or decoding (for consumers). Decoding is available for free at https://trufo.ai/verify.



Trufo's model for secure watermark provenance: authentication checks occur at the circular nodes.

For the user—the content producer or the content consumer—there is only one simple interaction to worry about: encoding (for producers) or decoding (for consumers). Decoding is available for free at https://trufo.ai/verify.

Encoding entails more steps. Individual users can create an account at https://trufo.ai/ or download the mobile app. Enterprise partners can request custom solutions. For the actual encoding, the various annotations and associations discussed in §4 and §5, respectively, are available and can be configured. Default settings are provided; they generally include: attribution and copyright information, content hashes for authentication, a C2PA metadata manifest, and an all-purpose watermark.

Since the C2PA metadata manifest can be included in the annotation, the full provenance chain from the cryptographic metadata can be preserved. Since a perceptual hash can be included in the annotation, the searching and matching operations of fingerprint detection can be supported. Depending on the specific Trufo product, these functionalities may or may not be available.

7 Recommendations

The problem of content authenticity is a broad one, and the ultimate solution will be a joint effort between many players, from individuals to governments, from corporations to institutions, from public figures to the average person. From this white paper, a couple of conclusions can be drawn.

1. The problem is urgent, and there will likely be a significant acceleration in the development, implementation, and adoption of solutions within the next few years.
2. Provenance-based assisted analysis is far more effective than unassisted analysis in the long run; however, an investment in content labeling must first be made.
3. Such a labeling system requires both robust annotations and robust associations to work well. This is best achieved through secure and scalable infrastructure.

There is already significant investment in such solutions. For example, a blockchain-based solution that enables C2PA-like provenance trees but in Web3 is already valued at $2 billion before launch.[48] Given that blockchain is only a small part of the overall infrastructure, the total economic and social value that content provenance solutions bring may be tremendous.

In particular, content with provenance will become more valuable than content without provenance. One recommendation, therefore, is that most content producers should invest in labeling content. This will provide both immediate value—in asserting IP, in proving claims—and serve as insurance against malicious actors, AI advancements, and regulatory changes.

For a deeper dive into existing joint efforts, C2PA and the broader CAI group are excellent starting points. Note that C2PA is a standard, not a product; for actual usage, the recommendation is to try out a C2PA provider such as Trufo.

Lastly, for those looking for a provenance solution, check out Trufo! Trufo's platform combines cryptography, watermarking, and other suitable components into what is arguably the most robust solution to the content authenticity problem.[49]

8 References

Much thanks to the team at Trufo for providing consultation throughout the compilation of this white paper. For further inquiries, feel free to email tech@trufo.ai or press@trufo.ai.

[1] Tate Ryan-Mosley. *MIT Technology Review*. October 4, 2023. https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/.

[2] *Department of Homeland Security*. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

[3] Ilia Shumailov et al. *Nature*. July 24, 2024. https://www.nature.com/articles/s41586-024-07566-y.

[4] Michael Grynbaum, Ryan Mac. *The New York Times*. December 27, 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[5] *The Economist*. January 9, 2023. https://www.economist.com/science-and-technology/2023/01/09/proving-a-photo-is-fake-is-one-thing-proving-it-isnt-is-another.

[6] Nadeem Badshah. *The Guardian*. March 21, 2024. Celebritiesandpoliticianswillwanttodefendagainstdeepfakeattacks.

[7] *OpenAI*. May 13, 2024. https://openai.com/index/hello-gpt-4o/.

[8] Stuart Thompson. *New York Times*. June 24, 2024. https://www.nytimes.com/interactive/2024/06/24/technology/ai-deepfake-facebook-midjourney-quiz.html.

[9] Dina Berrada. *Google*. April 10, 2024. https://blog.google/products/photos/google-photos-editing-features-availability/.

[10] Audrey Schomer. *Variety*. May 10, 2024. https://variety.com/vip/lip-sync-dubbing-beta-tests-begin-hollywood-1235994811/.

[11] Matthew Loh. *Business Insider*. May 25, 2024. https://www.businessinsider.com/upenn-student-olga-loiek-youtube-china-social-media-stole-face-2024-3.

[12] *Europol Innovation Lab*. 2022. https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf.

[13] *Originality.AI*. https://originality.ai/ai-content-in-google-search-results.

[14] Shirin Ghaffary. *Vox*. June 3, 2023. https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe.

[15] Renee DiResta, Josh Goldstein. *Stanford*. March 18, 2024. https://cyber.fsi.stanford.edu/io/news/ai-spam-accounts-build-followers.

[16] *European Parliament*. June 18, 2024. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[17] *European Parliament*. December 9, 2023. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

[18] *The European Union*. June 13, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

[19] *Cyberspace Administration of China*. November 11, 25. https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm.

[20] Xuezi Dan, Yan Luo. *Covington*. August 30, 2023. https://www.insideglobaltech.com/2023/08/30/labeling-of-ai-generated-content-new-guidelines-released-in-china/.

[21] *The White House*. October 30, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[22] *Department of Commerce*. February 7, 2024. https://www.commerce.gov/news/press-releases/2024/02/us-commerce-secretary-gina-raimondo-announces-key-executive-leadership.

[23] *Congress*. April 8, 2021. https://www.congress.gov/bill/117th-congress/house-bill/2395/text.

[24] *Congress*. July 23, 2024. https://www.congress.gov/bill/118th-congress/senate-bill/3696/text.

[25] Emma Roth. *The Verge*. December 27, 2023. https://www.theverge.com/2023/12/27/24016212/new-york-times-openai-microsoft-lawsuit-copyright-infringement.

[26] *Senate Committee on Commerce, Science, and Transportation*. July 11, 2024. https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content.

[27] *NIST*. February 7, 2024. https://www.nist.gov/artificial-intelligence-safety-institute/aisic-working-groups.

[28] *OpenAI*. May 7, 2024. https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/.

[29] Stuart Thompson, Tiffany Hsu. *New York Times*. June 28, 2023. https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html.

[30] *Meta*. June 12, 2020. https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

[31] *OpenAI*. July 2023. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/.

[32] Philip Marcelo. *AP News*. May 23, 2023. https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4.

[33] Cath Virginia. *The Verge*. July 11, 2024. https://www.theverge.com/2024/7/11/24196769/copied-act-cantwell-blackburn-heinrich-ai-journalists-artists.

[34] David Klepper. *AP News*. March 21, 2024. https://apnews.com/article/artificial-intelligence-deepfakes-congress-biden-7026e6223c64a042bd434b11c211b753.

[35] Hope King. *Axios*. May 31, 2024. https://www.axios.com/2024/05/30/google-cloud-pension-error.

[36] *BBC*. March 4, 2024. https://www.bbc.com/mediacentre/2024/content-credentials-bbc-verify.

[37] Trisha Datta, Binyi Chen, Dan Boneh. *Stanford*. May 1, 2024. https://ee.stanford.edu/event/05-01-2024/veritas-verifying-image-transformations-scale.

[38] Lorenzo Grassi et al. *USENIX*. 2019. https://eprint.iacr.org/2019/458.pdf.

[39] Corey Seeman. (Photo, CC BY-NC-SA). https://www.flickr.com/photos/cseeman/50707338788/.

[40] Neal Krawetz. *Hacker Factor*. December 13, 2023. https://www.hackerfactor.com/blog/index.php?/archives/1013-C2PAs-Worst-Case-Scenario.html.

[41] Antigone Davis, Guy Rosen. *Meta*. August 1, 2019. https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

[42] *OpenAI*. June 28, 2021. https://github.com/mlfoundations/open_clip.

[43] Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin, Adam Smith. *Advances in Cryptology*. May 2004. https://www.cs.bu.edu/~reyzin/fuzzy.html.

[44] Invisible Watermark. https://github.com/ShieldMnt/invisible-watermark.

[45] *Trufo*. March 21, 2024. https://medium.com/trufo/dynamic-embedding-extraction-4a941b7807ea.

[46] Ivan Selesnick, Richard Baraniuk, Nick Kingsbury. *IEEE Signal Processing Magazine*. November 2005. https://eeweb.engineering.nyu.edu/iselesni/pubs/CWT_Tutorial.pdf.

[47] Sage Lazzaro. *Fortune*. March 7, 2024. https://fortune.com/2024/03/07/ai-watermarks-arent-just-easy-to-defeat-they-could-make-disinformation-worse/.

[48] Sam Reynolds. *CoinDesk*. August 21, 2024. https://www.coindesk.com/business/2024/08/21/story-protocol-developer-raises-80m-series-b-led-by-a16z-for-intellectual-property-chain/.

[49] Jen Miller. *Princeton Alumni Weekly*. July 24, 2024. https://paw.princeton.edu/article/cindy-han-22-working-protect-artists-theft-ai.