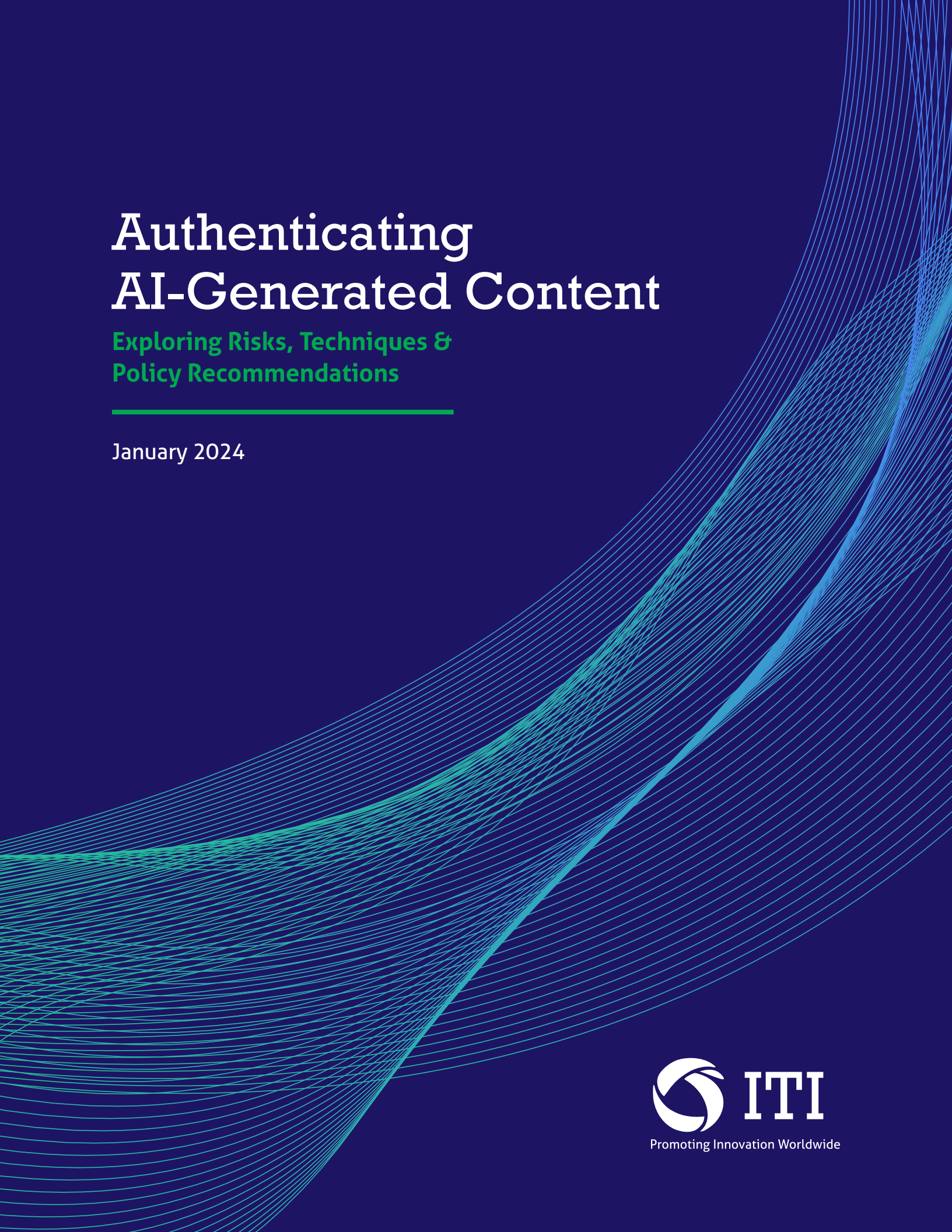
Authenticating AI-Generated Content

Exploring Risks, Techniques & Policy Recommendations

January 2024



ITI

Promoting Innovation Worldwide

Background

Since the rapid popularization of artificial intelligence (AI) chatbots like ChatGPT and image generators like Stable Diffusion, ***generative AI**** has become a global phenomenon. ChatGPT itself, created by OpenAI, became the fastest-growing consumer application in history after reaching over 100 million monthly active users in just two months.[1] While chatbots have captured much of the public attention focused on generative AI, there are many forms[2] of AI-generated content.

Many of these AI applications offer new advantages, particularly for accessibility and to augment human creativity.[3] While AI-generated content is not new, its growing sophistication and adoption — and the response from the general public, industry, and governments — has cultivated a sense of urgency to leverage the technology for social benefit and to minimize the harms that could come from its use.

This paper examines some of these benefits and risks, offers potential solutions for authenticating AI-generated content, and provides policy recommendations for advancing best practice.

**Terms that are bolded and italicized are defined in the Glossary on page 18.*

What is AI-Generated Content?

*Broadly, AI-generated content can refer to any kind of content or media (text, images, video, audio, sometimes **multimodal**, or a combination of mediums) that has been wholly or partially created with use of generative AI/ML techniques, like **natural language processing** (NLP) and/or **computer vision**. Examples of generative AI systems and AI-generated content include:*

1. Image generators which leverage text to image generation: DALL-E (OpenAI), Midjourney, Stable Diffusion (StabilityAI)
2. Chatbots that generate conversational text: ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google)
3. AI-generated audio and voice recordings, like musical samples, voice imitation technology, and audio tags to describe content for people with visual impairments
4. AI-generated video and video recordings, like video samples of lip-sync with audio dub overs, video re-ordering, video imitation, deep fake videos
5. Multimodal generation – combination of image, audio, video generated completed by AI (e.g., Noonoouri)

What is AI Authentication?

***Authentication** is a well-known practice in computing and cybersecurity, which involves "verifying the identity of a user, process, or device"[4] — and in this case — data, models, and outputs. AI authentication refers to determining when content is AI-generated, or otherwise verifying the validity or provenance of layers of the AI value chain,[5] leveraging authentication mechanisms like cryptographic methods and human verification.*

AI authentication is a new and emerging field. While there are a variety of techniques that can be used to authenticate AI, this guide explores the most robust techniques currently in development, including watermarking, provenance tracking, metadata auditing, and human authentication. Some of these techniques overlap along the AI value chain or leverage similar processes. One instance of this is ***labeling***, a process used in a variety of techniques like watermarking, provenance and metadata auditing but each with different methods — resulting in distinct forms of labeling solutions across distinct AI authentication techniques. This dynamic and other techniques will be explored later in the report.

Importantly, a combination of these methods will be the most effective way to validate and authenticate AI-generated content. However, because of the emerging nature of these techniques, as well as the dynamic nature of AI-generated content, it is unlikely that one solution alone — technical or manual, involving human intervention — will fully address the risks posed by AI-generated content.

Risks from AI-Generated Content & Value Proposition of AI Authentication Techniques

As is the case with most AI capabilities, generative AI and AI-generated content can create distinct benefits and risks. Many of the risks associated with AI-generated content — like mis- and disinformation[6] — are not new and have been observed and well-documented as problems with human-generated content on social and traditional media.[7]

Risks from AI-Generated Content

While misinformation and disinformation are often used interchangeably, they are not the same thing. It is important to distinguish between the two so that implications regarding intent or accountability are clear.

Misinformation

Misinformation refers to false, misleading, inaccurate or out of context information shared without an intent to deceive. Examples include:

- Rumors
- Satire and parody (where the reader may not understand the joke or underlying message)
- Misleading information such as comments or opinions presented as facts
- False connections or information with false content, like factually accurate content shared in an irrelevant setting (i.e. when headlines, visuals or captions do not support the content)
- Sponsored content, when the reader or consumer is not able to distinguish factual or substantive content from an advertisement
- Errors in reporting or sharing of content[8]

Disinformation

Disinformation is false information that is shared with the intent to deliberately mislead, deceive, or cause harm. Examples include:

- Propoganda
- Hoaxes
- Fabricated or completely false content
- Manipulated content, which has been purposefully distorted or changed from its initial form
- Imposter content: the impersonation of real or credible sources

Generative AI's capability to produce content at scale introduces new concerns for the spread of mis- and disinformation, which can contaminate or harm the integrity of the AI value chain, creating issues for consumers, and undermining trust in the technology ecosystem. Understanding the way in which generative AI presents risks related to mis- and disinformation requires a nuanced analysis, as generative AI can create or perpetuate mis-and disinformation in several ways through both the inputs and outputs of generative AI systems:

→ **Inputs** *(i.e., data, datasets, underlying models, or model weights):*

- **Inappropriate or under/over representative data**: Data unintentionally trained on misinformation can create wrong or inaccurate responses. Data trained on disinformation can create intentionally false outputs or encourage a model to come up with more extreme false embellishments.
- ***Data poisoning***: Whereby humans inject false or wrong data to skew a model in a way that affects the accuracy and legitimacy of its output, sometimes as a means to promote various mis- and disinformation campaigns. Data poisoning can be done with both training data and on inferencing data. Poisoning can also occur by manipulating the order in which training data is presented.
- ***Data evasion***: Whereby a model is fed an adversarial example — a carefully manipulated input that looks and feels exactly the same as its untampered copy — to cause the model to misclassify the data or intentionally fail during inference.

→ **Outputs** *(i.e., a photo, video, sound bite):*

- ***Hallucination***: Sometimes dubbed "hallucinations," AI systems can produce outputs or responses based on misinformation.[9] These hallucinations can range from models trying to fill in gaps of knowledge they do not have with false or irrelevant context, to leveraging information that is erroneous.
- ***Deepfakes***: Mis- and disinformation can also be spread through the human creation of deepfakes or other purposefully manipulated or altered content. Deepfakes are especially concerning as a tool to spread election disinformation and have already been used to manipulate footage or commentary of candidates.[10]
- **False or human-generated content that is spread or adapted by AI**: Just as AI can create false or misleading content, so can malicious human actors. If not properly labeled or moderated in a way machines can understand, AI can spread fake content generated by humans.

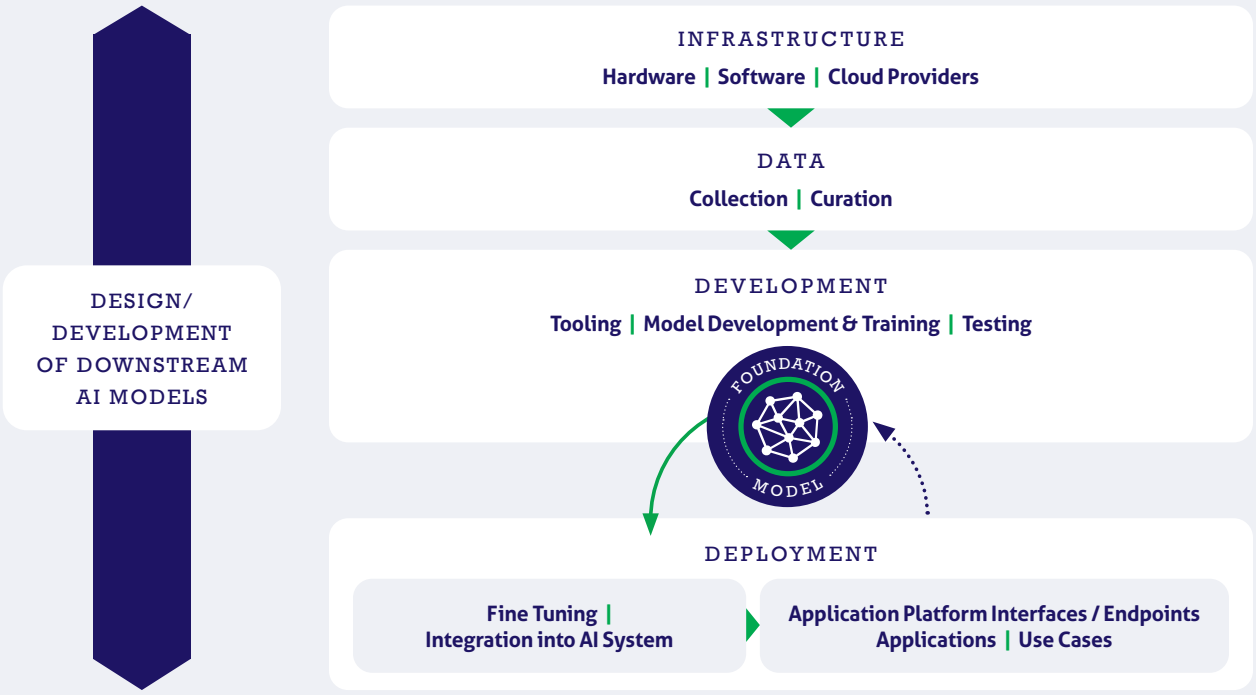
Malicious generative AI tools can be used by bad actors to generate content intended to **manipulate individuals in consumer scams**. For example, generative AI has been used to place scam calls affecting parents and the elderly. AI scam calls provoking elder fraud cases have risen with the increase of voice alteration techniques used to mimic the voices of trusted loved ones in an attempt to get a ransom or personal and/or banking information.[11] These new and sophisticated threats created by malicious actors negatively impact consumer protection by expanding the threat landscape. This can lead to an overall decrease in trust in the technology value chain.

AI-generated content, like traditional AI applications, can also perpetuate **bias and discriminatory misrepresentations**. Traditional AI applications in some contexts preserve or exacerbate societal bias in much the same way as AI-generated content can reproduce or create themes or mischaracterizations of underrepresented or marginalized groups.[12] AI algorithms are only as good as the information that is available to train them, and sometimes may be trained on content which is satirical or used to promote hateful stereotypes of individuals from historically marginalized groups in a way that AI models do not understand. This can result in AI-generated content that is inaccurate or includes unfair representations of marginalized people or cultures. Issues of bias and discrimination often arise from the training data or how the model was reinforced to understand data inputs.

Harm Reduction Through AI Authentication

To promote transparency and awareness, of AI-generated content, AI authentication techniques aim to minimize these risks and their impacts. AI authentication techniques can apply to different parts of the AI value chain, from data inputs to the model, to the application, and even to the output itself. Authentication techniques boost trust in the AI value chain, including between businesses, among consumers, and across the information ecosystem — among journalists, traditional media, and other information access points. Content verification can improve trust in generative AI and allow for healthier engagement and digital experiences online. Importantly, ***the earlier in the AI development lifecycle authentication happens, the more helpful it is to a broader range of stakeholders****. This is particularly true in instances with direct consumer or societal harm, such as fraud and election-related misinformation.*

Below is a graphic that illustrates different components of the AI value chain, including areas where authentication techniques may be able to be incorporated.



While AI authentication can help build trust in AI-generated content and the technology ecosystem, **AI authentication is still a very new field, requiring investment and implementation from government and industry. Additionally, it will require input from and engagement with consumers to promote education and awareness regarding risks and the consumer-facing authentication techniques discussed in this paper, such as visible watermarking (discussed further below)**. Similar approaches to boosting awareness have become common practice in cybersecurity, where techniques like two factor authentication and/or HTTPS browsing are now better understood by consumers as security best practices.

Survey of AI Authentication Techniques

AI authentication techniques vary in utility and ease of implementation across the AI value chain, depending on the application and use context. Below we highlight four techniques that have made headway in research and use, detailing an overview of each technique, key uses and best practices, limitations, and applicable tradeoffs.

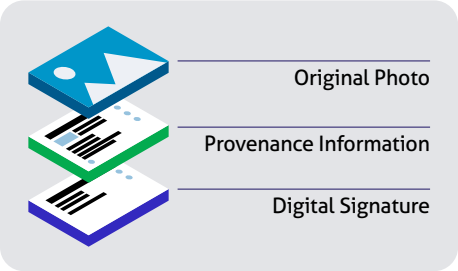
Provenance Generation and Tracking

***Provenance**, or provenance tracking, is a common technique for AI authentication that allows for the tracing of the history and quality of a dataset. Digital content provenance in particular references the verifiable chronology of the origin and history of that content, such as an image, video, audio recording, or electronic document.[13] Generally, provenance includes the origins of a dataset, any changes that have been made to alter that dataset, and other relevant details needed to ensure the integrity of the dataset and data within it. Provenance can be accomplished in a variety of ways, including embedding a signature on the metadata of the dataset which is being used to train an AI model. Each level of the data set can have a signature or content credential to ensure that the data the product used has not been tampered with. This helps to demonstrate that the training data is not compromised before it is fed into larger models, and to validate whether content has been created by an AI model trained on such data. For AI-generated content, provenance can refer to signals embedded in the dataset used to create the content as well as **content provenance:** the content's source and history, and modifications to the content subsequent to its creation.*

Like watermarking, content provenance has been a primary focus of leading industry groups focused on authentication of AI generated, synthetic, or manipulated images. In 2007, the Joint Photographic Experts Group (JPEG), a joint committee formed by international standardization organizations ISO/IEC JTC 1/ SC 29 and ITU-T Study Group 16, published the international standard "Secure JPEG 2000," which provides a syntax for security services such as source authentication and integrity verification to be applied to JPEG 2000 coded image data.[14] In 2020, JPEG began an exploration on "Fake Media" that resulted in a call for proposals for a standard that can facilitate the secure and reliable annotation of media asset creation and modifications, and intends to address use cases that are in good faith as well as those with malicious intent.[15] JPEG received six different proposals from the Coalition for Content Provenance & Authenticity (C2PA), Huawei, Sony, Vrije Universiteit Brussel/imec, UPC, and Newcastle University, the majority of which are expressly based on provenance information, and JPEG continues to work on developing a new standard to combine the best elements of each proposal.[16]

Likewise, C2PA, which includes corporations of all sizes, leading journalism entities, not-for-profits, and academics, was formed in 2021 to address concerns related to misleading information and to develop technical standards for content provenance, called Content Credentials.[17] The C2PA's voluntary, open standard, or specification, works with visual, video and audio content by binding provenance information to a piece of media at its point of creation or when it is altered and attaches to the content with Content Credentials, which is a combination of secure metadata and watermarking.[18]

At the point of media creation, the accompanying metadata or provenance information about the origins, creator, and other relevant information is attached through Content Credentials and affirmed by the media publisher or creator with a digital signature. The provenance information, signature, and media are then bound together and cryptographically sealed into a tamper-proof "manifest" that is tied to the piece of content forever. Each time the media is edited, a new, resealed manifest is created and resealed, which then accompanies the newly created media, preserving its history. These steps are visualized in the graphics below.[19]



Initial Layers of C2PA's Content Provenance Specification



Added Layers or manipulation



Cryptographically resealed media with completed provenance

As discussed in more depth later in this paper, attaching provenance to the metadata of digital content allows anyone engaging with that piece of content to understand how, when, and by whom that digital content was altered. This focus on transparency helps to foster confidence for any person, business, or government agency engaging with a piece of digital content thereby helping to engender trust in the entire content value chain, particularly AI-generated content, and allowing for consideration of the authenticity of any piece of digital content. Examples of deployment of the C2PA Content Credential technology at scale include Leica Cameras with digital photos, inclusion in various Microsoft products, such as the Bing search engine, and Adobe Firefly, Photoshop, Lightroom, and Stock.

As mentioned above, the C2PA also introduced Content Credential iconography with a unique symbol to promote tagging and transparency around AI-generated content. The icon, a small "CR", which stands for Content Credentials, will be included with the metadata of any content to clarify who owns and created the data, including regarding the AI tool used to make it. This recognizable icon or symbol will be useful to developers but also to general consumers, who will be able to hover over the recognizable mark embedded in the image to glean information about its source.[20]

Key Uses and Best Practices

1 **[Content] provenance should happen at multiple levels of the AI value chain** — at the data level, in the metadata and dataset itself, as well as on the model output.

2 **The amount of detail that is provided as part of the provenance information should be tied to the use case.** While including more detail can provide more visibility into the provenance of the content, in some use cases such information can create significant privacy concerns and may conflict with data minimization efforts. Provenance information in general should never include personal data or data that could be used to identify a person when combined with other data sets (for example location or time data) without the express affirmative consent of the user, and should be limited to the minimum amount of data required for the specific attestation and use case. If provenance information is visible on the output or content itself, closely tailoring the provenance information to the use case could also reduce the information that must be displayed and make it easier for a broader audience to understand the history of the content.

3 **Collaborative approaches to content provenance** can help to ensure that the industry is developing best practices together to promote interoperability and that they are unified in their approach.

Limitations and Trade-offs

- **Insufficient to mitigate all harms**. Though provenance can be used to authenticate that a dataset was not tampered with and/or whether it has been used to train an AI model generating content, neither provenance nor watermarking will help reduce bias in the model or dataset inputs if in fact the model or data inputs are biased or are not representative. Analysis of the data itself to ensure its integrity should be incorporated to mitigate this trade-off. While provenance tracking can also help to establish information about the origin and history of a piece of content, it cannot independently prove that the content is truthful.

- **Effective only if the vast majority of a content category has provenance data**. In order to be effective at detecting synthetic content, a consumer must be able to assume that either (i) **every** generative AI model attaches provenance data to synthetic content, such that any content without "synthetic" provenance data can be presumed to be authentic, or (ii) **every** image or document capture device attaches provenance data to authentic content, such that any content without "authentic" provenance data can be presumed to be synthetic. This is only possible in an ecosystem with an extremely high adoption rate.

- **Multiple potential points of failure**. A single weak link in the provenance chain could allow a malicious actor to alter data or insert changes without detection or insert fake attestations to sow doubt in the content. The more widespread the adoption must be, the greater the risk there is that a weak link is created.

- **Software vs hardware implementations**. Content provenance standards that rely or allow for software-only attestors at some (or all) points in the provenance chain are vulnerable to manipulation by privileged processes/actors that could allow malicious actors to manipulate these attestors into signing false attestations. Hardware based attestation may be significantly more resistant to such attacks.

Watermarking

Watermarking *is a technique that involves embedding a signal in a piece of text or an image with information. It can be visually apparent, such as an overlay showing information about its source or creator, or to identify whether it was AI-generated. Conventional examples of watermarking include a stamp on text output (i.e. a "DRAFT" or "Not for Release" stamp on a Word Document or an embargoed press release), or photo (for example, Getty images[21]).*

Alternatively, a watermark can be invisible or imperceptible. Invisible or imperceptible watermarks are a small amount of data embedded in the pixels of an image. Sometimes watermarking is also used to describe metadata, which is information in a special segment of a file that is part of an image but not embedded in its pixels. This data can be bound to the pixels via cryptographic ***hash***, and digitally signed to provide assurance that no manipulation has occurred after creation or signed modification.

Watermarking is unique among techniques in the sense that depending on the underlying tool used, it can create a level of transparency for a broad range of users allowing consumers and end users (in addition to developers) to know that a piece of content (image or text) was generated using AI. Some emerging examples of watermarking AI-generated content are described below.

Watermarking AI-generated content is a ripe field of research in the AI community and among large tech companies seeking to provide consumers and governments with information flagging AI-generated content. Watermarking is perhaps the most discussed approach and viewed as an evolving industry best practice. A number of coalitions involving the tech industry, in collaboration with partners from media and journalism, have formed to examine watermarking and other authentication techniques. For example, the Partnership on AI's synthetic media framework is a collaboration between 50 different companies working together to responsibly develop, create, and share AI-generated content.[22] Leading AI labs in the Frontier Model Forum will likely also share best practices and techniques for watermarking.[23]

Individual organizations are likewise experimenting with AI watermarking techniques. SynthID, a solution developed by Google Deepmind and Google Cloud, labels metadata of AI-generated content created with Google's developer platform, Vertex.[24] OpenAI's ChatGPT is considering the use of watermarks in the text that it generates.[25] Meta is exploring Stable Signature, which is a method for watermarking images created with open source AI.[26]

Key Uses and Best Practices

1 **Steganography** is a technique which hides a watermark or content information file inside a primary media file. One of the more common types of steganography involve embedding this hidden or secret information in the Least Significant Bit of a media file, which is done by slightly modifying or adding additional information to bytes (or bits on those bytes) of data within pixels in a media file.[27] This ensures that the watermark and relevant verification information is not visible to the eye to maintain the usability of the image and/or text itself. Steganography is often conflated with cryptography. While the techniques are similar, cryptography involves changing or scrambling information into ciphertext, which can only be unscrambled with a decryption key. Steganography involves hiding or concealing information in a way that is not readily apparent, without needing to decrypt it. Since steganography requires careful application involving specificity at the bit, byte, pixel and file level, this approach requires robust industry collaboration to be successful.

2 **Invisible forensic watermarking** can be used on video and image content by creating and encoding a unique mark into a video file that acts as a unique identifier of the recipient or consumer of the content. Should a piece of video content be AI-generated, the mark in the video file would allow owners of the content to track its spread. Like with steganography, invisible forensic watermarking promotes content use as the marked content is accompanied with verification. This is a very popular tool used in the entertainment industry for tracking copyrighted productions and videos and to prevent unauthorized pirating or distribution.[28]

3 **Dataset watermarking** can be used to track and verify input and training data that will be used for model development by tracing the dataset back to its original owner or creator. Watermarks are typically embedded into a subset of a dataset during a model training process, and distributed to ensure they remain even when filtering or reorganizing the data (and so that they can't be easily removed). Watermarks in datasets help organizations track dataset ownership and provenance and can be used for ML model verification.[29]

4 **Model watermarking** can be used to track and verify ownership of an AI model or otherwise track its usage by embedding information within the model ***parameters*** or structure of the model itself. This technique primarily addresses concerns related to unauthorized use, distribution, or modification of a model.[30]

5 **Differential watermarking** combines various watermarking techniques by targeting them at different elements (data, metadata, content) of either inputs or outputs and ensuring each has a unique signal. A benefit of using differential watermarking is that it allows AI systems to cite the sources of factual text more accurately through the watermark itself, as it facilitates the direct sourcing of content. This is achieved by embedding the watermarks in the output and in the metadata as well.[31]

Limitations and Trade-offs

- Watermarks can **validate elements of the AI value chain to the extent that humans know that they are real**. Watermarks can be placed on AI-generated content and data, but do not prevent the spread of false information if the AI-generated content is in fact fake or if it relies on bad data (but is watermarked nonetheless).
- Visible watermarks on image content **can be easily cropped out or removed** with AI-enabled watermark removers, diminishing the value of a perceptible watermark. Additionally, watermarks **can be "mimicked"** on images, leading to the spread of false information that may lead others to think it is legitimate.
- Invisible watermarks on images are **easier to remove** than invisible watermarks on other types of content. For images, the invisible watermark is usually embedded into pixel bytes and bits of the image, so if the format of the file changes or is altered through a screenshot or other image format conversion, the invisible watermark is no longer effective.[32]
- Invisible or embedded watermarks created through labeling or adding additional information to a piece of content's metadata **can create *noise***, generally unnecessary information in a dataset which necessitates greater amounts of storage space. This can sometimes make it harder to download and access certain types of content, particularly with low-latency bandwidth or a system that does not have strong working memory or abundant storage.[33]
- **Visible watermarks on fluid content like video are much harder to remove.** However, this type of watermark can be distracting and make content consumption less appealing as it is visible to consumers of the content. Required watermarks can also distract from the primary content, particularly in advertising and marketing purposes where, for example, a business wants to generate an image or icon of a product to share in a campaign.

Figure 1: AI Authentication Along the AI Value Chain



Content (as seen by human eye)	Visible watermark	Getty Images tag
		Hover-over Content Credential
	Content label	
Metadata	Embedded watermark	Steganography
		Metadata Content Credential
	Metadata Auditing	
Model	Embedded watermark	Challenge/response signals embedded in the model
Training Dataset	Dataset watermark	Noise/signals distributed in data
Application / API	Embedded watermark	Metadata content credential
Computing Infrastructure	Attested SBOM	Metadata content credential with attestation certificate

Human Authentication

*Sometimes referred to in AI discussions as "human in the loop," **human authentication** requires human involvement to verify whether content has been AI-generated at or across certain points of the AI value chain. A straightforward example is requiring human validation of content that has been flagged as potentially AI-generated. Human authentication can be used in a broad range of contexts, to validate model inputs (datasets or metadata) or outputs (content).*

Human authentication is useful and already part of the AI value chain: generally, cleaning and labeling datasets or inserting credentials into metadata is a highly manual process requiring human oversight. While automation in the context of data mapping and governance is becoming more robust, labeling datasets, assessing outputs, validating models, and performing ***reinforcement learning with human feedback*** (RLHF) are all still very manual processes requiring human skill and involvement. Human intervention, including the involvement of diverse human perspectives in AI development generally, is a widely accepted mechanism to prevent issues like bias and real-world harm – and to improve product accessibility and use.

Key Uses and Best Practices

1 **Involves diverse groups of people in the authentication process**. A variety of people from a range of backgrounds and subject matter areas will improve the quality of human authentication processes. Some people will spot things others will not.

2 **"Selective intervention,"** by which human authentication is used only in essential parts of the AI authentication process, to ensure that human intervention is targeted and focused so as to maximize human resources.

Limitations and Trade-offs

- Human authentication and review is an inherently manual process. Human intervention **may slow or delay authentication** of AI-generated content. Despite this dynamic, human intervention is generally regarded as one of the most effective ways to minimize the unintended risks related to AI.
- Human authentication **may not be reliable** in instances where AI is purposely leveraged to deceive humans.
- Humans possess inherent and implicit bias which will impact any kind of human-involved process. Human intervention will **not fix the presence of inherent or implicit biases** and will certainly not be consistent across an organization.

Other Techniques of Note

Many techniques are emerging to authenticate AI-generated content, but not all have reached a level of robustness or attention comparable to the techniques discussed above. The techniques below have been proposed to help authenticate AI-generated content, but further research and interrogation is needed to determine their effectiveness.

1. **Distributed Ledger Tech/Blockchain Technology**: Provides tamper-resistant data storage and verification, making it difficult to alter data once it's recorded on the blockchain. As this technique leverages blockchain technology, it requires individuals to fully comprehend it and have access to the technology, making it a less effective tool. In addition to this barrier to entry, the same sort of metadata tracking can be done via technological mediums that don't require blockchain. Lastly, this technique also does not protect against bias in the data, and due to the anonymous nature of blockchain technology, the identity of who alters the ledger could be difficult to track.[34]

2. **Statistical detection**: A method of identifying statistical anomalies in the distribution of pixels, speech frequencies, etc. that correlate to artificially generated content. Statistical detection is difficult to implement because it requires a baseline of "normal" frequencies which is arbitrary in nature. Statistical detection carries a higher rate of false negatives (of AI-generated content), making it less ideal as a method of use for authentication purposes.[35]

3. **Behavioral analysis**: A technique that looks for natural patterns of imperfection, rhythm, cadence, or other typical behaviors in language or image generation to differentiate human output from AI output. Though it is similar to human authentication and can certainly be used as a type of human authentication, it is narrower in scope. Automating a behavioral analysis process would require enough — i.e., massive volumes — of training data for an AI tool to decide if the image or language content was artificially created. The technology has not yet matured enough to make these determinations independently.

4. **Turing testing**: A method to challenge AI-generated content or verification agents with CAPTCHA-style tests or open-ended questions a human would easily answer but an AI may struggle to solve. Turing testing is said to be a thing of the past because many AI algorithms can be trained to pass these confined or narrow tests.[36]

5. **Leveraging AI to detect AI-generated content**: At the same time that generative AI can contribute to misinformation, AI can also help to combat misinformation. Indeed, AI tools can help to assess the trustworthiness of digital content. Companies are in the process of building software and programs that analyze things like form and content.[37] However, more research and development is needed to hone these tools and improve their accuracy in detecting deepfake and other AI-generated images.

Policy / Best Practice Recommendations

Authenticating AI content is a newly emerging field — ripe for innovation and the development and adoption of best practices. In line with our 2021 Global AI Policy Recommendations[38] and Policy Guide for Understanding Foundation Models & the AI Value Chain published earlier this year,[39] ITI recommends that as governments consider AI regulation, they first scope and understand specific harms, keeping the impact on innovation top of mind. As policymakers contemplate the development of regulation and guidelines regarding AI-generated content and authentication, we offer the following considerations:

1 Enable and allow innovation in AI authentication to grow and thrive.

Policymakers should be careful not to be overly prescriptive in establishing regulation in this area because state-of-the-art technology for verification and authentication of AI content is still being developed and studied, and some authentication systems are only beginning to be deployed at scale. While AI authentication techniques offer promising solutions to verify and assure content, more research and investment are needed in this area before a definitive recommendation can be made as to the type of authentication that should take place. It may be the case that multiple types of authentication are appropriate for a given output. For example, the C2PA Coalition supports both watermarking and content credentials to include content provenance to engender public trust in digital content in the age of ubiquitous AI. Any approach to oversight or control in this domain has the potential to become quickly outdated, especially as authentication techniques will continue to evolve alongside the technology itself.

2 Promote consumer transparency and awareness around AI-generated content.

While industry is in the process of developing solutions for AI authentication, no one technical fix will be a silver bullet. Consistent with our AI Transparency Policy Principles,[40] we believe that consumers should understand when a piece of content is AI-generated, and also be familiar with the tools provided to them to do so. More broadly, users should be informed of the capabilities of the AI system, including what it can and cannot do. Governments should likewise support guidance and programs to consumers to spread consumer literacy and boost the public's ability to determine when content is AI-generated.

3 Leverage public-private partnerships to understand both the opportunities and limitations of various authentication techniques.

Public-private partnership and collaboration remains important to ensure that policymakers better understand the state-of-play of AI authentication tooling. Demonstrating the progress and limitations of various authentication methods will help inform policymakers as they consider relevant policy tools and interventions in this area, and boost trust in and collaboration among the public and private sectors.

4 Recognize that AI authentication is a shared responsibility.

Collaboration across industry and with other stakeholders like government bodies and academic researchers will be key to success in this area. Government-backed studies and stakeholder engagement via multistakeholder processes or working groups in collaboration with affected communities and the technical ecosystem is a helpful approach to advancing these techniques.

5 Foster interoperability and collaboration across AI authentication techniques.

One technique alone will not be enough to ensure that a piece of content has or has not been created by AI. Policymakers should ensure any regulatory approach takes into account the variety of techniques needed for effective content authentication. Similar to the zero-trust model in cybersecurity, there should be a mixed-method approach to AI authentication utilizing the techniques discussed above.

6 Incent and augment humans in the loop.

Human authentication is essential to minimizing the potential negative impacts of AI tools and content, particularly when it comes to unintended risks. Policy and research should explore how human authentication can be augmented at the data, model, and output levels as described (including with the help of AI).

7 Invest in the development of clear, voluntary industry consensus standards for AI Authentication.

While industry is working on the development of standards and regimes to watermark and apply provenance to and authenticate AI content, involvement from voluntary technical standards organizations, like the International Standards Organization (ISO) and experts from bodies like NIST to help develop and test best practices will help promote consistency, interoperability, and collaboration across techniques.

Glossary

Authentication: "An act, process, or method of showing something to be real, true, or genuine"[41] and in the case of AI, data, models, and outputs, the act of verifying or confirming the authenticity of content generated by AI models.

Computer Vision: A field of computer science that enables computers to identify and understand various objects and features in images and videos.

Content File: A digital file or set of files accompanying a piece of content or media which includes the content itself, metadata, and any other relevant content.

Content Provenance: The chronology of ownership, custody, or location of a piece of content.

Data Poisoning: Method used to manipulate or distort datasets and AI models by inserting corrupted information or data into the model training process. If performed maliciously, objectives of data poisoning are primarily to trick models to bias their outputs.

Deepfakes: The result of distorting or manipulating AI models or data to generate convincing image, audio, or video content that appears real but is in fact misrepresentative or false.

Disinformation: False information that is spread with the intent to mislead individuals into believing it is true.

Generative AI: A type of artificial intelligence that can create or produce different types of content, including text, imagery, audio, and synthetic data based on patterns and predictions in the data.

Hallucination: When an AI model generates false or misleading information but presents it as if it were a fact and/or truthful.

Hash or Hash Value: A numeric or alphanumeric value or sequence of a fixed length that uniquely identifies data.

Human-in-the-loop: The dynamic that describes the utilization or involvement of humans at any stage of the AI development or authentication process.

Labeling: A process used across a number of AI authentication techniques which involves sharing relevant context along with a piece of content or information, often in the form of a sticker, explanation, or tag embedded alongside the content.

Large Language Model (LLM): LLMs are a type of foundation model that can generate text after being trained on massive amounts of unlabeled, textual data.

Misinformation: False information that is spread without the intent to mislead individuals into believing it is true.

Multimodal content: Content that uses different mediums, i.e., a combination of text, audio, and video.

Natural Language Processing: An interdisciplinary subfield of artificial intelligence and linguistics that provides computers the ability to comprehend text and speech.

Noise: Unnecessary, excessive, or meaningless information in a piece of data or dataset, which necessitates greater amounts of storage and parsing, sometimes convoluting a dataset.

Reinforcement Learning with Human Feedback (RLHF): Process in machine learning training that involves "rewarding" a model when it performs well or as desired to optimize the model. RLHF can be highly contextual depending on the organization and model.

Parameter: A variable in an AI model, the impact of which can be adjusted during training to influence how new input data is transformed or analyzed into output.

Provenance: The chronology of ownership, changes / manipulation, and location of an object or thing.

Sidecar File: Files that accompany a content or source file which store metadata that is not supported by the format of a content or source file.

Watermark: A technique used to embed or attach a signal or information in a piece of content to identify its validity or source.

References

1 Smith, J. (2023, February 1). ChatGPT sets record as fastest-growing user base: Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

2 NVIDIA. (n.d.). Generative AI. NVIDIA. Retrieved 13 December 2023, from https://www.nvidia.com/en-us/glossary/data-science/generative-ai/

3 For example, AI-generated content can be used to create alternative text (alt text) to label any non-text content. See: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/use-case-alt-text

4 National Institute of Standards and Technology. (n.d.). Authentication. NIST Computer Security Resource Center. Retrieved 13 December 2023, from https://csrc.nist.gov/glossary/term/authentication

5 See page 5 of our report on Understanding Foundation Models & The AI Value Chain: https://www.itic.org/documents/artificial-intelligence/ITI_AIPolicyPrinciples_080323.pdf

6 Mis-and dis-information is referred to as false information that could be consequential or inconsequential.

7 Misinformation Integrity Institute Report. The New York Times. (2022, October 13) https://www.nytimes.com/2022/10/13/technology/misinformation-integrity-institute-report.html

8 UNHCR Fact Sheet on types of misinformation and disinformation: https://www.unhcr.org/innovation/wp-content/uploads/2022/02/Factsheet-4.pdf

9 Techopedia. (n.d.). AI Hallucination. Techopedia. Retrieved 30 November 2023, from https://www.techopedia.com/definition/ai-hallucination#:~:text=An%20AI%20hallucination%20is%20where,outputs%20from%20large%20language%20models

10 Swenson, Ali. (2023, August). FEC Warns of Potential for AI-Generated 'Deepfakes' in Election 2024. AP News. https://apnews.com/article/fec-artificial-intelligence-deepfakes-election-2024-95399e640bd1e41182f6c631717cc826

11 National Institutes of Health (NIH). (2023, December). Beware of Virtual Kidnapping Ransom Scam. Retrieved from https://ors.od.nih.gov/News/Pages/Beware-of-Virtual-Kidnapping-Ransom-Scam.aspx; Dudha, A. (2023, June 18). Scammers can easily use voice-cloning AI to con family members: expert. CBC News. Retrieved from https://www.cbc.ca/news/canada/saskatoon/fraudsters-likely-using-ai-to-scam-seniors-1.6879807

12 Qadri, R., Shelby, R., Bennett, C. L., & Denton, E. (2023, June 12). AI's regimes of representation: A community-centered study of text-to-image models in South Asia. Proceedings of the ACM Conference on Human Factors in Computing Systems, 1–16. [https://dl.acm.org/doi/abs/10.1145/3593013.3594016].

13 See S. 2559 - Deepfake Task Force Act, definition of digital content provenance.

14 International Organization for Standardization. (n.d.). Artificial Intelligence: AI/TS 37382 – Requirements for trustworthy AI. ISO. https://www.iso.org/standard/37382.html

15 Joint Photographic Experts Group (JPEG). (n.d.). JPEG Fake Media. Retrieved from https://jpeg.org/jpegfakemedia/

16 Joint Photographic Experts Group (JPEG). (2022, December 11). JPEG initiates specification on fake media based on responses to its call for proposals. Retrieved from https://jpeg.org/items/20221211_press.html: https://jpeg.org/items/20221211_press.html.

17 Joint Development Foundation. (n.d.). Coalition for Content Provenance and Authenticity. Retrieved from https://c2pa.org/.

18 Ibid.

19 See C2PA technical specification here: https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html.

20 Joint Development Foundation, Coalition for Content Provenance and Authenticity (C2PA). (n.d.). Introducing Official Content Credentials Icon. Retrieved from https://c2pa.org/post/contentcredentials/.

21 Getty Images. (2023, December 12). Getty Images launches innovative new watermark to combat image misuse. PR Newswire. https://www.prnewswire.com/news-releases/getty-images-launches-innovative-new-watermark-153637795.html.

22 See the Partnership on AI's Synthetic Media Framework here: https://syntheticmedia.partnershiponai.org/

23 See more information about the Frontier Model Forum here: https://frontiermodelforum.org/

24 Google DeepMind. (2023, August 29). Identifying AI-generated images with SynthID. Retrieved from https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

25 The New York Times. (2023, February 17). Can AI tell if text is real or fake? The race is on. Retrieved from https://www.nytimes.com/interactive/2023/02/17/business/ai-text-detection.html.

26 Meta AI. (2023, October 26). Stable Signature: A new method for watermarking images created by open source generative AI. Retrieved from https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/.

27 V. Verma, Poonam and R. Chawla, "An enhanced Least Significant Bit steganography method using midpoint circle approach," 2014 International Conference on Communication and Signal Processing, Melmaruvathur, India, 2014, pp. 105-108, doi: 10.1109/ICCSP.2014.6949808.

28 Moxion. (n.d.). Forensic Watermarking. Retrieved from https://help.moxion.io/article/121-forensic-watermarking#:~:text=Forensic%20watermarking%20is%20a%20security,the%20specific%20user%20viewing%20it: https://help.moxion.io/article/121-forensic-watermarking#:~:text=Forensic%20watermarking%20is%20a%20security,the%20specific%20user%20viewing%20it.

29 Wang, K., Liu, Y., & Zhang, X. (2022, February 16). On the effectiveness of dataset watermarking in adversarial settings. arXiv preprint arXiv:2202.12506.

30 Boenisch, F. (2021). A Systematic Review on Model Watermarking for Neural Networks. Frontiers in Big Data, 4:729663. https://doi.org/10.3389/fdata.2021.729663

31 Wiggers, K. (10 December 2022). OpenAI's attempts to watermark AI text hit limits. TechCrunch. Retrieved from: https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/

32 Emmons, L. (20 December 2021). Invisible Watermarks: Are They Worth It? LightRocket. Retrieved from: https://www.lightrocket.com/blog/invisible-watermarks-are-they-worth-it

33 Nanjundan, P., & George, J. P. (2022). Text Watermark Analysis – Concept, Technique, and Applications. In J. Sen & J. Mayer (Eds.), Information Security and Privacy in the Digital World - Some Selected Topics. IntechOpen. DOI: 10.5772/intechopen.106914

34 Finance Strategists. (7 September 2023). Blockchain and AI. Finance Strategists. Retrieved from: https://www.financestrategists.com/wealth-management/blockchain/blockchain-and-ai/; also see Futureworks.news. (n.d.). Blockchain-Based Digital Identity: Benefits, Risks, and Implementation Challenges. Retrieved from https://future.works/blockchain-based-digital-identity-benefits-risks-and-implementation-challenges

35 PromptEngineering. (n.d.). The Truth About AI Detectors: More Harm Than Good? PromptEngineering. https://promptengineering.org/the-truth-about-ai-detectors-more-harm-than-good/

36 Beiever, C. (25 July 2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI. Nature. Retrieved from: https://www.nature.com/articles/d41586-023-02361-7

37 European Commission. (2021 July 7). Can artificial intelligence help end fake news? Research and Innovation. Retrieved from https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news

38 Information Technology Industry Council. (2021 March). ITI's Global AI Policy Recommendations. Retrieved from: https://www.itic.org/documents/artificial-intelligence/ITI_GlobalAIPrinciples_032321_v3.pdf

39 Information Technology Industry Council. (2023 August). Understanding Foundation Models & the AI Value Chain: ITI's Comprehensive Policy Guide. Retrieved from: https://www.itic.org/documents/artificial-intelligence/ITI_AIPolicyPrinciples_080323.pdf

40 Information Technology Industry Council. (2022 September). ITI Policy Principles for Enabling Transparency of AI Systems. Retrieved from: https://www.itic.org/documents/artificial-intelligence/ITIsPolicyPrinciplesforEnablingTransparencyofAISystems2022.pdf

ITI
Promoting Innovation Worldwide

The Information Technology Industry Council (ITI) is the premier global advocate for technology, representing the world's most innovative companies. We promote public policies and industry standards that advance competition and innovation worldwide.