Watermarking and Metadata for GenAI Transparency at Scale: Lessons Learned and Challenges Ahead

Elizabeth Hilbert (Meta)**, Gretchen Greene** (Meta)**, Michael Godwin** (Magnit)**, Sarah Shirazyan (**Meta)
Correspondence to: kggreene@meta.com

Abstract

The proliferation of generative-AI ("GenAI") technology promises to revolutionize content creation across online platforms. This advancement has sparked significant public debate concerning transparency around AI-generated content. As the difference between human-generated and synthetic content is blurred, people increasingly want to know where the boundary lies. Invisible and visible watermarks, content labels, and IPTC and C2PA metadata are some of the technical approaches in use by Meta and by the industry at large today to enable transparency of AI-created or AI-edited content online. This paper examines Meta's approach to marking AI content and providing user transparency, highlighting lessons learned–and the challenges ahead–in striving for effective AI transparency, including suggestions for research areas most likely to advance industry solutions for indirect disclosure and user transparency for GenAI content. Key challenges have included the lack of robustness of metadata, imperfect robustness of watermarks, difficulty in defining "materiality" for AI edits, and how to provide users appropriate transparency, and evolving understanding and expectations over time. We provide details of Meta's experience launching labels for first- and third-party content–both fully AI generated and AI edited–at a global scale using GenAI signals from IPTC, C2PA, and known invisible watermarks and the challenge of meeting user expectations related to materiality of edits and choice of language, resulting in changes to our approach. This paper focuses specifically on transparency related to user generated content that is non-commercial in nature.

1 Introduction

Meta's platforms are used by billions of people to learn new things, create and share content, and connect with others. As a key player in the AI industry, Meta plays a dual role: it not only builds its own Generative-AI ("GenAI") products and foundational models but also distributes synthetic or AI-generated media. For instance, we develop technology that enables users to create synthetic media through our own GenAI features such as Imagine. Meanwhile, users and creators share synthetic media created outside of our platforms across our social networking and messaging products, including Facebook, Instagram, and Threads. This dual role has informed our approach to AI transparency.

We've implemented a number of transparency measures for GenAI content on our products and services to suit this hybrid role, and we employ both direct and indirect methods of disclosure to ensure transparency. The approach we take to transparency depends on whether content was created using our own AI tools ("first-party" or "1P" content) or with an outside tool and distributed on our platforms ("third-party" or "3P" content). This distinction is due to the asymmetrical technical limitations in the data available for first-party GenAI content versus that available for third-party content. Because we have more detailed information about content created with our own AI tools, we have more control over how we disclose when content has been generated or modified by AI.

First-party content. When photorealistic images are created using first-party GenAI features such as Imagine, we have tried several things to inform people that AI is involved, including putting visible burnt-in watermarks on the images, as well as embedding invisible watermarks *in image content* and metadata *within image files*. Using invisible watermarking together with metadata in this way is aligned with Partnership on AI's best practices for improving robustness (Building a glossary for synthetic media transparency methods. Partnership on AI, 2023) and the inclusion of metadata enables other platforms to easily identify GenAI content when it is shared online. We also may add a visible label indicating that AI tools have been used to shape content when that content is shared across our own platforms.

Third-party content. For 3P content (uploaded to our platforms), in May 2024, we launched industry-leading tools to identify invisible, "indirect" markers at scale–reading the "AI generated" and "AI edited" information in the C2PA and IPTC technical metadata standards for images, and provided the option for users to self-disclose that content they upload was made with AI. When we detect metadata signals indicating that content shared on our platforms was fully generated with AI, or if users self-disclose their content as such, we currently apply a visible label to that content. In addition, if we determine that digitally created or altered image, video or audio content creates a particularly high risk of deceiving the public on a matter of importance, we may add a more prominent label, so people have more information and context.

2 Putting transparency into practice

In our efforts over the last three years to deliver on user and stakeholder expectations for greater transparency about GenAI content, we have engaged with experts globally, surveyed academic research and talked with peer companies to understand what techniques for creating transparency might be available to us, and to understand which techniques might best enable our particular use cases. To summarize what we've recounted above: our users may engage with our platforms to receive content ranging from viewing compelling photographic imagery, to connections with family, friends, and groups that share common interests, to consumption of news and entertainment. They also may engage with creators who use our platforms and platform tools to create new content for both commercial and non-commercial purposes.

We also conducted our own research with users, creators, and external experts regarding GenAI transparency. Our research findings suggest that concerns about GenAI transparency are less about whether a piece of content has been GenAI-created or altered by GenAI tools and more about whether such content–particularly visual content–is "real." This is an important distinction because it suggests that transparency measures should be designed to solve a narrower problem than simply labeling GenAI content, and has led us to focus on GenAI tools that could make realistic visual content (e.g., Imagine), and on media types that may be more likely to mislead people and where provenance and transparency could help reduce that risk (i.e., images, video, and audio).

We started our content transparency and provenance work amid global concerns about GenAI content's potential impact on elections in 2024, a significant election year worldwide. In the United States, despite early fears of AI's potential to fuel disinformation campaigns and deepfakes (The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election. Yan et al., 2025), research and post-election analysis revealed these fears were largely overblown and AI did not significantly disturb elections (We looked at 78 election deepfakes: political information is not an AI problem. Kapoor & Narayanan, 2024). This shift in perception highlights evolving public expectations–not only about the technology itself but also about the type of transparency required and the need to remain flexible in this domain.

The available palette of transparency/provenance tools–including the IPTC metadata framework for images as well as steganographic and/or visible watermarking–was comparatively well-understood and technically straightforward to apply initially to the massive volume of images that was the most commonly uploaded or created type of digital content collectively produced by Meta's more than 3 billion users worldwide in 2024. (N.B.: in the same period in which Meta focused initially on provenance measures for uploaded and

on-platform-created images, Meta FAIR published research that advanced the scalability of digital-video–see Video seal: open and efficient video watermarking, Fernandez et al., 2024. Also in this period, Meta FAIR researchers developed a model for digital-audio imperceptible watermarking–see Proactive detection of voice cloning with localized watermarking, San Roman et al., 2024.)

We also understood that our system of invisible watermarks had asymmetric applicability–to the extent we wanted it to be useful for our provenance purposes, we couldn't share everything about our watermarking model with other actors in the same market without some risk of undermining that usefulness. If, for example, we shared our *particular* invisible-watermark (IW) model with others in the industry, perhaps through open-source licensing, disclosure of the particular elements of the model would run the risk of enabling bad actors to counterfeit or remove the marks intended to be signals about GenAI provenance for those images. So we looked to additional provenance and transparency systems that would allow us to communicate to external stakeholders the GenAI aspect of content created or hosted on our platforms. This led us to deploy the IPTC metadata system for content, already in wide use in the industry, for marking on-platform-created content and for reading GenAI signals in uploaded image content.

Although some invisible watermarks are comparatively robust against common transformations (e.g., screenshots) from benign users and even from adversarial attacks, metadata is more fragile, and can be more easily removed or altered, whether intentionally or unintentionally. So the price of engaging in a widely adopted industry metadata standard like IPTC, or in an emerging metadata standard like C2PA, would be the inherent (comparative) lack of robustness in provenance and transparency systems that rely on metadata alone.

Based on this, we initially deployed visible and invisible watermarks as well as IPTC metadata for on-platform (1P) content, and enabled reading IPTC and C2PA metadata and known invisible watermarks on uploaded image content. We ultimately produced a case study in 2024 for the Partnership on AI that provided an initial summary of what we learned in our iterative efforts to address concerns about transparency and authenticity. In the next section of this paper, we focus on that case study and what guidance it has given us, and perhaps may give to other researchers, as we continue to explore methods of helping users recognize when GenAI is used in content production or dissemination on our platform.

3 Lessons learned and challenges

We currently employ a multi-pronged approach to content transparency to increase robustness and deliver the correct level of transparency and disclosure across a wide variety of use cases for synthetic media. However, in a fast-evolving space, our journey is just beginning; we have already learned a number of lessons (see below) and made changes to our approach accordingly.

3.1 Reliably identifying GenAI content at scale is challenging

Robustness. We are using state-of-the-art invisible watermarking and metadata technical solutions, which are in line with Partnership on AI's best practices (Partnership on AI, 2023) for our first-party and third-party approaches, but–as discussed above–the technical solutions available for machine-readable signals all have moderate to severe robustness issues: unintentional and intentional stripping, editing, and counterfeiting. Visible watermarks can also be easily removed, changed, or counterfeited.

Security/accessibility tradeoff. Invisible watermarks are a valuable tool for platforms to understand content provenance, but they can be counterfeited if the ability to read them is shared. If a company or platform tells you how to read their invisible watermark, they've told you how to strip it and how to counterfeit it, creating a security/accessibility tradeoff that limits the use of invisible watermarks where they could otherwise be the preferred choice.

Incomplete coverage for third-party content. While many platforms and tools now attach industry standard markers such as C2PA and IPTC metadata, this coverage is not universal.

Many AI tools choose not to attach these signals for images, and such signals are not yet deployed at scale for video and audio.

3.2 Providing users appropriate transparency is also challenging

When Meta first launched transparency labels for third-party GenAI content distributed on Facebook, Instagram, and Threads in May 2024, we quickly learned a number of valuable lessons about how users perceive GenAI content transparency. These "Made with AI" labels, indicating that AI tools were used with the content, were automatically applied to images when we detected industry-standard AI image indicators, or metadata (such as from the C2PA or the IPTC). Soon after launch, many creators and users began expressing surprise to encounter a "Made with AI" label automatically applied to their images via metadata, because they either didn't believe that they had used AI to produce the content in question, or else didn't believe they had used it in any significant (that is, "material") way.

In response to this feedback, we began investigating the behavior of the labels, and found a common theme–the use of image-editing tools such as Adobe Photoshop. These images had C2PA metadata attached, as many GenAI assisted editing functions are now integrated into image editing tools. For example, some images that included minor, non-material AI modifications, such as basic retouching or color correction, included industry standard metadata indicating AI use and thus were labeled as having been "made with" GenAI tools. In some cases, creators knew that they were using an AI-powered editing tool, but didn't consider the degree of the edits to the original photo materially significant enough to warrant a label indicating that the content was made with AI. In other cases, creators were not conscious of any AI editing at all, but the image editing tools they used were powered with AI under the hood. We find the following lessons about transparency to be compelling and instructive for other practitioners.

Design of transparency labels matters–and can backfire. Our implementation of transparency labels for AI generated and edited content aimed to provide users with neutral information about content they saw to help them make informed decisions. Nevertheless, many users and creators, particularly those in creative industries such as fashion and entertainment, did not perceive the label as neutral–instead, they felt the label indicating AI use was stigmatizing and undermined their creative work.

This points to an important distinction between applying labels to explain the content-production process vs. applying labels based on the potential for content to mislead or deceive. Academics, particularly those specializing in cognitive science and social psychology, have emphasized the importance of clear labels that help explain the production process as a means of reducing the potential for post to mislead (Labeling AI-generated content: promises, perils, and future directions. Wittenberg et al., 2023).

In response to this user and creator feedback, we decided we would update the design of the visible label to be more clearly inclusive of both AI generated and edited content–changing from "Made with AI" to "AI Info". The intent behind this short-term change was for our labels to read more neutrally while we worked to improve transparency holistically. In parallel, we made updates to the visible burnt-in watermark on our first-party AI-generated media, testing a less obtrusive approach which encourages brand attribution and awareness of AI use, without stigma.

Placement of transparency labels matters. In response to user feedback about our labels not accurately representing the use of GenAI to create content, we aligned on an updated approach to labeling for content that was *edited with*, but not *fully created by*, GenAI. We approached this by differentiating between the industry-standard metadata types for AI edited versus AI created content. The label for edited content, accessed through contextual menus, would continue to provide transparency for those seeking more information about AI content. Labels for content wholly created by AI would remain in their original placement, on the surface of the content.

These decisions were informed by our user and stakeholder research indicating that people expect greater transparency for wholly created synthetic content, as opposed to edited content. We also performed an analysis of AI edited content receiving the "Made with AI" label, and found that a large majority of the AI edited content posted to our platforms contained edits

that would be considered cosmetic or artistic–in other words, edits that did not meaningfully change the context of the content.

Defining materiality is challenging. Users and stakeholders have told us that they have differing expectations for transparency based on the type of AI-generated or AI-edited content, and we've adjusted our approach accordingly. Specifically, we've learned that there are lesser expectations of transparency for certain types of content, such as content with minor AI alterations. We also frequently see this distinction with different treatment in draft legislation for "significant" AI edits, without defining "significant." However, the current state of the technology when it comes to metadata means that we–as a company and as an industry–are not able to differentiate between minor edits, like color correction, and more significant changes which could potentially mislead people–like using AI editing tools to portray an event which didn't happen, or a person doing something they didn't do.

This distinction isn't as simple as understanding the amount of pixels in an image that were changed–some changes which might seem minor, such as changing a small percentage of pixels or cropping an image, could significantly affect the content's meaning, depending on the context. Even with technical advancements to distinguish at scale between aesthetic and material edits, no single company should create their own definitions of materiality or nonmateriality of AI use in editing. Divergent definitions and transparency approaches across online platforms have the potential to confuse users and create high cognitive overhead, ultimately failing to deliver the needed context to help people make informed decisions about the content they see. That is why we're continuing to lean into partnerships with groups like the Partnership on AI, C2PA Steering Committee, and MLCommons, among others, to encourage the collaborative creation of best practices.

4 What we hope for as we look ahead

We are working together towards a standard but at least today, different technical approaches may work better for different companies. We have learned we can make progress towards cross-industry interoperability without necessarily converging on a uniform approach. For example, our organization reads GenAI signals from IPTC, C2PA, and known invisible watermarks. There remains a need for flexibility in accommodating rapidly evolving technology, as well as for industry-wide collaboration about common, shareable approaches to communicating content provenance and transparency regarding the use of GenAI tools.

The need for flexibility across industry and across time, interoperability, and consideration of all actors. Based on our work and collaboration with industry peers, a successful GenAI provenance ecosystem will need to include 1) flexibility, 2) interoperability, 3) iteration, and 4) consideration of all actors. All of the current solutions have considerable downsides depending on use case. Across the industry, we are already approaching alignment on a limited family of available technical approaches for indirect disclosure and, with a flexible approach, companies can pick the solution or solutions that work best for their business, lowering costs and increasing participation. GenAI provenance is an evolving field with developing standards and active work happening across industry. While companies should take steps today, we should not lock in any one approach, because a much better one may emerge soon, and the rapid pace of technical innovation and the advancement of capabilities may call for a different approach.

An evolving approach to GenAI-content labeling. Our approach continues to evolve as technologies improve; we learn from research and experience; and user knowledge and expectations change. As users come to expect GenAI content, they may feel less strongly about labeling and we expect labeling to be less useful as the percentage of online content which is GenAI increases. GenAI is embedded in much of the content and media people encounter online, and the prevalence should only be expected to increase. With different companies choosing different approaches, different actors in the ecosystem will need to read and exchange provenance information using more than one approach to have the overall desired effect. There are many actors in the ecosystem that can contribute to the trustworthiness of generative AI, including creators and users, model and tool makers, and platforms/distributors of content. They each have different and important roles to play.

The challenge of implementing transparency. This rapid change presents a challenge for the technology industry in providing meaningful transparency that allows users to make informed judgments about the content they are seeing, without overwhelming or confusing users. As we learned when launching our AI transparency labels for content, labelling content with minor AI edits, which do not present meaningful incremental risk as compared to traditional photo editing or retouching, was not in service of our transparency goals and did not benefit our users. Without industry alignment on technical solutions for signaling material editing, however, there are incremental risks of potentially misleading edited content not receiving prominent labels under our current approach.

Regulation. At the same time, regulators around the globe are eager to act to counteract risks associated with generative AI, including "deepfakes," but often find themselves ahead of the technical progress necessary to support their proposals. For our organization and the broader industry, the key challenge is finding alignment on standardized methods of providing the appropriate GenAI signals to stakeholders, without overlabeling (providing too much irrelevant information) or oversimplifying (not providing enough). Our goal is to create a more refined, nuanced approach to labeling GenAI content that helps educate the public about the growing sophistication and widespread use of these tools. However, as users engage with this technology more frequently, their expectations around transparency are likely to evolve, and we must be responsive to those changing needs.

The ongoing need for collaboration. In developing and refining our approach to GenAI content transparency, we have collaborated with other players in the industry peers through forums including Partnership on AI, C2PA Steering Committee, the Munich Security Conference's AI Elections Accord, and regular bilaterals with our peers. We will continue to watch and learn, and we'll keep our approach under review as we do. We will keep collaborating with our industry peers, and we will remain in a colloquy with governments and civil society.

A call for research. Our user research and risk analysis has led us to believe we should prioritize visual and audio watermarking. We need continued research to improve robustness and accessibility. We believe promising approaches may include introducing multiple watermarks (Watermark anything with localized messages. Sander et al., 2024; Practical deep dispersed watermarking with synchronization and fusion. Guo et al., 2023) with varying degrees of robustness, which could allow for detecting tampering via the less durable watermark(s) and provide provenance information via the more durable watermark(s). Alternatively, with multiple-watermark approaches, it might be possible for Company A to share one watermark with other companies to detect GenAI content made or edited with Company A's on-platform GenAI tools, while reserving one or more for Company A's internal provenance use (e.g., reducing risk of removal, alteration, and counterfeiting).

Exploring new methods. As an organization, we are continuing to explore methods to improve watermarking robustness, public/private key watermarking to address tradeoffs between security and accessibility, and improved watermarking methods for GenAI video. Continued exploration will also inform research-backed definitions of materiality which are operationalizable at scale and reflect user and policymaker expectations for transparency. As GenAI applications, both standalone and as part of larger creative systems, continue to grow, we also anticipate that there may be other applications in which watermarking research may provide useful tools in new contexts, just as it has done over the last three decades.

Expanding uses of digital watermarking. In particular, we predict there will be a growing range of forensic use cases in which digital watermarking will provide needed evidence regarding AI-model protection (Sander et al. 2024), data protection (Bouaziz et al. 9 Oct 2024), and data attribution (Asnani et al., 2024). Given this expanding range of uses we predict for provenance generally–and for digital watermarks in particular–we look forward to continued knowledge-sharing as our collective research into watermarking techniques and efficacy continues to advance.

References

Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, Shruti Agarwal. Promark: proactive diffusion watermarking for causal attribution. IEEE/CVF Conference on Computer Vision and

Pattern Recognition 2025. URL https://cvpr.thecvf.com/media/cvpr-2024/Slides/29862.pdf. Submitted on 9 Oct 2024. Accessed on 2 April 2025.

Adam Block, Aush Sekhar, and Alexander Rakhlin. GaussMark: A Practical Approach for Structural Watermarking of Language Models. URL http://doi.org/10.48550/arXiv.2501.13941. Submitted on 17 Jan 2025. Accessed on 27 March 2025.

Wassim Bouaziz, El-Mahdi El-Mhamdi, Nicolas Usunier. Data taggants: dataset ownership verification via harmless targeted data poisoning. URL https://doi.org/10.48550/arXiv.2410.09101. Submitted on 9 Oct 2024. Accessed on 2 April 2025.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. URL https://doi.org/10.48550/arXiv.2303.15435. Last revised July 26, 2023 (version 2). Accessed on 27 March 2025.

Pierre Fernandez, Hady Elsahar, I. Zeki Yelniz, and Alexandre Mourachko. Video Seal: Open and Efficient Video Watermarking. URL https://doi.org/10.48550/arXiv.2412.09492. Submitted on 12 Dec 2024. Accessed on 27 March 2025.

IPTC. Photo Metadata. URL https://iptc.org/standards/photo-metadata/. Accessed on 2 April 2025.

Hengchang Guo, Qilong Zhang, Junwei Luo, Feng Guo, Wenbin Zhang, Xiaodong Su, and Minglei Li. Practical Deep Dispersed Watermarking with Synchronization and Fusion. In *MM'23: Proceedings of the 31st ACM International Conference on Multimedia,* pp. 7922–7932. URL https://dl.acm.org/doi/10.1145/3581783.3612015. ACM 27 Oct 2023. Accessed on 27 March 2025.

Sayashi Kapoor and Arvind Narayanan. We looked at 78 election deepfakes: political information is not an AI problem. Knight First Amendment Institute at Columbia University. URL https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem. 13 Dec 2024. Accessed on 3 April, 2025.

Gregory Kang Ruey Lau, Xinyan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Brian Kian Hsiang Low. Waterfall: Framework for Robust and Scalable Text Watermarking and Provenance for LLMs. URL https://doi.org/10.48550/arXiv.2407.0411 Accessed on 27 March 2025.

Partnership on AI. Building a glossary for synthetic media transparency methods. URL https://partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/ 23 Dec 2023. Accessed on 30 March 2025.

Partnership on AI. How Meta changed its approach to direct disclosure based on user feedback. November 2024. Accessed on 1 April 2025.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, & Hady Elsahar. Proactive Detection of Voice Cloning with Localized Watermarking. URL https://doi.org/10.48550/arXiv.2401.17264. Submitted on 30 Jan 2024. Accessed on 27 March 2025.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze & Teddy Furon. Watermarking makes language models radioactive. URL https://doi.org/10.48550/arXiv.2402.14904. Submitted on 22 Feb 2023 (v1), last revised 24 October 2024 (v2). Accessed on 2 April 2025.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, & Matthijs Douze. Watermark Anything With Localized Messages. URL https://doi.org/10.48550/arXiv.2411.07231. Submitted on Nov. 11, 2024. Accessed on 27 March 2025.

Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, & David G. Rand. Labeling AI-Generated Content: Promises, Perils, and Future Directions. URL https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf. MIT Schwarzman College of Computing. 28 Nov 2023. Accessed on 27 March 2025.

Yan, H. Y., Morrow, G., Yang, K.-C., & Wihbey, J. (2025). The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election. *Harvard Kennedy School (HKS) Misinformation Review,* 6(1). URL https://misinforeview.hks.harvard.edu/wp-content/uploads/2025/01/HKSMR_yan_origin_of-public_concerns_20240130.pdf. Received 15 August 2024. Accepted 13 Jan 2025. Published 30 Jan 2025.