PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How Meta changed its approach to direct disclosure based on user feedback

Meta

This is Meta's Case Submission as a Supporter of PAI's Synthetic Media Framework.
Learn more about the Framework

1 Organizational Background

1. Provide some background on your organization.

Meta builds technologies that help people connect, find communities, and grow businesses. Our mission is to give people the power to build community and bring the world closer together. More than three billion people around the world connect using our Family of Apps (Facebook, Instagram, Threads, WhatsApp, and Messenger) and Reality Labs.

AI has been a critical technology at Meta for over a decade, powering a number of important functions on our platforms, such as personalized ranking and recommendations for users, enabling accessibility tools like video subtitles, avatars and filters, and detecting violating content. With the development of our open-source Llama large language models, we've begun integrating generative AI into our Family of Apps in new ways. For example, we've integrated our latest models into Meta AI, which we believe is the world's leading AI assistant. It's now built with Llama 3 technology and available in more countries across our apps.

People are using Meta AI on Facebook, Instagram, WhatsApp, Messenger, and the web to get things done, learn, create, and connect with the things that matter to them.

2 Framing Direct Disclosure at your Organization

1. Are you writing this case as a Builder, Creator, and/or Distributor of synthetic media as defined in PAI's Synthetic Media Framework?

Meta acts as a Builder, Creator, and Distributor of synthetic media, as defined in Partnership on AI's (PAI) Framework, at different points. We build technology allowing users to create synthetic media through AI features like Imagine. We also create synthetic media directly through Imagine for You. Additionally, users and Creators distribute synthetic media made outside of Meta products across our Facebook, Instagram, Threads, Messenger, and WhatsApp.

2. Please elaborate on how your organization provides direct disclosure (as defined in our Glossary for Synthetic Media Transparency Methods) to users/audiences. What criteria does your organization use to determine whether content is disclosed? What practices do you follow to identify such content?

Meta uses a combination of direct and indirect disclosure methods across our apps. Our disclosure method differs depending on whether synthetic media was created using a Meta AI tool, or created by a third-party tool outside of Meta and shared on our apps. This divergent approach is a result of technical limitations on the information we have available to share about first party content vs. third-party content.

When photorealistic synthetic images are created using Meta AI, we take several actions to make sure people know that AI was used. This includes adding a visible burnt-in watermark on the images, and by adding labels to the image when they are shared or posted across Facebook, Instagram, and Threads, to let people know they were created using Meta AI. International Press Telecommunications Council (IPTC) metadata is also added to these images to allow for transparency when content is reshared outside of Meta's apps. Since not all online platforms currently label AI content based on metadata, applying a visible watermark provides additional transparency when AI content generated by Meta's products

is shared elsewhere on the internet and is maintained if the image is screenshotted.

For AI content created outside of Meta's platforms and shared to Facebook, Instagram, or Threads, we have built industry-leading tools that can identify invisible markers at scale via C2PA and IPTC technical standards, which allows us to label AI images if they incorporate these standards. Companies like Google, OpenAI, Microsoft, Adobe, Midjourney, and Shutterstock have all adopted these standards enabling us to label a significant amount of third party content.

We have also added a feature for people to disclose whether a piece of content was AI-generated when they share it on Facebook, Instagram, and Threads so we can add a label to it. We require people to use this disclosure tool when they post organic content with a photorealistic video or realistic-sounding audio that was digitally created or altered, and we may apply penalties if they fail to do so.

We also require that advertisers using third-party generative AI tools disclose AI use when they digitally create or alter a political or social issue ad. This disclosure is verified as part of our normal ads verification and review process.

---

3. Does your organization implement direct disclosure in a manner that, per the PAI Framework, strives to: "mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation?"

Meta's overarching goal of labeling AI content is to provide transparency around content that was created using AI, since many people are coming across this content for the first time on our platforms.

To achieve this, we have additional transparency and safety measures in place beyond the AI transparency labels and watermarks we discuss above. For example, we may apply a more prominent label for digitally created or altered content that creates a particularly high risk of materially deceiving the public on a matter of importance, so people have more information and context.

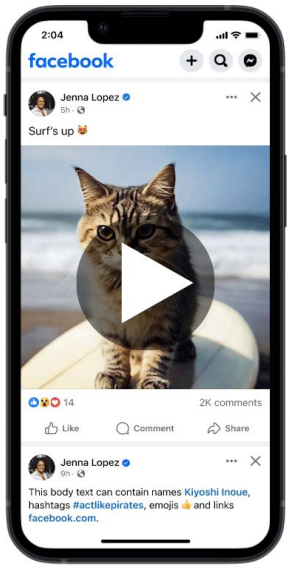
It's also important to note that all content on our apps is subject to third-party fact-checking, whether it's created with AI or not. AI-generated content is still eligible for rating by fact-checkers when it makes a claim that risks misleading people about something consequential that has no basis.

Image taken from "Our Approach to Labeling AI-Generated Content and Manipulated Media."

4. What, if anything, from your organization's approach to direct disclosure is missing from this NIST taxonomy below? Should it be added to a taxonomy of direct disclosure? If so, why?

From NIST's Reducing Risks Posed by Synthetic Content:

> The most commonly used techniques to *directly disclose* to the audience how AI was used in the content creation process include:
>
> - content labels (e.g., visual tags within content, warning labels, pre-roll or interstitial labels in video and/or audio, and typographical signals in text highlighting generated AI text with different fonts),
> - visible watermarks (e.g., icons covering content indicating AI usage where the bigger the icon, the harder its removal), and
> - disclosure fields (e.g., disclaimers and warning statements to indicate the role of AI in developing the content, and acknowledgments to provide more context to the AI contribution and credits to reviewers).



Meta uses direct disclosure labels as an entrypoint to user education. When people click on the label, they'll see a bottom sheet providing more context about why and how Meta is applying these labels, with more information about how AI is used to create or edit images.

Video taken from "Our Approach to Labeling AI-Generated Content and Manipulated Media." *Click to play. Requires Adobe Acrobat.*

5. Per the Framework, PAI recommends disclosing "visual, auditory, or multimodal content that has been generated or modified (commonly via artificial intelligence). Such outputs are often highly realistic, would not be identifiable as synthetic to the average person, and may simulate artifacts, persons, or events." How does your organization's approach align with, or diverge from, this recommendation?

Our direct disclosure approach differs depending on whether the content was created with a Meta AI tool ("first party") or created using a non-Meta AI tool and then posted to Facebook, Instagram, or Threads ("third party"). This divergent approach is a result of technical limitations on the information we have available to share about the creation of first party content vs. third-party content.

When photorealistic images are created using our Meta AI feature, we do several things to make sure people know AI is involved, including putting visible burnt-in watermarks on the images, and IPTC metadata embedded within image files. In addition, we place labels on content letting users know which Meta AI tool, such as Imagine, was used to create it.

For AI content created by non-Meta tools and posted to Meta platforms, our AI labels are applied in one of two scenarios:

1. When we detect industry-standard metadata signals on images indicating that AI was used in the creation of the content.
2. Or, when users self-disclose the use of AI on content they upload to Facebook, Instagram, or Threads.

This approach to direct disclosure is broader than PAI's recommendation because we don't require distributed content to be photorealistic for it to receive a label. For distributed content, transparency labels based on AI metadata received from the third-party tool used to create it are applied regardless of whether or not the content is photorealistic.

3 Real World, Complex Direct Disclosure Example

1. Provide an example in which your organization applied (or did not apply) a direct disclosure to a piece, or category, of content for which it was challenging to evaluate whether it warranted a disclosure (based on your organization's policy). This could be because the threshold for disclosing was uncertain, the impact of such content was debatable, understanding of how it was manipulated was unclear, etc. Be sure to explain why it was challenging.

In May 2024, Meta began rolling out direct disclosure labels for AI content distributed on Facebook, Instagram, and Threads. We began applying these labels to images when we detected industry-standard AI image indicators or metadata, such as from the C2PA or the IPTC, from companies such as Google, OpenAI, Microsoft, Adobe, Midjourney, and Shutterstock. When our systems detected metadata indicating that content was created or edited using AI, a label was automatically applied.

At the same time, we launched the ability for users to self-disclose that they were uploading AI content and add a label to their posts, stories, and Reels.

Soon after launch, creators and users expressed surprise to see labels indicating their content was created with AI applied to their images via metadata. Creators were confused about our approach, as they did not believe that they used AI to produce the content in question.

In response to this feedback, we began investigating the behavior of the labels, and found metadata indicating AI usage present on all the images we looked at.

Further investigation and conversations with creators revealed a common theme — the use of image editing tools such as Adobe Photoshop. These images had C2PA "edited" metadata attached, as many AI assisted editing functions are now integrated into image editing tools. Like others across the industry, we found that our labels based on these indicators weren't always aligned with people's expectations and didn't always provide enough context. For example, some images that included minor modifications using AI, such as the use of retouching tools, included industry standard metadata and were then labeled as "Made with AI."

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 significant enough to warrant a label indicating that the content was made with AI. For example, photos where an AI editing tool had been used for color corrections, removing distracting details in an image's background, or adding detail to clothing were labeled as "Made with AI" based on the metadata we received. In other cases, creators were not conscious of any AI editing, but the image editing tools they used were powered with AI under the hood.

Several factors made this a particularly challenging problem. Photo editing is very common

online, and photo editing that is not powered by generative AI has historically not been broadly labeled. Questions were thus raised about the marginal risk introduced by these AI-edited images — many of which, as stated above, contained immaterial editing.

In addition, the current state of the industry on AI metadata means that the signals Meta received from tools used to produce images did not differentiate between minor edits, like color correction, and more significant changes which could substantially mislead people — like using AI editing tools to portray an event which didn't happen, or a person doing something they didn't do.

As a result of our internal investigations and conversations with users and Creators, we began discussing solutions to improve the labeling process. Our goal was to find a way to apply labels that better matched our intent of labeling content created with AI, and that created less confusion for users and Creators.

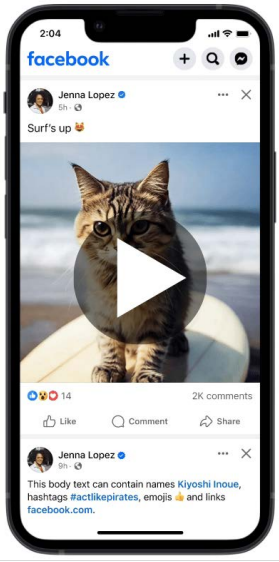
To address these concerns, we ultimately decided to implement two changes to the product.

1. In the short term, we would update the label to be more clearly inclusive of both AI-generated and -edited content ("AI info"). The intention behind this short-term change was for our labels to read more neutrally while we worked to improve transparency holistically.

2. In the longer term, we aligned on a different approach to labeling for content that was edited with, but not fully created by, AI. This label, accessed through our three-dot contextual menus, would continue to provide transparency for those seeking more information about AI content. Labels for content wholly created by AI would remain on the surface of the content.

Making this change meant we would continue to label content based on metadata received from Adobe, Google, OpenAI and other creation tools, and would begin to differentiate between the industry-standard metadata types for AI-edited versus AI-created content.

This decision was informed by feedback indicating that people expect greater transparency for wholly created synthetic content, and an analysis of AI-edited content receiving the AI transparency label. We found that a majority of the AI-edited content posted to Facebook, Instagram, and Threads contained edits that would be considered cosmetic or artistic — in other words, edits that did not meaningfully change the context of the content. This included things like color correction, de-noising, and other aesthetic editing commonly done by photographers.



Video taken from "Our Approach to Labeling AI-Generated Content and Manipulated Media." *Click to play. Requires Adobe Acrobat.*

2. Where possible, what were some of the rejected solutions for directly disclosing this content? Please provide details on your organization's reasoning for rejecting those solutions.

When discussing potential solutions, our priorities were to balance the need to continue providing transparency around AI-created or -edited content, while reducing confusion for users and creators. To inform the options we considered, Meta analyzed platform data on the amounts and types of AI-generated and -edited content distributed on our platforms, and considered the current state of metadata-based AI content labeling on other online platforms.

Other potential approaches to this problem could have included entirely ceasing metadata-based labeling of content that was edited with, but not wholly created by, AI. However, content does not have to be totally AI-generated to be what people view as AI. For instance, AI editing to swap a background or add other people to photos could meaningfully change the context of an image. In the context of global elections in 2024, we did not want to take such a drastic step away from direct disclosure, even if a large portion of the AI-edited content we were seeing posted to our platforms was benign.

Another potential approach to this problem could be two different direct disclosure labels — for example, one label for AI-generated content and a different label for edited content. However, such an approach is more complex and would place an additional mental burden on users.

Though some creators and users requested the ability to remove the label on request or be added to an allow-list of creators who would not be labeled, we rejected those approaches. As our labels are intended to be neutral signals to viewers and allowing individual users or creators to opt-out was not aligned with our transparency principles. Additionally, we would have no way to verify at scale that creators on such an "allow list" were not using AI in their content.

3. How was this piece/kind of content identified?

This issue was identified through feedback from users and **Creators** on Facebook and Instagram.

4. Was there any potential for reputational (e.g., negative impact on your organization's brand, products, etc.), societal (e.g., negative impact on the economy, etc.), or any other kind of harm from such content?

Our intention in labeling AI-generated content is to provide users with neutral signals around how AI is used to create or edit content.

However, many **Creators** in creative industries, such as film and television, video games, and fashion, did not perceive AI labels neutrally. In certain artistic and creative communities, generative AI tools are viewed negatively—as a shortcut, or even as a risk to the livelihoods of artists, writers, and designers. Certain film industry labor unions have restrictions on the use of AI, while some members of the creative community stigmatize the technology, believing it is harmful to artists.

In addition, we had a broad concern that if labels were applied in cases where users would not reasonably consider the content in question to be generated by AI, the result could be label fatigue, or a loss of trust in the label.

This disconnect in the labels' intent and how some users received them highlights the need for industry-wide alignment on labeling standards, which provide transparency that users find valuable while avoiding stigma, confusion, or fatigue due to over-labeling.

5. What was the impact of implementing this disclosure? How did you assess such impact (studying users, via the press, civil society, community reactions, etc.)? Did the disclosure mechanism mitigate the harm described in the previous question (3.4)?

Prior to launching these labels, we conducted research on the expectations of transparency around AI content and considered external research on the subject. We also consulted with academics, civil society organizations, and other experts, as we typically do, to inform new products and policies. We'll continue to monitor the overall impact on user sentiment and experience to understand how users respond to this type of direct disclosure.

6. Is there anything your organization believes either the Builder, Creator, or Distributor of the content (aka others in the content pipeline) should have done differently to support your implementation of direct disclosure?

Overall, this case emphasizes the need for increased industry-wide alignment in several areas. First, the industry should work together to align on definitions differentiating between AI-created and AI-edited content, and principles on how this nuance can be embedded in industry-standard metadata signals. Alignment is also necessary when it comes to how and if direct disclosure approaches should differ for those two content categories. Differentiation between cosmetic or artistic and more contextually meaningful edits should be a conversation at the industry level—no one technology company should implement its own unilateral definition that may not align with the definition of other players.

7. In retrospect, would your organization have done anything differently? Why or why not?

We expect that we will continue to review and evolve our approach over time as the cutting edge of what's technically possible develops. Throughout that ongoing process, this experience highlights the importance of extensive user testing, beta feedback, etc., prior to the large-scale rollout of novel direct disclosure mechanisms.

8. Were there any other policy instruments your organization relied on in deciding whether to, and how, to disclose this content? What external policy may have been helpful to supplement your internal policies?

It would be helpful to have broad external alignment among producers of AI media on differing methods for disclosure depending on degree of creation or modification of content. This case also highlights the need for the industry to coalesce around standards for what constitutes "material" vs. "immaterial" edits. Much of the challenge that Meta faced in developing a response to these issues stemmed from what we can and cannot tell about AI use to create or edit content at scale.

Based on the current state of industry metadata, a photo edited with an AI tool to brighten colors and reduce blur would, via metadata, appear to be exactly the same as a photo edited using AI to depict a real person in a wholly made-up scenario.

9. What might other industry practitioners or policymakers learn from this example? How might this case inform best practices for direct disclosure across those Building, Creating, and/or Distributing synthetic media?

This case highlights the importance of alignment on differentiation of created vs. edited synthetic content. When considering implementation of direct disclosure, other distributors of synthetic media should consider the relative prevalence of AI-created vs. -edited content on their platforms. Other technology companies should also note the importance of extensive testing and feedback prior to the rollout of direct disclosure mechanisms.

Distributors and platforms should also consider the attitudes of their user communities towards synthetic content and how that may affect their reactions to direct disclosure. As we've seen, creative communities don't necessarily have a neutral view of synthetic content—to name just one example. Thus, a label that was intended neutrally wasn't always perceived that way.

Finally, this lesson can and should extend to regulation as policymakers around the world grapple with how best to govern provenance-related issues. Policies should reflect, and be sufficiently flexible to address, the technical challenges involved with differentiating between wholly generated and edited content. To the extent relevant, policymakers should also provide clear guidance on what constitutes "material" vs. "immaterial" edits. By working with practitioners and sandboxing solutions, policymakers can provide valuable direction for assessing the relevant technical, social, and commercial trade-offs.

Distributors of synthetic media should consider the relative prevalence of AI-created vs. -edited content on their platforms.

4 How Organizations Understand Direct Disclosure

1. What research and/or analysis has contributed to your organization's understanding of direct disclosure (both internal and external)?

Prior to launching our AI transparency labeling program, Meta's team completed consultations with over 120 stakeholders in 34 countries in every major region of the world.

Throughout these conversations, a majority of stakeholders agreed that content removal should be limited to only the highest risk scenarios where content can be tied to harm, since generative AI is becoming a mainstream tool for creative expression. This aligns with the principles behind our Community Standards — that people should be free to express themselves while also remaining safe with our services.

As part of this policy stakeholder engagement work, we heard a wide spectrum of expert feedback on whether and when Meta should label AI-generated content. We asked for feedback on the principles of when Meta should be applying a label, outlined options on the size and placement of labels, and who should be responsible for labeling (e.g., Meta applying a label vs. asking users to self-disclose).

Overall, we heard broad support for labels and strong support for a more prominent label in high-risk scenarios. Many stakeholders were also receptive to the concept of users voluntarily self-disclosing content as AI-generated.

Some social media researchers have argued self-disclosure is an opportunity to encourage users to take responsibility for their content, could prevent content from being taken out of context and going viral, and would empower users to control their experience on the platform. Other experts argued that voluntary self disclosure unfairly placed the onus on users to identify AI-generated content.

We also conducted public opinion research and asked people how social media companies, such as Meta, should approach AI-generated content on their platforms. A large majority of survey respondents favored warning labels for AI-generated content that depicts people saying things they did not say.

Aligned with this finding, users have told us that they have higher expectations for transparency around certain types of content — such as content depicting human subjects or current events. Other types of content have less expectations for transparency, such as content with minor alterations by AI.

When it comes to direct disclosure approaches, many consumers of synthetic content have told us they prefer obvious synthetic content labeling, e.g., those that contain both text and an icon. Many participants explained they didn't have the context to understand icon-only watermark transparency approaches.

We'll continue to conduct user research and expert engagement to understand how attitudes and preferences toward this content develop over time. In addition, we'll continue to monitor the development of the regulatory landscape in this space and evolve our product approach accordingly.

2. Does your organization believe there are any risks associated with either OVER or UNDER disclosing synthetic media to audiences? How does your organization navigate these tensions?

There are potential risks associated with over-disclosing synthetic media. For example, users could become blind to or distrust disclosure labels if they are over-applied to content that people don't consider to be made with AI. This is related to the risk of **Creator** and user confusion when labels are applied to content that most people would not consider to be made with AI.

A primary risk of underdisclosure of edited synthetic content, which we considered in our response to this case, could be that AI-edited content has the potential to be misleading, even if it is not wholly created by AI. This highlights the need for the industry as a whole to align on a standard for material vs. immaterial content edits.

3. What conditions or evidence would prompt your organization to re-calibrate your answer to the previous question (4.2)? E.g., in an election year with high stakes events, your organization may be more comfortable *over labeling*.

As 2024 saw a historic number of elections around the world, we deemed it important to prioritize greater transparency, which made us more comfortable with potentially over-labeling. We also considered the novelty of generative AI technology and the fact that many users are coming across this type of content for the first time on Meta's apps. In this context, user expectations around transparency are heightened.

As mentioned in the previous answer, our internal research indicates that users have higher expectations for transparency around photorealistic content, content depicting human subjects, or content depicting public figures or current events — all of which are types of content that are even more important to provide transparency on in the context of significant elections.

4. In the March 2024 guidance from the PAI Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Creative uses of synthetic media should be labeled, because they might unintentionally cause harm; however, labeling approaches for creative content should be different, and even more mindfully pursued, than those for purely information-rich content."

   Does your organization agree? If so, how do you think creative content should be labeled? What is your organization's understanding of "mindfully pursued"? If your organization does not agree, why not?

As part of our internal user research, we found that users had reduced expectations of transparency for AI content that was designed to entertain or amuse, as opposed to content intended to inform or to impact behavior.

In the short term, our labeling strategy for distributed content does not differentiate between creative uses of synthetic media as opposed to any other applications. This is currently an industry-wide technical limitation — we're not able at scale to differentiate between photorealistic and non-photorealistic distributed AI content. While there are separate "edited" and "created" C2PA metadata fields, there's no such distinction between "photorealistic" and "non-photorealistic."

However, we do distinguish between photorealistic and non-photorealistic content in our requirements for users to self-disclose the use of AI in content they share on our platforms. Artistic or stylized AI-generated content shared by users is not required to be labeled if it's not photorealistic.

As generative AI technology is novel and many users are encountering AI-generated media for the first time on Meta's apps, we consider it important to over-index on transparency. We regularly consult with academics, civil society organizations, and other experts to inform our products and policies, to help ensure we get this balance right. When we roll out new product changes, we survey people to understand how these changes impact peoples' experiences.

As we mindfully pursue the right balance for transparency on creative content, we'll continue to pay attention to how this technology develops and work with our industry partners to evolve our approach. We expect that we will continue to review and evolve our approach over time as the cutting edge of what's technically possible develops.

5. Overall, what role(s) does your organization believe Builders, Creators, and Distributors play in directly disclosing AI-generated or AI-edited media to users?

Overall, we are aligned with PAI's Responsible Practices for Synthetic Media and adhere to those best practices in our roles as **Builders**, **Creators**, and **Distributors** of synthetic media. We also believe that users can play a role in creating norms around transparency around the use of AI-generated or AI-edited media through self-disclosing AI content they share online.

6. How important is it for those Building, Creating, and/or Distributing synthetic media to all align collectively, or within stakeholder categories, on a singular threshold for:

   1) the types of media that warrant direct disclosure, and/or
   2) more specifically, a shared visual language or mechanism for such disclosure?

   Elaborate on which values or principles should inform such alignment, if applicable.

We believe this is a critical step for the industry to take. As we've indicated elsewhere in this case study, a shared set of standards could reduce the informational overhead on users and make it easier for users to make informed judgements about the content they're seeing.

A shared set of standards could reduce the informational overhead on users and make it easier for users to make informed judgements about the content they're seeing.

5 Approaches to Direct Disclosure, in Policy and Practice

1. What does your organization believe are the most significant socio-technical challenges to successfully achieving the purpose of directly disclosing content at scale? (Refer to question 2.3 for reference to PAI's description of direct disclosure)

A significant challenge is the speed of technical innovation and advancement of capabilities. AI development is moving fast. Soon, AI will likely be meaningfully embedded in much of, if not most of, content and media people see online. In this context, the technology industry faces a challenge in providing meaningful transparency that helps people make judgments about the content they're seeing, in a way that is helpful and not over-burdensome to users or **Creators**.

At the same time, regulators are eager to act in this space, and are often getting ahead of the technical progress and alignment that would be needed to accomplish their proposals.

2. What is your organization hoping to accomplish by implementing direct disclosure? (Note: you may have already discussed this in detail in Section 2, so feel free to call upon that.) Does your organization believe directly disclosing ALL AI-edited or generated media, is useful in helping accomplish those goals?

We believe it's important to provide transparency when we know that media was created with or edited by AI, whether by Meta's AI tools or outside of our platforms. As we discussed in Section 2, we've found through internal user research that users expect direct disclosure from platforms, especially since generative AI technologies are nascent and many people are coming across this content for the first time.

As we continue to learn more about the development of AI capabilities, including how people are sharing and engaging with AI content, and what kind of transparency approaches people find useful, we expect to continue to evolve our approach. In doing so, we'll monitor the ways in which adversarial actors seek to deceive people with AI content and keep looking for ways to stay one step ahead.

3. Please share your organization's insight into how direct disclosure can impact:

    1) Accuracy
    2) Trustworthiness
    3) Authenticity
    4) Harm mitigation
    5) Informed decision-making

    Note: You can also discuss your understanding of the relationship between these concepts (for example, authenticity could impact trustworthiness, harm mitigation, etc.)

1. Accuracy

Our AI transparency labels are not fact-checking labels, and we've taken care not to position them as such. Meta maintains a separate policy and product approach for fact-checked misinformation in partnership with our trusted network of third-party fact checkers. All content on our platforms, whether created with AI or not, is eligible to be reviewed and rated by fact-checkers.

2. Trustworthiness

That being said, our internal research has affirmed that users do expect tech companies to provide transparency around AI content when possible — particularly around realistic content, and content depicting human subjects, public figures, or current events. Many researchers and academics in the field are also concerned about the "liar's dividend", wherein people will begin to broadly "doubt the authenticity of content because of the plausibility that it's AI-generated or AI-modified." Providing direct disclosure of confirmed AI-generated or -edited content could help to combat this.

3. Authenticity

We view our current approach to direct disclosure as a first step in a much larger ongoing project related to communicating the authenticity of content. As this case illustrates, AI

generation of content is more of a spectrum than a binary. As smartphone makers like Apple and Google integrate AI photo cleanup and editing features directly into their phone cameras, these lines will continue to muddle. A longer-term vision is to provide far more information about the full process of creating and editing of a given piece of content.

4. Harm mitigation

When it comes to AI-generated or -edited content that could be misleading — photorealistic content, content depicting public figures, etc. — direct disclosure labels can mitigate harm posed by the spread of digital misinformation. In this use case, AI direct disclosure operates in tandem with Meta's existing third-party fact-checking labels, which apply to all content whether created with AI or not.

5. Informed decision-making

All of these concepts can be applied together to help online platform users make informed decisions about content they encounter online.

4. Does your organization believe there will be a tipping point to the liar's dividend (that people doubt the authenticity of real content because of the plausibility that it's AI-generated or AI-modified)? Why or why not? If yes, have we already reached it? How might we know if we have reached it?

Previously, we've cited the challenges posed by the high velocity of innovation in the generative AI space. It's likely too early to tell how content created using AI will impact people's perceptions of the online content ecosystem as a whole. However, this challenge highlights the need for cross-industry alignment on direct disclosure of AI content and content authenticity more broadly. A patchwork of direct disclosure approaches by tech companies, where users currently don't have strong mental models for what indicates authenticity, will make it easier for bad actors to mislead.

There's a similar risk that a patchwork of laws in this area — where different direct disclosure requirements are adopted in various countries or U.S. states — will add to such confusion.

5. As AI-generated media becomes more ubiquitous, what are some of the other important questions audiences should be asking in addition to "is this content AI-generated or AI-modified," especially as more and more content today has some AI-modification?

Our learnings from this case highlight the importance of not only alignment on whether content has been modified by AI, but to what extent. The large majority of the content modified by AI that we reviewed was not likely to mislead people or to create any societal risk — most edits were cosmetic or artistic as opposed to edits that significantly changed the context of the content.

It will also be important for the industry to align on the marginal informational risks posed by AI-edited content and approaches to disclosure. Photo editing not powered by generative AI has the potential to mislead people, but historically has not been broadly labeled online.

6. How can research help inform development of direct disclosure that supports user/audience needs? Please list out key open areas of research related to direct disclosure that, the answers to which, would support your organization's policy and practice development for direct disclosure.

Research will be needed to inform user expectations of transparency in a world where AI is likely to be part of most online content. For example, how would users expect **Builders** or **Distributors** to communicate the use of AI in a ten-minute video that contains some synthetic content, while most of the video is not AI-generated?

It will be beneficial to the industry as a whole to develop consistent direct disclosure approaches to such challenges, grounded in user research and expert feedback.

6 Media Literacy and Education

1. In the March 2024 guidance from the Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Broader public education on synthetic media is required for any of the artifact-level interventions, like labels, to be effective."

   Does your organization agree? If so, why? Has your organization been working on "broader public education on synthetic media"? How? (please provide examples.) If your organization does not agree, why not? What responsibility do organizations like yours (identified in the Framework as either a Builder, Creator, or Distributor) have in educating users? What about civil society organizations?

Users have told us that they desire more information about generative AI content. That has informed our full scope of transparency measures, including labels, visible watermarks, and educational bottom sheets.

In addition to in-product transparency on AI content, through our Privacy Center, system cards and model cards, we're taking steps to explain how our AI systems work in a way that experts and non-experts alike can understand.

It's important to note that any user education — whether by technology companies, civil society organizations, or educational institutions — will be less effective if people have to familiarize themselves with a different transparency system on every social media platform or website they use.

Any user education will be less effective if people have to familiarize themselves with a different transparency system on every social media platform or website they use.

2. What would you like to see from other institutions related to improving public understanding of synthetic media? Which stakeholder groups have the largest role to play in educating the public (e.g., civic institutions, technology platforms, schools)? Why?	Distributors or hosts of synthetic content will have an important role to play as the venues where most people will be encountering synthetic media. User education and investment in media literacy initiatives will be key here — both directly on apps where people encounter this content, as well as in partnership with organizations like Partnership on AI and other industry groups.
3. What support does your organization need in order to advance synthetic media literacy and public education on evaluating media?	Alignment on direct disclosure approaches across the industry would be an important component of public education. As we discussed earlier in this case study, if users need to familiarize themselves with a disparate labeling and disclosure approach for every different online space, disclosure will be less effective. A well-aligned cross-industry transparency approach would require less cognitive overhead from users and lead to less confusion. Additionally, regulation grounded in the cutting edge of what's technically possible will benefit the industry, as many regulators are currently eager to intervene in this area, but what's technically possible in some cases lags behind their policy ambitions.