PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA

CASE STUDY

How Google's research informed its approach to direct disclosure

Google

This is Google'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.

Google's mission is to organize the world's information and make it universally accessible and useful. People around the world use our family of products, which include Search, YouTube, and Google Ads. AI has been central to our products for many years, powering Google's ability to provide the most relevant and helpful information to our end users. In recent months, we've watched people bring their ideas to life with help from our generative AI models: YouTube creators are exploring the creative possibilities of video backgrounds for their YouTube Shorts, and many use our Gemini App or Google Labs tools to create text and media content for a broad range of creative, personal, and business goals. As we continue to bring AI to more products and services to help fuel creativity and productivity, we are focused on helping people better understand how a particular piece of content was created and modified over time.

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?

Per PAI's Synthetic Media Framework, our services place us as **Builders of Technology and Infrastructure** and **Distributors of synthetic media**; however, we will focus on the distribution aspect for the purpose of this case study — and specifically on direct disclosures in the context of Google Search, Google Ads, and YouTube.

2. Please elaborate on how your organization provides direct disclosure (as defined in our Glossary for Synthetic Media Transparency Methods) to users/audiences.

As of December 2024, Google uses direct disclosures in the following ways (please note that changes may have been made since December 2024):

- YouTube requires creators to disclose when they've uploaded meaningfully altered or synthetically generated content that seems realistic. To help creators understand what is meant by content that "seems realistic," YouTube provides examples for when creators are required to disclose (e.g., showing a realistic depiction of a missile fired toward a real city) and when they are not required to disclose (e.g., using an AI-generated animation of a missile in a video) (see YouTube's product policy page for details). Based on this disclosure, YouTube adds a label for disclosed content in the video's description. For sensitive content, YouTube also displays a more prominent label on the video player for added transparency.
- Google Ads requires certified election advertisers to prominently disclose when their ads include realistic synthetic content that's been digitally altered or generated to "inauthentically depict real or realistic-looking people or events." We provide disclosures for that information; in formats where Google labeling is not available, advertisers are required to add a label to the asset. (Click here for more details on included and non-included formats.)
- The "About this image" feature in Google Search helps people assess the credibility and context of images they see on the open web, including by providing information

as to whether an image is generated by Google's own AI systems based on its SynthID watermark, and how the image was made or edited based on standards from the International Press Telecommunications Council (IPTC) and the presence of Coalition for Content Provenance and Authenticity (C2PA) manifest data (example).

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?"

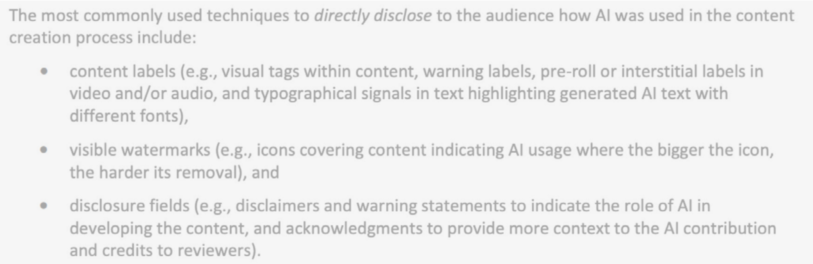
At Google, we see information about whether a piece of content is edited or generated by AI as one of multiple potentially important elements (e.g., content's source, date of publication, corroboration of claim) that can help users determine whether they want to trust what they see online.

In other words, knowing if content was made with generative AI answers the question "How was this made?" but does not necessarily answer the question "Is this trustworthy?" Not all AI-generated content is deceptive, and not all deceptive content is AI-generated. In fact, we expect that over time a large amount of high-value content will be AI-generated or augmented, and we want to make sure users can easily discover and benefit from that content as needed.

Accordingly, we strive to use disclosures as part of a broader set of contextual features, ranking heuristics, and policies that are geared toward helping users make informed decisions about what content to trust. Sometimes, prominent disclosures about the method of content generation may be particularly helpful to users; other times, they may be unhelpful or confusing. (We will elaborate on this in the case study section.) We calibrate our approach to optimize for helpful use cases and limit the risks of unhelpful ones.

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?

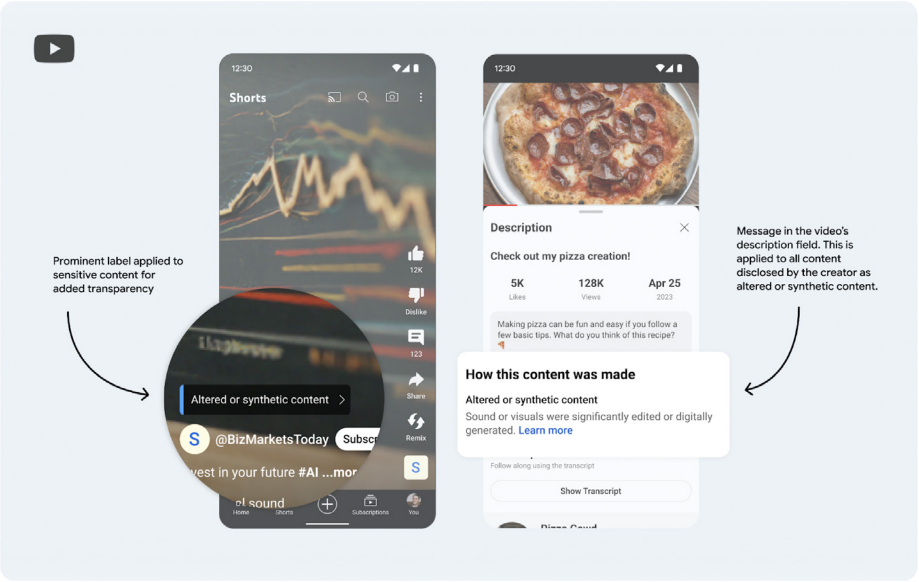
From NIST's Reducing Risks Posed by Synthetic Content:



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).

Given the risks of direct disclosures that solely indicate AI use (which we will outline in the case study), Google recommends rigorous analysis of the downsides and upsides of user-facing disclosures in each individual use case where they are considered. In most cases, we use a generic header on a landing page such as "How this content was made" or "About this image"; context is provided about how the content was made, along with other information known to be useful in determining the trustworthiness of content such as information about the source. However, in such cases where risks of deception are high and the harms of such deception would be significant, we may deploy prominent user-facing disclosures indicating that content is altered or synthetic. (See below.)



Prominent label applied to sensitive content for added transparency

Check out my pizza creation!

Making pizza can be fun and easy if you follow a few basic tips. What do you think of this recipe?

How this content was made

Altered or synthetic content

Sound or visuals were significantly edited or digitally generated. Learn more

Follow along using the transcript

Message in the video's description field. This is applied to all content disclosed by the creator as altered or synthetic content.

5. What criteria does your organization use to determine whether content is disclosed? What practices do you follow to identify such content?

Google centers its decisions about direct disclosures on user need to determine the trustworthiness of content. Our research suggests that, in some cases, directly disclosing that content was made with generative AI can confuse or mislead an end user and undermine their ability to determine the trustworthiness of content. (We will elaborate on this research in the case study section.) Because of these risks, we currently employ the use of disclosures sparingly. We employ prominent direct disclosures (i.e., a disclosure that is placed directly on the surface of content) only in cases with sensitive content in which there is high likelihood of harm from deception (e.g., an election ad). We employ low-prominence direct disclosures (e.g., a disclosure accessible through a semantically neutral entry point that indicates "learn more") more frequently. The use and design of direct disclosures indicating that content is made with AI vary, based on:

- **Sensitivity:** Google is more likely to provide prominent disclosures on topics wherein the harms of deception are likely to be highest, such as realistic civics or elections content. A prominent disclosure is one that is placed directly on top of content.
- **Reliability of the signal:** Google is more likely to provide prominent disclosures if the signal is very reliable and gives us high confidence in the accuracy of the information relayed to users than if the signal is unreliable and does not give us high confidence.
- **Product experience:** Google's products vary in terms of the specific user needs they are meeting and therefore vary in the design system and user experience. Given this variability, products design their direct disclosure solutions based on the design language of the product itself, weaving direct disclosures into the fabric of the product experience. Because of this customization, the specific design of direct disclosures may vary across products like YouTube and Search.

See above (Question 2) for a sense of how our products currently implement disclosures in light of these considerations.

Our approach may change over time as use of generative AI becomes more prevalent and user understanding of generative AI evolves; any user-facing disclosure strategy should dynamically adapt to changes in technology and in people's expectations.

6. 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?

Google's approach to direct disclosures anchors on the potential harm from deception that may come from synthetic content. Therefore, realism is among a set of factors that we consider when determining the use of direct disclosures. We will expand on this in the case study.

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.

As the use of generative AI to create content has increased, users, press, and policymakers are becoming more concerned about the potential for deception arising from generative AI content. Google's services have long invested in efforts to empower users with the kind of context that helps them make their own informed trustworthiness decisions about online content, and in recent years we started exploring whether there was more we could do when it comes to generative AI content specifically.

Our first step was to confirm the primary user problem we were aiming to solve: helping users determine for themselves the trustworthiness of content they encounter online.

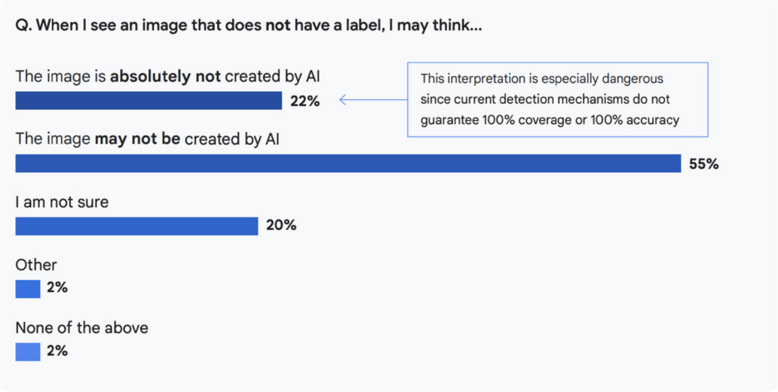
How that content is made constitutes one important factor to that end, (though not necessarily determinative); accordingly, we engaged in a deeper exploration of the ways we could help users understand whether and how generative AI was used to create content.

With these goals in mind, we initiated research projects involving candidate user-facing labels (i.e., "Created by AI," "Not created by AI," "Made with AI," "Edited with AI," and "Altered or synthetic content"), and tested them with research participants:

1. Which label was most comprehensible to end-users
2. The impact the label had on user perception of the content
3. The impact the label had on user perception of unlabeled content

This research demonstrated that people can struggle to understand the concrete implications of any variant of the labels we tested, and that a lot of nuance and precision was lost in the process. For example, not everyone has the same understanding of what it means for content to be "altered or synthetic." Similarly, users were not clear how AI was used to generate or alter the content. In both cases, a range of provenance signals may indicate that an AI tool was used to edit or create the content, but they don't help the user interpret that information. A label short enough to be superimposed on top of content was found to be too short to accurately communicate this nuance.

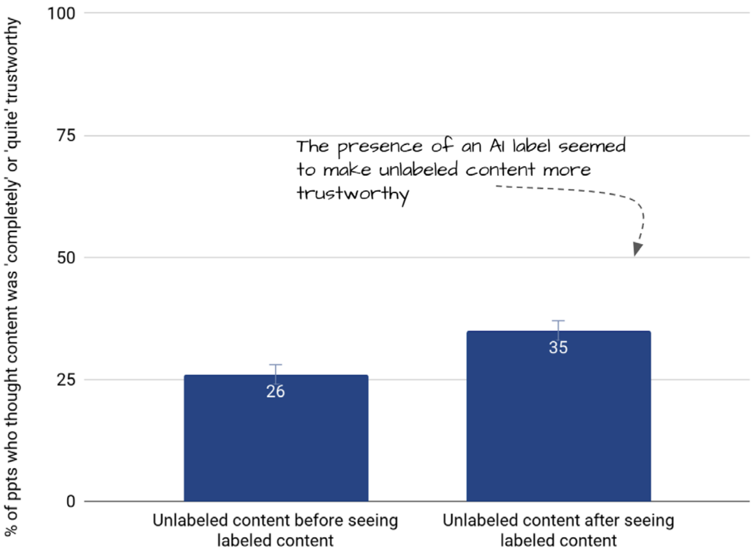
In addition, this research demonstrated the "implied authenticity effect" — the phenomenon that when some content is labeled as AI-generated or edited but other content is not — a significant slice of participants (more than 20 percent in our research, with a sample of 2,700) assume that unlabeled content must be authentic. This is a dangerous interpretation, because no detection mechanism is 100 percent accurate or 100 percent comprehensive; therefore, all that it means for content to be unlabeled is that it is not known how the content was made.



Google UXR Study 2024, U.S. sample, n=2,700

Finally, the research replicated the implied truth effect found in fact-check labeling research, with participants being more likely to identify unlabeled content as trustworthy after seeing other content with a generative AI label. This seems to suggest that these participants assume that content denoted as being AI-generated is inherently untrustworthy, and/or that content that is captured, rather than generated, is inherently more trustworthy. This is inaccurate and risks leading people to flawed conclusions about online content. For example, an unedited photo captured by a camera can be rendered deceptive by using simple techniques such as cropping or metadata editing, or by presenting it in a way that misleads about its context. By contrast, AI-generated content can accurately and helpfully represent various complex phenomena (e.g., in physics or biology), or it can be made to be creative and artistic rather than deceptive. Deceptive content predates generative AI:

If labels can mislead users into thinking they should trust any content that is not created with a generative AI tool, they represent a significant risk for users' ability to make informed decisions.



Google UXR Study 2024, U.S. sample, n=8,000, where ppts means participants

With these insights, Google services including Google Ads, Search, and YouTube began to develop solutions to help people better understand how the content they see was captured or generated, leveraging the mechanisms available to, and best suited for, each of these different products. Those typically include a mix of signals:

- Some products, like YouTube, can engage an active community of creators who are eager to share more information about how they created their content.
- In addition, a lot of online content comes with metadata, information about a piece of content encoded in open and interoperable formats. Such standards as IPTC-enabled tools allow creators to attach information about how content was made in ways that can be read by downstream viewers or distributors. In some cases, as with the C2PA standard, that metadata can be cryptographically protected, meaning that any tampering would be evident.
- Finally, watermarking technology such as Google DeepMind's SynthID can enable companies like Google to invisibly mark up content that their generative AI tools generate, or edit to allow users to be able to recognize the content as such once distributed.

Self-disclosure allows Ads and YouTube to ask more about the content itself and how a generative AI tool was used. It offers the content creator the opportunity to be part of the labeling process. Due to this additional contextual information, and in light of the specific risk profiles of these services, we determined that the benefits at this time of placing a prominent label on sensitive content such as election ads outweighed the downsides outlined above. Importantly, this strategy must be allowed to evolve as use of generative AI tools changes. For instance, if in the future all content is touched by generative AI tools, it may become meaningless to prominently label content as such.

Search must rely on data sources such as SynthID, IPTC metadata, and C2PA manifest data, given that it is not a hosting platform. While each has varying degrees of comprehensiveness and accuracy and all provide some useful information about how content is made, none has the granularity of context that comes with creator-powered self-disclosure. Furthermore, the risks of deception in a user-directed image search engine differ from those of an advertising service or a community like that of YouTube. Given these specificities, Search discloses the use of generative AI within a tool called “About this image.” Users can access this tool by clicking on three dots next to an image search result. Then, with additional space to explain what the disclosure means, “About this image” can put forth the provenance information.

Our learning from developing and launching these solutions has led us to issue the following recommendations when considering direct disclosure on generative AI content:

1. **Thoroughly analyze benefits and downsides of prominent labels for your use case**
   - Because of the downsides our research showed, we recommend the use of a prominent label only when we know the risk of harm from deception is high.
   - Evolve prominent labeling approach as use of AI increases.
   - Evolve prominent labeling approach as user understanding changes.
2. **Where possible, have an entry point to “learn more” about content**
   - Include information about provenance and such other useful information-literacy cues as source information.
   - Ideally, the entry point is as consistent and clear as possible but can vary depending on product constraints.
3. **Provide sufficient UI scaffolding to forestall any misinterpretation of provenance information**
   - Give enough context for the provenance data to prevent misunderstanding. This requires more space than can often be provided on the content itself.
   - Avoid presentations of provenance information that could be used as inaccurate heuristics (e.g., 10 percent of this image was edited with AI).

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.

In light of the implied authenticity and implied truth effects that resulted from labeling some content as made with generative AI and leaving other content unlabeled, one potential solution was to label all content. The hypothesis was that if you place a label directly on top of content that states if it is known to have been created or edited with generative AI, captured authentically, or that its technical provenance is unknown, any implied effects would be mitigated because all content would be accompanied by a declarative statement.

When evaluating this solution, we went back to the primary user need we were hoping to address: to help users determine the trustworthiness of content. Because knowing how something is made does not necessarily tell you if you should trust it, highlighting this information prominently on all content did not meet — and could actively contravene — the primary user need to determine the trustworthiness of content. In other words, knowing how something is made is merely one piece of the puzzle in determining content trustworthiness; it shouldn't be highlighted as if it were the only piece of the puzzle. In addition to this concern, we considered the risk that users would become blind to a label that appeared on all content. Therefore, placing a label on top of all content to indicate whether it was made with generative AI or has unknown provenance was rejected as a solution.

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

Currently, Ads, YouTube, and Search employ direct disclosure solutions on content for which provenance information is available. For Ads, that includes self-disclosed data. For YouTube, it includes self-disclosed data and C2PA manifest data. For Search, it includes SynthID, IPTC metadata, and C2PA manifest data.

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?

Propagating false information or harmful manipulated media (e.g., footage taken out of context or misleading clickbait) on Google platforms undermines our users' trust and causes harm to individuals and society. Google takes this threat very seriously because it cuts to the core of our mission, which is to organize the world's information and make it universally accessible and useful. As a leader in AI and a vocal proponent of its positive impacts, Google is committed to help society tackle problems that might arise from AI's misuse.

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 (3.3)?

We measured the impact of our direct disclosure solutions via user studies prior to launch. Our launched solutions were the most comprehensible of those tested. The implied effects and misinterpretations we uncovered were mitigated by using prominent labels only sparingly. In addition, we consulted information-literacy experts, to inform our direct disclosure principles. We will continue to monitor the impact of our direct disclosure solutions.

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?

Our explorations and experience launching direct disclosures have highlighted the importance of industry wide alignment on the goals of direct disclosures. For example, alignment on determining content trustworthiness as a primary goal would help establish such goal metrics as user comprehension of provenance text and such non-goal metrics as whether a label inadvertently implies anything about the trustworthiness of content.

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

One primary learning from our research is that this problem area is evolving rapidly, and continuous research and iteration is needed to create the healthiest information ecosystem for our users. We will continue to evaluate and learn from our direct disclosure solutions as user understanding and use of generative AI technology evolves.

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?

See answer to Question 2 in Section 1 above.

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?

The principles we've extracted from this case study are pasted again below. We hope they can be helpful to other practitioners and policymakers, and we welcome discussion:

1. **Thoroughly analyze benefits and downsides of prominent labels for each use case**
    - Because of the downsides our research showed, we recommend the use of a prominent label only when we know the risk of harm from deception is high.
    - Evolve prominent labeling approach as use of AI increases.
    - Evolve prominent labeling approach as user understanding changes.
2. **Where possible, have an entry point to "learn more" about content**
    - Include information about provenance and other useful information literacy cues such as source information.
    - Ideally, the entrypoint is as consistent and clear as possible but can vary depending on product constraints.
3. **Provide enough UI scaffolding to prevent misinterpretation of provenance info**
    - Offer enough context for the provenance data to prevent misunderstanding. This requires more space than can often be provided on-content.
    - Avoid presentations of provenance information that could be used as inaccurate heuristics (e.g., 10 percent of this image was edited with AI).

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)?

We described much of the research that informed our organization's understanding of direct disclosure in the case study section. In addition to that research, we relied on highly informative external research. We have summarized our learning and position, as well as compiled a list of citations in Google's white paper, "Determining trustworthiness through context and provenance."

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?

Yes, as described in the case study, we believe some risks of over labeling are:

- Labels emphasizing how content was made will imply that this is the information users should use to determine content trustworthiness.
- Unlabeled content will be seen as "real" or "authentic."
- Labels will be ignored or become useless. As more and more content is created with AI, labels will become less useful as a tool for discerning the differences among content.

We agree that users sometimes want to know how content is made, and we think they should be able to find that information. This is why we advocate for low-prominence direct disclosure whenever possible.

Finally, because the primary user problem we aim to help address is determining the trustworthiness of content, any solution will have to be dynamic and multifaceted.

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*.

We feel strongly that only synthetic content, wherein the risk of harm arising from deception is highest, should be prominently disclosed to our users.

We would be very concerned about over labeling in light of the risks outlined in the case study section and Section 2 above.

In addition to those risks, a further risk of over labeling is that we can show only so many pieces of context to users very prominently; focusing on the AI-generated or edited nature of content risks depriving us of the opportunity to signal more relevant or helpful information to users (e.g., information about meaningful changes to the content, its source, when the content was first posted, etc.) depending on the surface.

As Claire Leibowicz of PAI has noted, user-facing labels "can often be perceived as paternalistic, biased, and punitive, even when they are not saying anything about the truthfulness of a piece of content."

We expect that this could evolve due to any set of factors including new legislation, new research, or risks we have not yet considered.

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?

In general, Google and YouTube have long been attentive to the preservation of content that is Educational, Documentary, Satirical, or Artistic (EDSA). For instance, many of our policies have exceptions that apply to these types of content.

When it comes to creators signaling that a piece of content may be AI-generated or edited, however, we are keen to encourage broad adoption of best practices (including self-disclosing meaningfully edited, realistic EDSA content that average people might mistake to be authentic), especially since it is not the case that every use of AI would result in a prominent disclosure or any sort of penalty on our services.

Furthermore, there is no easily scalable method for categorizing content as either "creative" or "purely information rich." Five hundred hours of video are uploaded to YouTube every minute, and the Google Search index covers hundreds of billions of web pages. Machines do not currently have the capability to make a sufficiently accurate distinction between "creative" and "purely information rich" content, especially at this vast scale and given the subjective interpretation required in drawing such a distinction. It would therefore not be practical to apply separate policies to one category of content versus the other.

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?

We believe all layers in the provenance chain have a role to play in helping users determine what to trust online:

- Builders, inasmuch as possible, should include interoperable and verifiable provenance signals in content generated or edited with their models. (Exceptions include some B2B uses.)
- Creators should proactively disclose AI-generated or meaningfully edited content that's realistic, especially when it applies to verticals in which the most harm would arise from deception.
- Distributors should make informed and reasonable decisions on what information to impart to users and in what format, helping them decide what to trust while minimizing the risks of inviting confusion, fatigue, or ancillary harms such as the implied authenticity effect.

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 it is important for industry to continue to discuss and share best practices or technological advances and research in the space of responsible marking and labeling of synthetic media.

In particular, alignment on the exact meaning of back-end signals — which PAI has described as indirect disclosure — is essential if they are to be understood across the industry. For example, if someone uses the "Digital source type: Trained algorithmic media" field in IPTC, is the meaning of that field understood in the same way by the **Builder** who appends that information and the **Distributor** who receives it? And if that field is used to generate a direct disclosure, how should that be communicated to end users?

However, we believe this issue tends to be the exception rather than the rule. In general, different services may take varied approaches toward what media warrant disclosures and, in particular, what visual language makes sense in the context of their own products.

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)

PAI's description of the purpose of direct disclosures is to "mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation."

If we understand the stated purpose of direct disclosure correctly, direct disclosures are one among many signals that can help users determine what to trust online. Given this interpretation, the challenges would include:

- Reliability and interoperability of signals powering disclosures;
- Adapting to shifting user expectations and literacy over time; will users need as much information on AI-generated content three to five years from now, or will they learn to cross-reference content even when it seems real?;
- Risks that over labeling will trigger the implied truth and implied authenticity effect;
- Risks that the "liar's dividend" will settle in, whereby bad actors suggest that authentic content is AI-generated and simply bypass detection mechanisms; and
- More broadly, mismatches in expectations between what disclosures can achieve and what many would want them to achieve — they are not a silver bullet, and no approach to disclosures will fully resolve the risk of users being deceived by generative AI content.

If direct disclosures are rather meant to exist as an end in themselves — suggesting that they are the most important/essential piece of context users should be aware of regarding any piece of content — challenges would include:

- Fundamental mismatches between the intended goal (helping users determining what to trust) and likely outcome (user confusion, as most of the time it won't be clear what to make of a notice that something has been AI-generated); and
- Higher-magnitude versions of the implied truth and liar's dividend risks outlined above.

2. What is your organization hoping to accomplish by implementing direct disclosure? Does your organization believe directly disclosing ALL AI-edited or generated media, is useful in helping accomplish those goals?

See answers to Section 2 and Section 5, Question 1.

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
6) Anything else we're missing that is relevant here

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 direct disclosures about how content is made are not intended to communicate anything about the accuracy of the content on which they are applied. We are concerned about unintended implications on user perceptions of content accuracy and this is why we recommend exercising caution when prominently displaying provenance direct disclosures.

2. **Trustworthiness:** Similar to accuracy, our direct disclosures are not intended to communicate that content is more or less trustworthy. We have evidence that direct disclosures about content provenance can imply that unlabeled content is more trustworthy. This is why, again, we recommend exercising caution when prominently displaying provenance direct disclosures.

3. **Authenticity:** We emphasize that just because something is made without the use of AI does not mean it is "authentic." However, surfacing signals that communicate that a photo was taken and no edits were made can be useful to end users.

4. **Harm mitigation:** We encourage weighing the risks of direct disclosures that we outlined above with the potential harm that might come from deception. As stated above, for sensitive topics it may be worth the risk of implied effects and some user confusion in order to be able to assert the provenance of sensitive content.

5. **Informed decision-making:** We believe that provenance information is one important signal that can help users make informed decisions about what they trust online. We believe it is one signal, but not the only signal that is important to inform this decision-making.

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?

As defined by its authors, the liar's dividend has the potential to make it more challenging to determine the trustworthiness of online information; accordingly, it can result in making it harder for people to make informed decisions about topics of great importance for their livelihood or civic participation. Google does not form predictions or assessments about "tipping 'points" on this issue. That said, the concern that people may grow to doubt the authenticity of "real" content is one of the reasons we are adopting a cautious approach to direct disclosure for AI-generated content.

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?

We believe the most important questions people routinely ask themselves when encountering online content are:

1. Do I care about this content's reliability/trustworthiness? (For example, if a piece of content is consumed as satire/entertainment, its reliability may not matter to the people engaging with it).
2. Assuming I do care, how do I know whether I can trust this piece of content?

Over the years, Google has invested significant effort in empowering users to address Question 2. We believe the best we can do to that end is provide them with clear and understandable context on which to base judgment. We try to provide answers to questions including:

- How old is this piece of content?
- What is the source of this piece of content?
- Where else might this piece of content appear?
- How are other sites/domains referring to this piece of content and/or to the domain where it is hosted?
- What are other sites/domains saying about the topic of this piece of content?
- Is there authoritative information pertinent to this piece of content that I should be aware of (e.g., an election date or modalities; medical information about a disease, etc.)?

We've long worked with media literacy experts to develop and refine such contextual features as "About this result" or "About this image" in Google Search, or our "Information Panels" in Google Search and YouTube. We will continue to do so over time to help users navigate the questions above.

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.

We strongly believe further research is required in all areas discussed in this document including, but not limited to, better understanding the needs of users including creators, advertisers, and consumers; literacy; comprehension; and impact around disclosures.

There also needs to be *ongoing* research in light of the rapid pace of evolution in all the above, as methods that may be fit for purpose in one year may be out of touch the following year, given how fast users' AI literacy and expectations of online content are evolving.

Sample research questions could include:

- Which types of information help an end user discern between trustworthy and untrustworthy content?
- At what point of ubiquity does a disclosure become unhelpful?
- What media literacy interventions help the greatest number of people increase their ability to discern between trustworthy and untrustworthy content?
- How do disclosures impact creators and advertisers?
- Who should deploy those solutions for the most public good (while mitigating the likelihood of public harm)?

> *We strongly believe further research is required in all areas discussed in this document including, but not limited to, better understanding the needs of users including creators, advertisers, and consumers; literacy; comprehension; and impact around disclosures.*

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?

We are strong believers in the value of helping the public to be more informed consumers of online information, including but not limited to synthetic media.

This has included efforts like:

- Developing contextual features (like those mentioned above) to foster education by means of product design — including tools like "About this result" or "About this image," based on the SIFT framework developed by Mike Caulfield; and
- Investing in efforts of our own such as YouTube's "Hit Pause" campaign.

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?

We believe it is important to collaborate across society on media literacy campaigns that are empowering versus alarmist and that involve creators and voices that are relevant to their recipients.

3. What support does your organization need in order to advance synthetic media literacy and public education on evaluating media?

We are always looking for more ways to contribute to public education in this space and for more partners to relay our work, as described above.