PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How Truepic used disclosures to help authenticate cultural heritage imagery in conflict zones

truepic

This is Truepic'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.

Truepic provides enterprise-grade authenticity infrastructure solutions for digital content and workflows, combining advanced security, dedicated customer support, and customized integrations for the world's largest enterprises. Truepic's tools allow any organization to add and display transparency in authentic and synthetic media across the entire digital media lifecycle. As a founding Coalition for Content Provenance and Authenticity (C2PA) member, Truepic ensures seamless adoption and deployment of provenance (what Partnership on AI [PAI] calls indirect disclosure) technology, maintaining content authenticity at scale in an evolving digital landscape. Recognized as one of TIME's Best Inventions and Fast Company's World-Changing Ideas, consumers and businesses rely on Truepic's technologies to make informed decisions.

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 or as another kind of stakeholder that the Framework does not specify? Should the Framework be edited to add your category of stakeholder? Why or why not?

Truepic does not create or distribute synthetic media. Instead, we provide indirect and direct disclosure mechanisms to promote transparency in all digital media, including synthetic media. The PAI Framework meaningfully includes stakeholders working on transparency mechanisms to support both indirect and direct stakeholders.

2. Please elaborate on how your organization supports direct disclosure (as defined in our Glossary for Synthetic Media Transparency Methods) in the technology ecosystem.

Truepic's indirect disclosure tools, such as software development kits (SDKs), libraries, and other tools, are designed to facilitate direct disclosure. Truepic's SDK facilitates direct disclosure by using a built-in C2PA secure signing solution, which adds indirect cryptographic hashes (Content Credentials) to digital content.

Additionally, Truepic maintains direct disclosure code, which is publicly available for anyone to integrate. Truepic's Content Credentials Display validates and, as the name suggests, displays C2PA provenance history. Both are currently being used by partners such as Microsoft, Hugging Face, G&L, and many others for synthetic and authentic media disclosure.

3. Does your organization understand the goal of direct disclosure as specified in the PAI Framework: "to mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation?"

Yes. Truepic was founded in 2015 to help build a more transparent world. We believe that enhancing transparency and authenticity in the content we see and hear online is the most feasible approach to improving our information ecosystem. To successfully achieve this goal, we also believe that interoperability is critical as digital content moves throughout the internet. Without interoperability, fragmented approaches could lead to confusion, inconsistency, and a breakdown in the shared understanding necessary for maintaining a reliable online environment.

Truepic helped co-found the Coalition for Content Provenance and Authenticity (C2PA) with a wide range of partners to create an interoperable specification to share digital content and ensure direct disclosure at scale is possible. We agree with and support PAI's stated direct and indirect disclosure goals.

4. What, if anything, from your organization's approach to supporting direct disclosure is missing from this NIST taxonomy? Should it be added, 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).

Truepic's direct disclosure addresses nearly all of NIST's suggestions including content labels and disclosure fields. We consider the Content Credentials signal — "CR" — to be a visible watermark that is displayed on pieces of digital content in Truepic's direct disclosure library. Although Truepic provides direct disclosure services to partners, it does not operate a consumption or distribution platform. That's why we recognize that each platform may have differences in direct disclosure. In terms of direct disclosure, we believe that this guidance is largely sufficient. However, one point we would like to emphasize is that direct disclosure should be backed by secure indirect disclosure signals to ensure accuracy. We explain this concept below.

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." Does this align with how your organization understands direct disclosure? What criteria does your organization recommend be used to determine whether content is directly disclosed?

Yes, Truepic aligns with this recommendation. Truepic's tools, specifically its secure indirect disclosure tools, are actively used by AI platforms to power direct disclosure on visual and auditory media that is produced. One of the most notable examples of Truepic's implementation of PAI's recommendation is with Hugging Face, where we partnered to power two different spaces where **Creators** can generate synthetic images and immediately view their direct disclosure thanks to C2PA's Content Credentials. We are encouraged to see some of the most used AI platforms deploying indirect and direct disclosures. We also recognize that countless AI platforms, including open-source models, are used without disclosure mechanisms.

Adding disclosures to non-AI content, or authentic camera-generated material, is equally important to identifying AI content. This approach to non-AI content can help increase overall transparency online, giving content consumers visibility into the origin and history of digital content. Further, while the zeitgeist is related to transparency in AI-created content, "cheapfakes" or rudimentary content manipulation is also an ongoing challenge. In terms of the criteria on what content should apply direct disclosure labels, we believe that all AI-related content should be signed with indirect and direct disclosures at the point of creation. There should be no uniform criteria for non-AI content, but we believe content creators should have the option to add credentials at creation — which we explain further in Section 3.3.

3 Real World, Complex Direct Disclosure Example

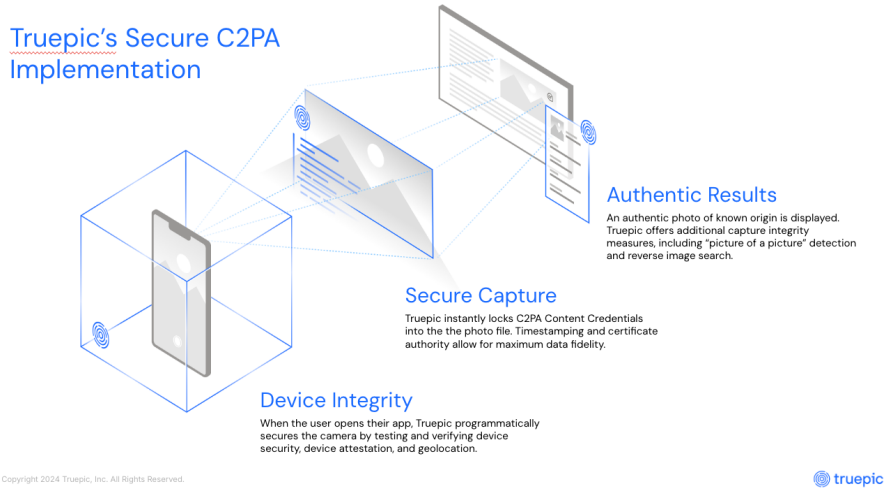
1. Provide an example in which your organization supported direct disclosure on a piece, or category, of content. Was it challenging to support direct disclosure in this example? If so, why? If not, can you provide an example of a detail that, if different, would have made it challenging?

Project Providence, the first end-to-end image documentation platform collaboratively developed by Truepic and Microsoft, is dedicated to preserving cultural heritage and nationally significant infrastructure in Ukraine. This initiative empowered its U.S. Agency for International Development (USAID)-funded implementing partner, the Anti-Corruption Headquarters, to meticulously document over 500 attacks throughout Ukraine, capturing a wealth of visual evidence through more than 1,500 images and videos. Furthermore, prosecutors across Ukraine have effectively leveraged the media outputs from Project Providence, incorporating its robust C2PA indirect and direct disclosure outputs in no fewer than ten high-profile accountability cases. These images have been instrumental in bolstering legal proceedings, awareness, and accountability claims. In addition, the Anti-Corruption Headquarters has extended the reach of this crucial documentation by sharing these authenticated images with multilateral authorities worldwide. This global dissemination underscores the platform's pivotal role in fostering international cooperation and support for Ukraine's cultural heritage preservation efforts.

2. How did technical mechanisms help accurately establish the origin and history of the content to enable direct disclosure?

Truepic's tools adhere to the C2PA open specification, which we believe is the optimal approach for scaling transparency in digital content because it is interoperable and tamper-evident. We have implemented this technology in various scenarios, with a particular focus on our collaborations with Microsoft on Project Providence in Ukraine and Content Integrity Tools (CIT). The Truepic Lens SDK enables the authentic capture of images, videos, and audio, enhancing transparency in conflict zones and global elections. Lens leverages the C2PA specification, which prescribes the use of cryptography to secure indirect disclosure signals within the media being captured.

Having secure, attested, and authenticated metadata that is cryptographically hashed in the form of indirect disclosure allows for accurate results to be displayed in the form of direct disclosure.



Truepic's technology ensures indirect and direct disclosure are accurate by authenticating metadata at device or creation system level.

In the Project Providence example in Ukraine cited above, Microsoft had its own direct disclosure mechanism built into Project Providence. For Content Integrity Tools, Microsoft has built its own direct disclosure tool. Truepic also supports direct disclosure for other partners through a hosted JavaScript library that can be integrated into any website or platform and used on Hugging Face spaces at no cost. This allows a display for direct disclosure directly on any platform or website. For use cases in which content is viewed on mobile devices, the Lens capture app has a built-in display with disclosure for authentic media.

We successfully implemented a functional and reliable technical direct disclosure mechanism and aim to continue to improve its usability. It's important that the disclosure accurately represents the underlying metadata and that users clearly understand the information presented. This remains a challenge that our team, which also contributes to the C2PA's User Experience (UX) Task Force, is diligently addressing.

Ensuring Accuracy

Accurately establishing data in disclosure is a pivotal issue. Truepic staunchly supports the secure implementation of the C2PA specification to ensure accurate metadata is cryptographically secured into the media file. The Truepic Lens SDK, which underpins the use cases mentioned earlier, not only adds C2PA metadata but also safeguards the device against tampering and authenticates the digital content at the moment of capture. This includes verifying the integrity of the capture device's operating system, ensuring that location information is untampered, and adding an original timestamp from a server rather than relying on device time, which can be altered.

Moreover, the SDK also frequently rotates the certificate used on each device to maintain anonymity while allowing for granular invalidation of images if any issues are detected. For authentic image capture, the SDK secures the capture process and associated sensors to ensure accurate pixel contents and metadata, including date, time, and location. Another critical area and common concern in C2PA is key security. Hardware-backed keystores provide a higher level of security as the keys are stored in isolated hardware designed to be tamper-resistant.

The C2PA has focused on this critical question as well and established its conformance working group to develop best practices and guidance on the optimal implementations of the C2PA specification. This will help guide implementations across the industry to raise the level of accuracy and security in direct disclosure mechanisms aligning with the C2PA standard.

---

3. 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.

Direct disclosure for AI platforms should be standard and we do not see significant reputational harm in disclosing that a piece of content is AI-generated. For non-AI content, we do believe that the **Creator** must opt in or choose to implement direct disclosure because there could be scenarios in which disclosing metadata could be harmful for **Creators**. Truepic has deployed its technology with documenting teams in conflict zones such as Syria, Ukraine, and Somalia. One priority consideration is the potential sensitivity of direct disclosure on the content captured, it is very possible that content creators may not want credentials or disclosure associated with authentic content — this is especially true in high-threat environments like conflict zones because the location or timestamp of content could be sensitive. **Creators** must be given the option to "opt-in" and clearly understand what details are associated with the content they are creating. In all scenarios, our partners have been educated on the details and opted in for secure metadata to be indirect and/or directly disclosed. This is a key programmatic consideration for disclosure.

Technically, disclosure implementations that fail to accurately reflect secure and verified metadata can cause reputational damage and negatively impact both the implementer and confidence in the broader specification itself. This is a genuine concern, which is why Truepic advocates and implements secure and authenticated disclosure methods (see above) to mitigate the risks associated with adding transparency markers to digital content.

Pointing back to the Ukraine example or any situation in which documentation is critical, secure indirect disclosure undergirds the value of the disclosure. Therefore, if an insecure method of indirect or direct disclosure displayed metadata that was not securely hashed and tested, and it would then misinform content consumers with incorrect time, date, creation, system, or other information. Ultimately, this would lead to a decrease of faith in the ultimate goals of the organization documenting content with indirect and direct disclosure technology. This would allow bad actors to deploy the liar's dividend like was done in past conflicts such as Syria. The existence of unproven or potentially faulty data would be used as the pretext to undermine all claims in the non-permissive environment. This would make the claims of the documentation groups harder to leverage in accountability hearings or procedures. This has happened in events such as natural disasters, in which manipulated or AI-generated images of destruction backfired to hurt the initial cause reputationally.

Achieving Secure and Accurate Disclosure

We believe there are two key concepts to help achieve the most secure disclosure outputs:

1. **Metadata:** Ensuring that indirect and direct disclosures are accurate, attested, and aligned with confirmed details such as time, date, and location.

2. **Certificate:** Ensuring that the operator of the disclosure tools has been vetted and issued a certificate by a trusted authority, thereby upholding the integrity and trustworthiness of the system. This prevents the risk of unverified organizations using certificates from untrusted sources, which could undermine the credibility of the C2PA signing process.

---

4. 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 consider Project Providence to have been a resounding success, which led to the expansion of its work and continued use of technology for other documentation purposes like election monitoring around the world. Most recently, technologies with similar methodology and make-up to Project Providence have been used in the U.S., Venezuela, Mexico, and soon other countries for election monitoring. Most projects are ongoing and the results have not been established yet.

---

5. Is there anything your organization believes the Builder, Creator, Distributor, or another stakeholder in the content pipeline, should have done differently to support direct disclosure?

In the context of Ukraine, we believe the right stakeholders were involved. The project was not meant to scale across the internet, but rather give documenters risking their lives to capture digital content on the ground the most transparent and secure technology to do so. In short, this was a controlled case that gave the **Creators** the right tools to deliver the images where they desired.

In general, with regard to scaling direct disclosure, we believe that much more can be done to advance the deployment of disclosure online. To sustainably create and expand a more transparent information ecosystem with disclosure at an online scale, adoption and

education must increase across the board.

- **Adoption:** The C2PA open specification needs to be adopted quickly, and C2PA-compliant technology should be deployed globally across the digital ecosystem in the form of direct disclosure. Without widespread adoption across the internet's most-used platforms, inadvertent edits and media processing workflows may disrupt provenance trails, thereby limiting the scale of disclosure. Many distribution platforms strip metadata out of digital content posted on their sites. By opting to support C2PA, they can effectively disclose to their users whether content has been AI-generated or not, compared to manually labeling content, which is sometimes done through content moderation. When media files contain Content Credentials, it is for the explicit, net-positive purpose of transparency. If an authentic capture has Content Credentials, it means the **Creator** knowingly opted to add transparency to their work. If synthetic content has Content Credentials, the generative media platform adhered to best practices and was committed to directly disclosing its outputs.
- **Education:** There is a clear need to educate enterprises, publishers, and governmental agencies on why digital content transparency is a best practice and should be widely adopted and implemented online. Additionally, increasing media literacy efforts for content consumers is critical so the general public knows to look for the provenance of digital content before making any consequential decisions. Information integrity initiatives should be funded globally to provide awareness, availability, and training on how provenance tools and other direct disclosure mechanisms can better inform consumers about the origins of online content.

---

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

We would not have done anything differently in our work in Ukraine. We moved slowly and methodically to ensure that programmatic and technical design practices met the needs and security requirements for our partners and program. While getting tools to our partners even sooner than delivered would have been ideal, we remain confident that the methodical approach was the most prudent method. In terms of the level of disclosure, we are satisfied with the style and substance of direct disclosure as it was highly correlated with our partners needs and interests.

With regard to scaling direct disclosure across the internet, we remain confident in our decision to prioritize secure indirect and direct disclosure mechanisms, as they are fundamental to scaling transparency across the internet. However, we were taken aback by the extent of conflation among various transparency mechanisms, including provenance (secure metadata), watermarking, fingerprinting, and detection services. While PAI's Glossary provides a helpful framework for these concepts, there is a pressing need to enhance education and awareness. This will help stakeholders better understand the distinctions, nuances, and overlaps between these mechanisms. We have already adjusted the course by working with partners and in various forums to help educate and raise awareness as a priority.

7. What technical mechanisms did your organization rely on to ensure content was accurately disclosed? Are there any policy instruments that could either 1) support technical methods for organizations to better authenticate content or 2) better connect technical authentication to direct disclosure? If not, should there be?

The Project Providence platform and Content Integrity Tools leverage Truepic's SDK for secure image capture and Microsoft's Azure cloud to securely store and display data. Both are fully interoperable due to adoption of the C2PA open standard. Secure signing occurs locally on verified mobile devices without compromising privacy thanks to automatic daily key and certificate rotation. With this technology, media modifications can be detected using C2PA validation and the authentic media source can be proven. We believe that relying on secure environments and attested systems to implement indirect disclosure is the optimal way to ensure accuracy in direct disclosure.

There are various policy instruments that have been established or are being discussed at the federal, state, and multilateral levels, most notably direct instruments like:

- **U.S. Federal:** Executive Order on AI, White House Voluntary Commitments, Protecting Consumers from Deceptive AI Act, and National Defense Authorization and Act (NDAA) all mandate the use of disclosure in AI outputs.
    - NIST's review and work examining optimal transparency standards is a key forum supporting this work.
- **State:** Various state initiatives are supporting disclosure in Gen AI, like Utah's SB 131 mandate's use of provenance ahead of elections; and a variety of pieces of legislation in California such as AB 3211 and SB 942.
- **Europe:** The E.U.'s AI Act Article 50 mandates transparency in generative outputs in Europe and the U.K. Safety Bill.

Further, indirect policy instruments like "voluntary commitments" in the U.S., Europe, and Australia have also helped push the industry forward on disclosure. For example, the industry-led AI election accord made on the margins of the Munich Security Conference in 2024 helped drive some of the most significant disclosure projects ahead of elections.

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

We believe that Project Providence showed that secure indirect and direct disclosure can have significant utility for authentic documentation in any use case. Policymakers should consider how operations, systems, and digital processes could be improved by this approach or technologies increasing transparency in digital content. Collecting, crowdsourcing, creating, and distributing media backed by secure indirect disclosure and then directly disclosed to audiences has efficacy proven by the examples from Ukraine. While this belief is supported at the highest levels of government worldwide to protect against AI misuse, policymakers should also consider its utility for non-AI content, too — like Project Providence.

We live in a digitized world where the online information we access is crucial to our decisions, and this year is pivotal, with billions voting globally. Disclosure can also be considerably important in conflict zones, which are often non-permissive and become fertile ground for deceptive media and misinformation. Truepic and Microsoft aimed to showcase this in both Project Providence and Content Integrity Tools.

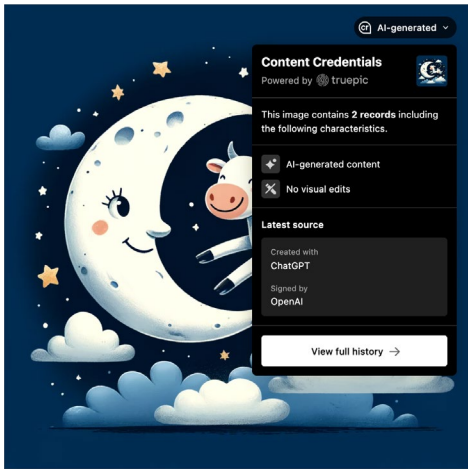
Due to the interoperability of the C2PA open standard, the newly captured digital content and its credentials can be visible on any compliant website, platform, or verification page. By embedding Content Credentials into the media at the moment of capture, each piece of content can be verified for authenticity and origin, thereby helping prevent and pre-empt tampering but also to help stem the spread of the liar's dividend — bad actors claiming that authentic images are fake.

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)?

Early user research helped us understand the current level of media illiteracy and empathize with content consumers who did not spend too much time examining details about content. We found that information needs to be front and center in the user interface, not buried in a display. There was inherent acceptance to believe displayed details shown because accuracy was generally assumed. Viewing a direct disclosure display overwhelmingly helped improve our understanding of the media; we understand this from a variety of internal research and focus groups. Also, external early research on the efficacy of Content Credentials also helped contribute to our understanding. Our current display library is a result of our early research and information, such as "AI-generated content," is displayed up front for content consumers to view in addition to the creation system and signer.

Further, the C2PA's UX task force has implemented guidance and suggested best practices for optimally effective disclosure and user understanding. PAI has also worked on disclosure formats, which helped inform our thinking. Other studies, such as those completed by researchers at Oxford University and the University of Washington, also helped increase our understanding and approach to direct disclosure. As stated earlier, this is an area of continued growth, and we believe that a multistakeholder approach will help its refinement.



Prompt "Cow Jumping Over the Moon," entered into Open AI's DALL·E. Indirect disclosure via C2PA Content Credentials are embedded into all DALL·E outputs.

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, there are risks associated with over-disclosing synthetic media to audiences. In a world where every editing tool uses AI, a certain amount of flagging blindness will develop, and people will no longer see every disclosure relating to AI-generated content. Over time, we'll need more granular logic regarding whether to show AI disclosure or not. The C2PA specification will likely evolve to show more nuanced regions and areas of interest in a media file. Users will eventually be able to determine if the AI edit was significant or not and view other rating levels of how many modifications were made to the image. Truepic is actively involved in the C2PA's various committees, such as the UX, technical working group, and conformance committees, to navigate these discussions and continually iterate on the specification.

We also believe there are risks associated with under-disclosing synthetic media to audiences. Today, some Content Credential displays show only surface-level data of the most recent modification, which may not involve the use of any AI system. Details about the use of AI can easily be buried too deep in the manifest, potentially suppressing important file details that are not shown to the user. Another example is when the C2PA signing process occurs after a file has already been altered by AI. If C2PA Content Credentials are not attached to that file upon creation prior to any modifications, the absence of AI callouts in the history will give the incorrect impression that AI was not used.

We constantly iterate on our display to balance any known risks associated with disclosing synthetic media to audiences. Truepic regularly uses feedback, like partner comments, and considers the risks stated from C2PA working groups to make assessments. We implement and automate C2PA signing across diverse platforms to seamlessly incorporate signing with Truepic certificates into server- and cloud-based media workflows. Original details of a media file are verified and cryptographically sealed upon creation, AI-generated files are signed in real-time, and modifications to the file are recorded in its history.

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 continuously recalibrate and refine our approach to ensure that our disclosure mechanisms evolve alongside audience understanding and engagement. As adoption increases, we anticipate that consumers' ability to comprehend and digest these mechanisms will also grow and change. Therefore, Truepic is committed to responding swiftly to feedback, monitoring discussion forums, and incorporating guidance and user experiences to maximize the efficacy of our disclosures. While we do not believe that content is currently over-labeled, it is crucial that any type of label we or our industry partners deploy is comprehensible and effective in achieving our shared goals. For example, if consumer feedback and research shows that more or less information should be displayed on the "Level One," or first level, technical information on the display, then we will adjust our display to address the concerns. This highlights that users or consumers can see different pieces of information, such as if AI or camera captured, date, time, location, and edit history at different display stages based on what they click into. Similarly, the same can be done on the "Level Two" information, which would be when the Content Credentials are expanded to provide more information on the content.

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?

We ultimately believe that a tiered labeling system will be necessary for digital content. The security, use cases, and systems delivering transparency are all variables that can impact the spirit of the disclosure. Creative works are certainly a part of this discussion and we understand PAI's position that there should be a distinction for how they are labeled from information-rich content. Organizations should consider the system, context, history, and other aspects of the creative work before addressing how it should be disclosed. However, we would emphasize that other pressing labeling issues must be addressed first, such as the quality and security of the disclosure data. For example, is the data being disclosed securely captured and verified? If so, that label should appear differently to non-securely captured data.

In reference to general labeling differences for creative works, there are some challenges at the moment to truly introduce this tiered system of labeling:

- The C2PA spec does not yet have a tiered system of labeling to enact.
- Labeling is not yet uniform and may never be. Guidance may be in the form of best practices but not conformance. Therefore, platforms will likely have different disclosure standards and best practices.
- As the specification moves to version 2.1, there is less emphasis on style or type vs. the system and creation origin of the content.

In terms of assessing the mindful pursuit of labeling requirements, we believe the iterative process that takes feedback from a larger community is the optimal way to achieve the highest efficacy label. High-efficacy labels should be widely understood across different platforms, regardless of UX differences in each implementation.

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?

Ultimately, the distribution platforms and their stakeholders (e.g., content delivery networks or CDNs) bear the primary responsibility for enabling disclosure. This is why we emphasize that adoption is the most critical factor in scaling disclosure. The entire distribution pipeline should strive to facilitate disclosure to achieve optimal outcomes. However, we maintain that **Creators** should have the autonomy to opt into disclosure for non-gen-AI creations like authentic imagery or video because there could be sensitivities associated with metadata like location. Disclosure should never be mandatory for non-AI content; thus, **Creators** must choose to participate. For AI-generated content, many platforms like DALL·E and Bing Image Creator implement indirect disclosure by default. **Creators** should be aware that they are using platforms that prioritize transparency and responsible practices regarding disclosure.

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.

1. **Builders**, **Distributors**, and others within industry should not determine what requires disclosure. That should come from stakeholders, including civil society, end users, consumer protection groups, and others. **Builders**, **Creators**, and **Distributors** should comply with the best practices established for labeling. It is critical that these groups align on best practices, standards, and/or thresholds for disclosure to allow for uniformity and consistency online. This will help accelerate education and understanding of Content Credentials.

2. In this instance, it would be helpful if there were shared understandings or parameters in which visual indicators and language were consistent across distribution channels. Shared UX will help drive understanding and digestion of direct disclosure material across industry, government, and populations. Further, it is a force multiplier in terms of adoption and compliance within industry categories.

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)

The primary challenges revolve around the interdependence between systems and the necessity for standardization in direct disclosure processes. The interdependence between various systems is necessary as content flows through multiple platforms and systems, each with unique protocols and standards. Ensuring seamless integration and communication between these diverse systems is critical for efficient and accurate content disclosure. This requires interoperability standards and frameworks to accommodate the technologies and platforms involved. Standardization in direct disclosure processes creates consistency and reliability. It is essential to establish and adhere to standardized protocols. This ensures all stakeholders can understand and trust the disclosed information regardless of their system or platform. Furthermore, the system must be capable of automating the disclosure process rather than relying on manual intervention.

Perhaps a larger social challenge, as stated earlier in the document, is education and understanding. Education is challenging because it requires significant resources and a holistic, unified approach to raise awareness among stakeholders across diverse industries, government, and society. Additionally, the rapid pace of technological advancements means that education must be continuous, which adds to the complexity and demands of maintaining an educated and informed user base. However, with a well-informed stakeholder community, the use of disclosed content will become ubiquitous and increase the chances of a more transparent internet.

2. When implementing direct disclosure, what goals should organizations be trying to accomplish? Does your organization believe directly disclosing ALL AI-generated or edited media, as several draft policies are recommending, is useful in helping accomplish those goals?

When implementing direct disclosure, organizations should aim for transparency and accuracy. Transparency involves providing clear and accessible information about the origin and provenance of the content. Accuracy ensures that disclosed information is precise and reliable, maintaining credibility and integrity. Full disclosure promotes transparency and accuracy by ensuring that audiences are aware of the involvement of AI in the media they consume. This openness creates a more informed and discerning public, enhances trust, and supports ethical content creation and dissemination standards. We do believe that, at minimum, disclosure that a piece of content is AI-generated and edited should be a standard in our information ecosystem.

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.)

For all the categories below, ensuring accurate and secure metadata is critical to ensuring that indirect and direct disclosure have positive effects. Likewise, unverified and faulty data that may be disclosed directly would have the opposite effect on both. The security and fidelity of the implementation are paramount.

- **Accuracy:** Secure and verified disclosure of digital content is expected to enhance the accuracy of the collective informational ecosystem. This in turn, will likely improve the decision-making accuracy of content consumers by providing them with verified, high-fidelity information.
- **Trustworthiness:** Early studies, such as those from Oxford University, indicate that verified disclosure can positively impact the trustworthiness of content. However, it's important to note that trust is a complex issue and cannot be addressed solely through direct disclosure. Further, we also recognize that some suggest that the increased prevalence of disclosure indicators would also degrade trust in standard images without disclosure.
- **Authenticity:** Version 2.0 of the C2PA specification enhances this by focusing on the content creation and handling system, making it easier to distinguish authentic content from manipulated material. By implementing C2PA 2.0, **Creators** and publishers can offer clear, verifiable information about their content, thereby increasing transparency and trustworthiness. This level of disclosure is especially important in an era where misinformation and digital manipulation are rampant. Users can readily access information about the content's provenance, ensuring they are engaging with genuine material. Furthermore, disclosure of synthetic media, such as AI-generated content, is equally vital by explicitly labeling AI-produced material with statements such as "authentically created from X platform," allowing consumers to differentiate between human-made and machine-generated content.
- **Harm mitigation:** Understanding the origins and history of a piece of digital content will help mitigate deceptive practices and fraud.
- **Informed decision-making:** Building on the response to "accuracy," we believe that most decisions of consequence have been digitized across industries and in our personal lives. Understanding and being able to view and curate content based on

verified history and creation will better inform decision-making by people, businesses, and governments.

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?

A critical tipping point is underway, occurring in the world's most crucial arenas, from elections to conflicts and accountability decisions. Bad faith actors have effectively derailed UN Security Council resolutions and accountability by undermining the authenticity of digital media from Syria, Burma, and other regions. Elections have been compromised by claims that security camera or CCTV footage has been tampered with, and defendants in criminal cases have alleged that video evidence of their crimes are "deepfakes." Additionally, businesses worldwide are adjusting their operating procedures based on the assumption that standard imagery is no longer reliable. In our view, these developments signal that we have reached this tipping point. There is no definitive endpoint to this tipping point; instead, it will only worsen as hyper-realistic deceptive content continues to proliferate and spread at an exponential rate. However, as trust in standard imagery continues to decline, the demand for transparency and disclosure in content will grow.

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?

It's important to consider whether the modification fundamentally changes the content or reality of the image. For example, adding snow to a summertime photo alters the scene. While enhancing the sky to make it brighter does not significantly change the content, it could actually enhance the image closer to what the human eye would see. This distinction is essential in understanding the extent of AI's impact on the media. Another critical question is, if the content is not AI-generated, where did it come from? Understanding the origin of the content helps assess its authenticity and reliability. Additionally, audiences should consider whether they trust the source of the content. Trust in the source is vital for evaluating the credibility of the information presented and for making informed judgments about the content's accuracy and intention. By asking these questions, audiences can better navigate the complexities of AI-generated and AI-modified media, ensuring a more discerning consumption of digital content.

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 can guide the development of direct disclosure practices that meet user and audience needs. Key research areas might include:

- **Efficacy:** Does the user access the disclosure information?
- **Understanding:** Is the information being understood by content consumers?
- **Challenges:** Is anything being misunderstood or perceived incorrectly?
- **Cross-culture:** How is disclosure perceived across cultures/languages?

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?

Yes, our organization agrees that broader education is required. We continuously seek to educate content consumers, enterprises, and government agencies on the importance of content transparency. For example, in May of this year, we participated as an exhibitor at the inaugural Special Competitive Studies Project's AI Expo in Washington, D.C. Over 13,000 people attended, including members of the general public, industry leaders, and policymakers. We also distribute our Trusted Future newsletter monthly to over 3,000 subscribers, featuring updates on key industry trends and developments. We've hosted educational sessions on Truepic's technologies at a workshop for Mexican photojournalists ahead of the Mexican election that recently took place. As a technical partner for Microsoft's Content Integrity Tools, we support their larger educational efforts by providing our specific technical knowledge and expertise.

Truepic has long worked with independent media outlets like the BBC, Canadian Broadcasting Company, and The New York Times to be a resource for educating the public about C2PA adoption and the importance of Content Credentials. We've worked with Project Origin to help establish the first C2PA-compatible list of verified news publishers. We have led and participated in educational workshops in the U.S., U.K., Mexico, Norway, Denmark, Australia, and New Zealand. We also work with many civil society partners to support education efforts around the world and run a social impact grant program, providing in-kind technology grants to support organizations doing critical work around the globe in securely documenting authentic media. To date, we have issued over $500,000 USD in technology grants to 12 partner organizations worldwide.

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 would like to see tech platforms, media creators, online publishers, news outlets, educators, and government agencies implement widespread adoption of digital content provenance and promote responsible practices for developing, creating, and sharing synthetic media online. We have seen a variety of mechanisms help promote implementation, including voluntary commitments like the AI Election Accord agreed to by over 20 companies. More investment is needed to train the general public about how to critically evaluate online content and how to navigate the digital landscape. We believe this is a unique position for public-private partnerships to make a difference by sponsoring such investments.

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

We would reiterate our response to Question 4 in Section 3 regarding the adoption of and education about disclosure. Those are most critical to deciphering — as a starting point — what is synthetic. From there, content consumers can make their own decisions on trustworthiness.

7 Commentary on the Framework's first set of cases (beyond direct disclosure)

1. The first round of cases did not just focus on direct disclosure, but also on broad exploration of several case themes: creative vs. malicious use, transparency via direct and indirect disclosure, and consent.

   We want to leave room for respondents to highlight any *other* areas of the Framework that can be deepened or improved upon to ensure its viability in a rapidly changing synthetic media ecosystem (related to the case themes above, and moving beyond the direct disclosure focus of this case template).

One area deserving a more in-depth analysis is the distinction between secure and attested indirect disclosure mechanisms, which are crucial for ensuring the accuracy of direct disclosure. As outlined earlier, a robustly attested indirect disclosure is essential to validate direct disclosure, thereby making them both accurate and effective in achieving the objectives of the PAI Framework.

Additionally, it is critical to expand on how disclosing non-AI-generated material can significantly enhance the effectiveness of AI content disclosure. Relying exclusively on responsible platforms and abiding by best practices are insufficient to combat the widespread threat of image and content deception. Malicious actors will continuously attempt to circumvent or ignore disclosure protocols, leveraging non-compliant platforms or models that eschew transparency measures to create deceptive content. Therefore, PAI should consider that enhancing disclosure practices for AI content to also include non-AI content can strengthen the overall ecosystem, making it easier to identify and isolate content that does not adhere to established best practices. PAI could consider that including disclosure on non-AI material could help further add transparency and better isolate deceptive content which will not have disclosure, particularly on critical use cases like elections, media, or commerce, etc.

2. Has the Framework improved any processes, procedures, or policies at your organization, or your organization's observations of those Building, Creating, and/or Distributing synthetic media?

The PAI Framework has been a very helpful baseline for Truepic and the industry in identifying and highlighting best practices in AI transparency. It helped create discussion parameters across policy circles supporting disclosure as a scalable and feasible approach. Moreover, the influence of this PAI Framework extends beyond theoretical discussions. It has actively shaped the strategies and operations of distribution channels, helping to set industry standards. We are particularly encouraged by the tangible outcomes linked to the Framework, including its direct impact on three of the largest social media platforms, which have now integrated disclosure methods as a key component of their operations. This correlation underscores the Framework's practical relevance and significant role in advancing transparency within the digital ecosystem.