PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How Microsoft and LinkedIn gave users detailed context about media on the professional networking platform

Microsoft

This is Microsoft'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.

Microsoft creates platforms and tools powered by AI to deliver innovative solutions that meet the evolving needs of our customers. We are committed to making AI widely available in a responsible way, with a mission to empower every person and organization on the planet to achieve more.

Microsoft attaches cryptographically signed provenance metadata to images generated with OpenAI's DALL-E 3 model in our Azure OpenAI Service, Microsoft Designer, and Microsoft Paint. This provenance metadata, referred to as Content Credentials, is based on Coalition for Content Provenance and Authenticity's (C2PA) open technical standard and includes important information such as when the content was created and which organization certified the credentials.

Users can verify provenance information using open tools such as Microsoft Content Integrity and the Content Credentials Verify tool. Provenance information based on the C2PA standard is automatically ingested by and displayed on our professional networking platform, LinkedIn. LinkedIn is the world's largest professional network with over one billion members in more than 200 countries and territories worldwide.

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 the Framework, Microsoft falls under the categories of both **Builder** and **Distributor** of synthetic media. We focus on our Distributor role in this case study since we are focusing on LinkedIn, a professional networking platform.

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

On criteria for disclosure:

We rely on machine-readable provenance information, per the C2PA open standard, to disclose provenance information. If such provenance information is detected, it will be directly disclosed.

To detect content that violates our policies (e.g., policies prohibiting fake profiles on LinkedIn, as described in the LinkedIn Professional Community Policies, we also use synthetic media detectors to identify synthetic profile images for removal (rather than disclosure). Notably, synthetic media detection is different from other mechanisms used for provenance verification. For example, as described in Partnership on AI's (PAI) Glossary of Synthetic Media Transparency Methods, **Builders**, **Creators**, and **Distributors** can proactively add signals to content, such as visible and invisible watermarks, fingerprints, and metadata, to assist in identifying synthetic media. See New Approaches For Detecting AI-Generated Profile Photos.

On practices followed:

The public can verify provenance information using open tools such as Microsoft Content Integrity and Verify. Microsoft's Content Integrity tool first checks the certification information, which includes when Content Credentials (based on the C2PA standard) were certified and by which company or organization. Then, it checks whether those basic details, or metadata, were, in fact, certified. If certified, the tool shows Content Credentials metadata linked to the file. The tool also shows if changes to the metadata are detected. This allows people to make an informed decision about the trustworthiness of the file that was checked.

We also provide direct disclosure to LinkedIn users by ingesting provenance information based on the C2PA open standard and displaying associated provenance information for media in a user's feed. Image and video content cryptographically signed using C2PA Content Credentials will be noted with the C2PA Content Credentials icon as pictured below:

Clicking on this icon will display the Content Credential and available metadata, such as:

- Content source: The origin of the content, such as the camera model used or AI tool used to generate all, or part, of the image.
- Issued by: The entity that created and signed the content credential, providing transparency about who is asserting the authenticity and provenance of the content. This could be an individual creator, an organization, or a trusted authority.
- Issued on: The date and time when the Content Credential was created and signed.

[See more here: Content credentials | LinkedIn Help; LinkedIn Adopts C2PA Standard]

---

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?"

Yes, we leverage direct disclosure based on the display of Content Credentials, as we believe providing rich context, when available, helps mitigate speculation about content. This could include, for instance, context about the base image used, modifications made with an AI tool (which may be minor or significant), or the AI system used. Showing content consumers the Content Credentials (CR) icon indicates that provenance information is available to explore. It can also be more helpful than just a high-level label (e.g., 'synthetic content'), which could be misleading, misinterpreted due to a lack of nuance, or unhelpful as more context is impacted by AI tools to some degree in ways that may or may not be significant or deceptive.

Further, since no indirect disclosure method is foolproof against removal or tampering, we allow users to explore these details and any caveats, such as the source that signed the metadata. This is more reliable than reading portions of the metadata (which may be inaccurate) and converting it to a label that's been removed from the context. Although Content Credentials are not fully resilient to manipulation, manipulation can be flagged to the content consumer. The C2PA requirement for Content Credentials to be cryptographically signed mitigates forgery.

To communicate uncertainty without furthering speculation, we can display caveats such as inconclusive or invalid Content Credentials to end users (e.g., if an image has been tampered with after signing and hasn't been resigned or in cases where the signer cannot be verified on our Content Integrity Check site). If the Content Credential is not valid (e.g., it was tampered with) or if the signer cannot be verified by LinkedIn, the Content Credentials will not be displayed in the user interface.

---

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).

Within the 'content labels' category, the technique of using an icon (e.g., CR) to indicate provenance information is available for exploration could be explicitly noted. In addition, our approach to mitigating synthetic content risks and supporting media integrity includes two aspects that are currently missing:

- Disclosure of not just the history of the content (e.g., how AI was used in the content creation process) but also the source, in a privacy-respecting way (e.g., noting the system used to create the media or signer — if they wish to tie an assertion to their identity).
- Complementary disclosure of provenance information for non-synthetic content.

---

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

Our Content Integrity Check site verifies when visual, auditory, or multimodal content has associated provenance information per the C2PA standard, with support for the same formats supported by the Content Credentials Verify site. LinkedIn is starting its implementation by supporting Content Credentials icon display on images and videos.

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.

Microsoft has applied direct disclosure for multiple products, including LinkedIn and Microsoft Office's Word and PowerPoint. For LinkedIn specifically, the challenge was deciding what content to disclose to audiences, as detailed by the C2PA technical standard and UX guidelines. LinkedIn began ramping up C2PA disclosure for a subset of images and videos uploaded with Content Credentials source and history information. As LinkedIn learned from this implementation and user feedback, the team continued iterating and ramping to a phase two release whereby Content Credentials are now displayed for all images and videos with cryptographically signed metadata uploaded to LinkedIn's feed. Information currently displayed in the user interface includes:

- Content source: The origin of the content, such as the camera model used or AI tool used to generate all or part of the image.
- Issued by: The entity that has created and signed the Content Credential, providing transparency about who is asserting the authenticity and provenance of the content. This could be an individual creator, an organization, or a trusted authority.
- Issued on: The date and time when the Content Credential was created and signed.

Content Credentials are currently a relatively unfamiliar technology to the average consumer, which can impact their efficacy. Microsoft's user research teams have conducted studies on Content Credentials, showing that very subtle changes in language (e.g., "certified" vs. "verified" vs. "signed by") can significantly impact the consumer's understanding of this disclosure mechanism. As such, LinkedIn is currently disclosing the information listed above, while recognizing that the language for such disclosure may need to be updated as user education expands and understanding changes. LinkedIn has also provided an FAQ page on Content Credentials and necessary definitions to help users better understand the nuances in provenance information.

In this direct disclosure scenario on LinkedIn, the "content source" and "issued on date" components were challenging to implement. For content source, and specifically for the disclosure of AI-generated content, LinkedIn initially decided to disclose the general use of AI in the generation of an image, whether that generation applies to all of the content (i.e., a fully AI-generated image) or part of it (e.g., a modification to an image made with an AI tool). The Microsoft Responsible AI Standard around disclosure is designed, in part, to be general and scalable across various product scenarios, with greater nuance applied to synthetic media considerations, such as when a piece of content was entirely generated, partially-generated, or edited by AI.

The Content Credentials details LinkedIn discloses via the user interface may be refined in the future as C2PA guidance changes and more media content is partially AI-generated. Similarly, timing of when the Content Credentials were issued was also in flux when LinkedIn started displaying Content Credentials. LinkedIn initially displayed the month and year

when the underlying certificate was issued, not when that content was signed with Content Credentials or when it was created.

The C2PA also recognized that timestamp disclosure was confusing to the public as the creation and C2PA signing times may not be exactly the same. Based on this feedback, C2PA updated its timestamp recommendations in the latest iteration of the technical standard, v2.1, and LinkedIn updated the timestamp accordingly with its phase two release.

---

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.

From a **Distributor** perspective, direct disclosure based on probabilistic detectors has been rejected to date. We've chosen to lean on C2PA for direct disclosure to wide audiences and use methods such as synthetic media detectors to help assess content when provenance metadata is lacking (as discussed in Section 2, Question 2).

We also had to weigh potential solutions and their trade-offs, when thinking about LinkedIn's implementation in particular. For LinkedIn, as we considered what Content Credential information to display to end users, we initially planned to include manifest information (i.e., metadata assertions about content provenance) and signer information (i.e., the party that cryptographically signed the Content Credential and its assertions). However, we carefully considered concerns that arbitrary information or false claims (such as who the media creator was) could be added to the manifest fields — for instance, if the signer was an untrusted party. As a result, we focused the user interface around signer information, as signer verification (i.e., verifying that a Content Credentials signer matches an entity on a C2PA trust list per a valid certificate authority) is what LinkedIn is equipped to do. The platform is not technically capable of verifying the validity of metadata assertions found within Content Credentials.

Soliciting and carefully following feedback from LinkedIn's first iteration, the team heard that more information would be helpful to users. As a result, we started including dynamic information that is not deemed high-risk in the user interface for our second iteration. While our first iteration notes if AI was used to generate part or all of an image, our second iteration provides more granular details for synthetic content: (1) AI was used to generate part of this image, (2) AI was used to generate all of this image, or (3) AI was used to edit this image.

User experience implementation guidelines were nascent when we first launched support for Content Credentials display on LinkedIn, making implementation challenging but also raising an opportunity to share learnings and open questions where C2PA guidance would be helpful for platforms. As a result, during our second iteration, the C2PA produced guidelines for UX implementation.

3. How was this piece/kind of content identified?	LinkedIn is currently only adding direct disclosure for user-generated content (UGC) that is uploaded and published with pre-existing C2PA Content Credentials. This means that if LinkedIn users publish media content that do not already have C2PA Content Credentials, LinkedIn currently will not add new Content Credentials. Thus, it serves only as a pass-through mechanism and display platform. In determining the information that is displayed, LinkedIn is implementing the C2PA specification and UX guidelines for Level 1 and Level 2 display. Level 1 information consists of an indication that C2PA data is present and its cryptographic validation status. Level 2 consists of a summary of C2PA data available for a given asset that provides enough information for the particular content, user, and context to allow the consumer to sufficiently understand how the asset came to its current state.
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?	The risks here include both potential reputational harm to the Microsoft and LinkedIn brand as well as potential societal harm if the UI implementation was incorrect or misleading or if attack vectors purposefully distort the information. One example of an attack is if someone screenshots a LinkedIn member's post with Content Credentials displayed, alters the Content Credentials information displayed within that screenshot using a tool that does not support C2PA, and then reposts that altered screenshot to a social media platform. It is difficult to prevent such situations from occurring, though efforts like increased media literacy campaigns, which Microsoft supports, can serve as helpful mitigation measures by encouraging content consumers to verify the original media and check the source.
5. What was the impact of implementing this disclosure? How did you assess such impact (studying users, via the press, civil society, community reactions, etc.)? Did the disclosure mechanism mitigate the harm described in the previous question (3.4)?	As LinkedIn is the first and only professional networking platform that has implemented direct disclosure of content provenance, the intended impact is to provide LinkedIn users with the proper tools to formulate their own opinion on whether to trust the media content they see. As for whether the disclosure mechanism mitigates the harm described previously, no, a purely C2PA Content Credentials solution does not mitigate the malicious scenario above. C2PA Content Credentials, like all disclosure methods, is not foolproof. A greater industry movement toward the development and adoption of complementary provenance and watermarking technologies would be helpful, and media literacy efforts (as described in Section 6) will be especially important in building societal resilience against deceptively manipulated media and provenance attacks.

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?

There are a couple of different actors in the content distribution pipeline that can take additional actions to help support LinkedIn's implementation of direct disclosure.

Software and hardware agents can provide insights into where the content was originally sourced — whether the camera platform that took a photo, AI generation model that created an image, or software platforms that provide creation and editing tools, such as Adobe Photoshop. As the C2PA provenance trust model requires broad industry adoption to have as much source content certified with Content Credentials as possible, actors that provide the content source have a responsibility to implement C2PA Content Credentials. Right now, since C2PA is only at the early stages of adoption in the industry, the number of actors that have implemented Content Credentials to certify content source and origin is limited.

These other C2PA implementers can also help by correctly implementing both the required and recommended components of the C2PA standard. While the C2PA standard was designed to be interoperable, there could still be differences in implementation for certain components, such as in the edit assertions (i.e., what gets included in the Content Credentials and how it is displayed to end users). It would be helpful if implementers could cross-test their implementations with each other and with downstream displays on LinkedIn before release to ensure that the direct disclosure mechanism accurately reflects the provenance information.

Consumers and organizations that use the software and hardware agents described above could also share the responsibility of supporting LinkedIn's direct disclosure. As there are some platforms that already support C2PA Content Credentials, consumers and organizations could actively use these platforms to start the documentation of Content Credentials. For AI generation models this includes the use of OpenAI, Adobe Firefly, or Microsoft AI generation tools, such as Bing Image Creator or Azure OpenAI. For editing tools, this includes Adobe Photoshop and Microsoft's Designer or Paint tools. For camera platforms, this includes various models built by Sony, Leica, Nikon, and Canon. If consumers use software and hardware agents implementing the C2PA spec with the goal of publishing content on LinkedIn, this will allow for more content to be shared with direct disclosure attached to it.

Microsoft's and LinkedIn's goal is to contribute to industry adoption of direct disclosure of content origin and history. If software and hardware agents and consumers can proactively implement content provenance and use platforms that already support provenance, this could accelerate industry adoption and user education of direct disclosure mechanisms.

Finally, the C2PA itself should further evolve its user experience guidance to meet the needs of diverse creators and display platforms. With LinkedIn, we found many assertions might show up in a Content Credential, making it challenging to determine which ones should be shown and how. The current C2PA user experience guidelines are written with sparse information regarding what should show up at the L1 (top-level provenance details) and L2 (second level of granularity). Further, this might not match what LinkedIn would want to display per its users' needs. C2PA assertions could enable content creators to specify

what would be most helpful to display at the L1 or L2 level, per their use case. However, this would require a change to the C2PA specification. We provide an illustration of how such information may vary by use case below:

- **A content creator might want to include their social network handle, website, or other attribution data.** This is not included in current guidance for what should be displayed at the L1 or L2 level, so while a content creator might include this in their Content Credentials and want it to show up, it will not be displayed on platforms following the current UX guidelines. (To see such information, the media would need to be run through a C2PA verification site like Content Integrity Check or Verify to see the complete set of metadata associated with the media at all levels).
- **An advertiser might want to include their brand name at the L1 level to assert this information as top priority for their use case** — e.g., to assert to content consumers that a piece of media is truly associated with their brand and not a misrepresentation. This would be the case, for instance, if the signer of the Content Credentials was not the brand itself, but rather an advertising company. Such information would also not be displayed per current guidance.

We also think, based on our learnings, that it would be helpful for the C2PA to enable differentiating claims and verifications at the UI level. As alluded to in Section 3, Question 2, because LinkedIn is a professional network, we only validate the certificate of the party who signed the content, not the manifest itself.

If a signer's certificate has been validated, there's an implicit trust that the signer will include accurate information in the manifest; otherwise, their certificate may be revoked per C2PA directions under version 2.1 of the spec and trust list maintenance. If the signer included false information in the manifest or a malicious actor manipulates the media, making claims in an earlier manifest no longer valid, the situation would be beyond LinkedIn's control. Further, if the manipulations in question are occurring at the tool level by an abusive user, it would not make sense to penalize the company or tool that signed the media.

This will remain a gap even with revocations of signing permissions for actors abusing policies as trust lists are maintained. As such, it would be helpful at the UI level to enable social networks to make clear what they are vouching for (i.e., the signer was validated to appear on a list of trusted actors per its valid signing certificate) and what the signer is vouching for (i.e., the metadata information or assertions included in the manifest itself).

As we think about supporting a growing number of display needs, LinkedIn (among other display platforms) would benefit from others with a broad range of industry vertical, content creation, and distribution perspectives weighing in via the C2PA on what information should be displayed and how. As the number of entities implementing the C2PA standard grows, from tool builders to display platforms, and across verticals from news organizations to advertisers, we expect continued UX learnings and improvements — considering the needs of diverse content creators and consumers.

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

We recognize the value of as much UX feedback as possible. Looking back, we would have gathered more UX feedback from others around the company and in the C2PA.

Post-launch UI/UX for display is under constant consideration and evaluation. LinkedIn intends to respond to the results of its initial implementations to develop better data presentations for users over time, using the Check/Verify sites for forensic information and to provide feedback on implementation best practices to the broadder industry.

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?

One of Microsoft's long standing principles in responsible AI is transparency operationalized through Microsoft's Responsible AI Standard, which is a key policy instrument our organization uses to guide responsible AI in product development.

The third transparency goal of the Microsoft Responsible AI Standard, "Disclosure of AI interaction," speaks concretely to practices for direct disclosure by AI systems and helps ground our approach to disclosure in various AI system product scenarios, including novel GenAI systems and applications.

"Microsoft AI systems are designed to inform people that they are interacting with an AI system or using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic" — Goal T3: Disclosure of AI interaction.

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?

We believe this example illustrates the need for more sustained UX research and focus on how content should be presented on social or professional networks. Clear guidelines on how Content Credentials should be presented across the content ecosystem would be helpful, featuring one consistent design that reinforces what Content Credentials contain and what they do and do not indicate to users.

This example also demonstrates the value of providing more information to users in the manifest, rather than just a label lacking details so content consumers can make better decisions about what media they trust.

This case also illustrates a best practice of LinkedIn displaying Content Credentials for both authentic media (i.e., captured by a camera or recording device) and synthetic media (i.e., AI-generated or -modified). We believe greater encouragement from practitioners and policymakers to display credentials for authentic media would benefit the whole ecosystem; otherwise, the ecosystem risks over-indexing what AI content is and losing focus on what is authentic.

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)?

Microsoft's understanding of direct disclosure has been informed by work including:

- User research contributions from C2PA members informing the C2PA spec
- Insights on provenance from researchers at Microsoft (e.g., AMP | Proceedings of the 12th ACM Multimedia Systems Conference, Designing Media Provenance Indicators to Combat Fake Media | Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses)
- Research on the limitations of various media integrity technologies, including from a security lens. For instance, our own research, on attacks that are feasible when access to watermark detectors is made broadly available for detection has informed our caution around relying on these detectors for disclosure at scale.
- Considerations surfaced in other published guidance, such as PAI's Responsible Practices for Synthetic Media Framework and Glossary for Synthetic Media Transparency Methods.

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?

We generally don't see a risk in over-disclosure as long as the disclosures are accurate, and are provided to users in a format that is consumable and helpful based on the surface in which they are interfacing with it. We generally prefer to over index on accuracy over volume, allowing content consumers to be skeptical about media lacking verified provenance information rather than being misled by labels of varying accuracy applied to all content. This has helped inform our approach whereby we display Content Credentials and contribute to ongoing research on how to best signal warnings about unverified information (per the UX research links above); how we provide L1 and L2 information in social media feeds; and how to provide a complementary tool for forensic information including more granular L3 details.

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.*

For the reasons noted above, we generally err on the side of over-labeling when we apply provenance information as a **Builder** — but we remain cautious about both the content and form of what is displayed on our distribution platform and Content Integrity Check site. Conditions like election years with high stakes events have motivated us to invest heavily in and call for more media literacy, multistakeholder efforts to share abuse signals, and forensics efforts to disclose secondary and tertiary provenance signals to select parties who can take action on or share them more broadly when needed.

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 generally agree provenance information is beneficial for all content. As a practical matter, as a **Builder** and **Distributor**, we aren't well-positioned to determine what is considered a creative use when it comes to system output or media publishing. Thus, we label, for instance, all images produced by our GenAI systems, such as Dall-E in AOAI, Bing Image Creator, and Paint. We also display provenance information for images and videos containing C2PA metadata ingested by LinkedIn.

Considerations of the creative community highlight the importance of **how** provenance information is disclosed. Providing rich context where possible (e.g., L1 and L2 Content Credentials details on a social media platform to L4 details in a forensics tool), rather than applying a high-level label based on one metadata field pulled from the Content Credentials, can help mitigate unintended consequences such as:

- Potential stigma of having one's work labeled or credited as being made by AI rather than acknowledging human vs. system contributions (which may lead a creative wanting to strip indicators or evade a given label).
- Content consumers being misled regarding the degree to which AI played a role in the content (i.e., not understanding the degree to which AI modifications were minor or significant).

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 are aligned with the PAI Framework on roles these parties play.

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. Aligning on a threshold for direct disclosure will be important so that provenance information added to media will not just be displayed for synthetic content, but also non-synthetic authentic content (e.g., authentically captured by a recording device).
2. A shared visual language for disclosure is also important and should be informed by future user research at a national and international level. Alignment on how to represent the presence of provenance information and warnings/disclaimers to limit false assurances will be necessary.

5 Approaches to Direct Disclosure, in Policy and Practice

1. What does your organization believe are the most significant sociotechnical 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)

Key sociotechnical challenges include understanding whether disclosure is meaningful to content consumers and accurately interpreted.

While measuring technical robustness will be important, assessing societal resilience against deceptive AI content will be just as critical. Given the array of disclosure, verification, and detection methods that exist, it will be important to (1) educate the public on what labels or disclosures mean and do not mean (for instance, disclosure of origin or history is neither considered disclosure nor verification of veracity), and (2) understand how content consumers interpret signals that media is AI-generated or edited (including whether individuals are accurately interpreting labels or disclosures, to what extent this information is trusted, and whether trust varies depending on the actor that's labeling content).

Similar to sociotechnical literacy around security concepts like phishing prevention or SSL website security, there is a need for public tech literacy campaigns focused on provenance and various disclosure methods. Likewise, for detection tools, it will be important to understand the impacts of error rates, variations in practice for false positive and false negative thresholds, and how these collectively impact users' trust in the results that detection tools convey.

We recommended to NIST that evaluation infrastructure will be essential to study how provenance and watermarking technologies are being used, to what extent the technologies are understood and trusted when they are used as intended or misused/abused, and how trust in labels varies depending on the entity labeling content as AI-generated or edited.

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

See Section 5, Question 1, for more information.

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

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

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

1. Accuracy

Direct disclosure can impact the accuracy of perceptions when provenance information is reliably marked and accurately displayed.

2. Trustworthiness

Direct disclosure of both the source and history of content can empower consumers to determine the level of trust they have in the content.

3. Authenticity

Verifying and disclosing content authenticity can mitigate harms (e.g., potential deception and supporting incident response) and lead to better informed decision-making (e.g., per number two above).

4. Harm mitigation

We cover this in "Authenticity" above.

5. Informed decision-making

We cover this in "Authenticity" above.

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?

We recognize the implications of the liar's dividend and the extent to which even well-intentioned implementations of disclosure methods may backfire, increasing the impact of the liar's dividend.

While we believe providing transparency about media provenance will help build societal resilience against deceptive AI-generated content, we also understand that no disclosure method for AI-generated content is perfect and all are subject to attacks. These attacks include bad actors removing provenance information from AI-generated content to deceive the public into thinking it is authentic, as well as forging watermarks to mark authentic content as AI-generated. It will be critical to continually assess and improve the efficacy of disclosure approaches for AI-generated and manipulated content, to ensure that the transparency they offer is meaningful to content consumers, and that the capabilities and limitations of these approaches are well understood by the public. These measures will help guard against public distrust of digital content and individuals dismissing authentic content as manipulated. This could have grave consequences for our economy, courtrooms, the state of elections, and even national and global security.

To better sense if we are reaching a tipping point (i.e., if we are mitigating or exacerbating the issue), it will be important to assess and measure whether disclosures are meaningful to content consumers and if they are accurately applied and interpreted and trusted per the comments in Section 5 Question 1.

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?	What is the source of the content? And do I trust this source?
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.	See Section 5 Question 1 for more information.

Similar to sociotechnical literacy around security concepts like phishing prevention or SSL website security, there is a need for public tech literacy campaigns focused on provenance and various disclosure methods.

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 agree. We have been working on broader public education on synthetic media via our Societal Resilience Fund and initiatives run by our Democracy Forward team.

- In May of 2024, we launched a USD $2 million Societal Resilience Fund to further AI education and literacy among voters and vulnerable communities. We believe it is more important than ever to provide tools and information that will help people navigate an increasingly complex digital ecosystem and find authoritative resources, and that we have a responsibility to partner with a diverse set of global civil society organizations and academics on these efforts.
- Grants delivered from the fund will help several organizations, including Older Adults Technology Services from AARP (OATS), the Coalition for Content Provenance and Authenticity (C2PA), International Institute for Democracy and Electoral Assistance (International IDEA), and Partnership on AI (PAI) to deliver AI education and support their work to create better understanding of AI capabilities.
- More details on the above grants are provided here: Microsoft and OpenAI launch Societal Resilience Fund.

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?

As outlined in our whitepaper, Protecting the Public from Abusive AI-Generated Content, we think collaboration among a broad array of stakeholder groups is necessary to improve public understanding of synthetic media.

- **The Federal government** should publish and update best practices and standards for helping the public understand how to navigate synthetic content and lead a national research program to study synthetic content harms and media provenance.
  - NIST has a key role to play in collaboration with other agencies to leverage AISI findings to foster media literacy so citizens learn about the risks of synthetic content and tools to better protect themselves from being manipulated or deceived when such content is misused. This would help ensure citizens are critical content consumers and that, as provenance and other complementary disclosure methods are deployed at scale, they are easily digested and comprehended. NIST, through its AI Safety Institute and Consortium (AISIC), has already begun assessing best practices for synthetic content labeling, verification, and detection. It will be important to update these best practices annually as the sophistication and complexities of synthetic content increase, methods and tools for labeling and detection progress, adversarial attacks to deceive the public

about provenance evolve, and the public's understanding on labeling and detection approaches grows.

    - NSF, in partnership with AISIC, should lead a national research program to explore existing and emergent synthetic content harms, building an evidence base of where harms are manifesting and assessing how to best measure and mitigate them. In addition to harms directly related to synthetic content, this should include core methods, designs and signals for consumers, and an assessment of harms resulting from loss of trust in authentic content. Research should also assess how well tools for labeling and detecting synthetic content and display provenance are working in practice, and include a sociotechnical analysis of how they are used and perceived. Evidence on how well authenticity and provenance infrastructure is working in practice should inform ongoing public education campaigns and best practices for synthetic content disclosure.

- **Academia** should play an essential role advancing research on harms and best practices for the research program above.
- **Federal and state governments** should fund and create programs that educate the public about deceptive uses of synthetic content that present safety risks and harms, as well as approaches they can use to discern among digital content. This includes how to assess signals about whether content was authentically captured, or AI-generated or manipulated, what signals and tools can be used to see if it came from a source the content consumer trusts, as well as recommended practices to address the latest scams employing synthetic content.
    - Education campaigns can also target vulnerable demographics such as older adults and young people. According to the AARP, when it comes to new technology, most older adults are later adopters and have lower confidence in their digital literacy. Similarly, Microsoft's research, conducted in partnership with National 4H, shows that 72 percent of young people seek support from adults in learning how to use AI tools correctly and with confidence.

- **Civil society groups & news organizations** should inform education campaign content and play a role in disseminating it in local communities. Frontline actors, including local media and journalists, community leaders, and civil liberties and human rights groups should also benefit from these materials and forensics tools, to assess potential deepfakes and educate others as part of their work.

We believe legislation and Congressional funding would be helpful to achieve the objectives above.

3. What support does your organization need in order to advance synthetic media literacy and public education on evaluating media?	Support as outlined in Section 6, question 2 above would be helpful to meaningfully advance synthetic media literacy.