PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How Adobe designed its Firefly generative AI model with transparency and disclosure

Adobe

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

1 Organizational Background

A contextual introduction to the case study.

ADOBE'S RESPONSE

Since our founding in 1982, Adobe has pioneered multiple innovative technologies, bringing the world tools like Photoshop, PDF, Illustrator, and many others. With over 30,000 employees around the globe, we have been incorporating AI into our products for more than a decade to help creators and others realize their creative potential while respecting our customers. Adobe is uniquely positioned to address the recommendations around synthetic media in Partnership on AI's (PAI) Synthetic Media Framework. We are a **Builder** of AI technologies that allows anyone to create content, and we enjoy a rich tradition and trust within the creative community, with millions of creators around the world using our tools every day.

In March 2023, we introduced Adobe Firefly, our family of creative generative AI models designed for safe commercial use. Since launch, it has been used to generate nearly 6 billion assets.

As we developed our first Firefly model, there were several goals and principles we wanted to implement: we wanted it to be commercially safe, provide transparency to consumers, and respect the rights of artists and creators.

With our well-defined AI Ethics principles on transparency, responsibility, accountability, and the ongoing efforts around implementing Content Credentials — signals that allow consumers of content to understand the origins and changes made to digital files — we decided early on that every asset produced with Firefly would have an embedded Content Credential indicating that the file was created using AI, the model used to create it, and its version.

This decision was significant as it builds on our mission to ensure that AI tools like Firefly are used responsibly, giving viewers of digital content important context to help them understand what they are seeing and allowing them to make their own decisions about whether to trust the content.

In 2019, Adobe co-founded the Content Authenticity Initiative (CAI), a community of media and tech companies, NGOs, academics, and others working to promote adoption of an open industry standard for content authenticity and provenance, with the goal of restoring trust and transparency in digital content.

The CAI creates open-source tools based on an open technical standard developed by the Coalition for Content Provenance and Authenticity (C2PA) — which we helped found — called Content Credentials. Today, the CAI has more than 2,500 members globally across a variety of industries.

As we developed our first Firefly model, there were several goals and principles we wanted to implement: we wanted it to be commercially safe, provide transparency to consumers, and respect the rights of artists and creators.

2 Challenge

Elaborate on the challenge being addressed in the case study, i.e. the issue to which your organization is applying the Framework.

ADOBE'S RESPONSE

One of the main challenges with developing Firefly was how to provide transparency about the content created by our generative AI models.

As we have seen in the past year, generative AI has the powerful ability to create new content in seconds using just a few keystrokes. It is transforming the way we work, create, and communicate. For example, generative AI allows you to generate convincing synthetic images of political leaders, celebrities, and other imagined scenes almost instantly. Across all types of generative AI content, it is becoming increasingly difficult to distinguish between fact and fiction. While we are optimistic about the technology and encourage its use, we also wanted to provide the tools to ensure the transparency needed to help create trust.

We faced several challenges while building Firefly models that would meet our technical, legal, policy, and ethical standards, one of them being our desire to attach Content Credentials to Firefly-generated files in both web- and app-based environments.

Another challenge was how to develop a useful AI tool, while training it on a curated dataset. Adobe's first model was trained on Adobe Stock images, openly licensed content, and public domain content where copyright has expired. Training on curated, diverse datasets inherently gives your model a competitive edge when it comes to producing commercially safe and ethical results. Our team was able to strike the balance between developing this technology responsibly and still delivering a tool people would want to use.

Fortunately, we had several advantages:

1. **Company Buy-in:** In keeping with the PAI Framework and our own AI Ethics program, our leadership and product teams quickly made the decision that Content Credentials should be attached to content generated by Firefly.
2. **Existing Work on Content Credentials:** Adobe has been working on content provenance tools, like Content Credentials, since 2019, well before Firefly and other generative AI systems were made publicly available.
3. **Existing Work on Adobe Stock and Content Credentials:** Since 2022, all images downloaded from Adobe Stock have had Content Credentials automatically attached upon download to indicate that they were licensed from Adobe.

> *Across all types of generative AI content, it is becoming increasingly difficult to distinguish between fact and fiction.*

3 Objective

Describe what your organization is attempting to accomplish by addressing this challenge and/or furthering the opportunities.

ADOBE'S RESPONSE

We laid out our goals to address these various challenges around harm and risk through our own AI ethics principles, as well as signing on to policy efforts like the White House Voluntary AI commitments and PAI's Framework. By attaching disclosure mechanisms, as we have done with Firefly and Content Credentials, we see at least four objectives that we can accomplish:

- **Creating Trust by Providing Transparency:** By showing which generative AI tools have been used in the creative process, Content Credentials can show when something was created or has been edited with AI, and prove when something was human-created or captured in the physical world by a recording device. This allows responsible actors a means to create trust in their work with generative AI tools.
- **Mitigate Real World Harm:** Because generative AI tools can create realistic images capable of misrepresenting events or inventing scenes which never existed, the potential for misuse by bad actors is real and can influence perceptions of current events, political activities, and other areas where accuracy is important.
- **Protecting Against Reputational Risk:** Generative AI content has the potential to portray individuals, companies, and others in a negative or misleading light, as highlighted in Appendix B of the Framework, which could have serious consequences. Many of the creators and brands that Adobe partners with have expressed concerns that serious consequences such as "fake reviews, coordinated campaigns of misinformation on social channels, or the brand safety risk of ads appearing alongside malicious fake news generated at scale." (Source) That is why partners such as Publicis Groupe joined the Content Authenticity Initiative and are deploying Content Credentials to ensure brand authenticity throughout the content lifecycle.
- **Creator Rights:** Leveraging the Content Credentials technology, Adobe is working to enable and protect creators by attaching a "Do Not Train" tag to the metadata of their work so they can keep their content out of AI training databases, which are used to refine the output of generative AI engines. We will work to drive adoption of an industry standard for this technology. Notably, this is a form of disclosure/transparency that affects the **model development** for synthetic media, not necessarily transparency about the eventual artifacts that get **produced** using generative AI.

4 Framework Scope and Application

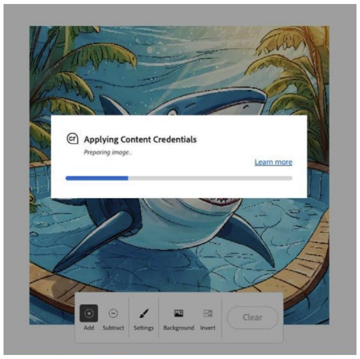
Identify which Framework principle was used to help address the challenge/ opportunity, how it was chosen and implemented, and describe how it was applied.

ADOBE’S RESPONSE

The primary PAI Framework principle we implemented was to “Take steps to provide disclosure mechanisms for those creating and distributing synthetic media.” We have done this by providing both Direct and Indirect Disclosure mechanisms.

As referenced in PAI’s Framework recommendation on Indirect Disclosures, we automatically attach cryptographic provenance data to Firefly output, consistent with the C2PA standard.

We also utilize Direct Disclosure methods. When a user creates and downloads an asset in Firefly, a notification appears to confirm that Content Credentials have been attached, letting people know that a generative AI tool has been used and which model. Content Credentials are then attached once the image is downloaded. The image can then be dropped into Adobe’s inspection tool, “Verify,” which will display this provenance information. Using open source tools, publishers are able to add an interactive Content Credentials icon to appear within the digital content, allowing the viewer to see this provenance information with a simple mouse click.

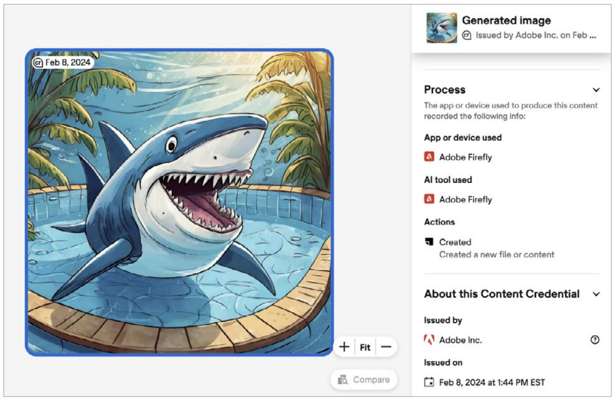


Applying Content Credentials

Preparing image..

IMAGE 1 (left).
A screenshot depicting the notification users get when Firefly is attaching Content Credentials to their content.

IMAGE 2 (below).
A screenshot depicting the provenance information users can see when using Adobe’s inspection tool.



Feb 8, 2024

Generated image

Issued by Adobe Inc. on Feb ...

Process

The app or device used to produce this content recorded the following info:

App or device used

Adobe Firefly

AI tool used

Adobe Firefly

Actions

Created

Created a new file or content

About this Content Credential

Issued by

Adobe Inc.

Issued on

Feb 8, 2024 at 1:44 PM EST

We have seen a lot of momentum around the need for transparency in AI-generated content. So far, many conversations have focused on invisible watermarking, which is important and included in the C2PA open standard. Imperceptible watermarking and cryptographically signed metadata are complementary and strong united measures for ensuring objective understanding of where content originates. As AI becomes more prevalent in the content we consume, we believe secure, signed metadata about how an image came to be will be critical to allow people to see content with context.

We also implemented PAI Framework principles around “providing transparency to users” and to “provide a published, accessible policy outlining the ethical use of your technologies and use restrictions that users will be expected to adhere to and providers seek to enforce.” All users of Firefly agree to Adobe’s Generative AI Additional Terms, which govern their use of generative AI features in our services and software and explicitly require that users “must not remove or alter any watermarks or Content Authenticity Initiative metadata (e.g., Content Credentials) that may be generated with the output, or otherwise attempt to mislead others about the origin of the output.”

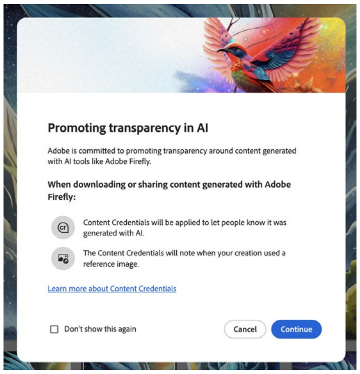
Our Generative AI User Guidelines also expressly prohibit the “dissemination of misleading, fraudulent, or deceptive content that could lead to real-world harm.” The Guidelines also explicitly state that “we may take action on your Adobe account if we discover content or behavior that violates these Guidelines.”

5 Obstacles

Elaborate on any internal or external obstacles intrinsic to the Framework that were overcome.

ADOBE'S RESPONSE

Internal feedback suggested that one of the obstacles we would face was external awareness and acceptance, which we hope to help overcome through our artist series (see further below). For many, their first encounter with provenance information in the form of Content Credentials is when they create an image in Firefly and download it. This underscores the importance of guiding principles like PAI's Framework, and our own transparency efforts. We want to clearly let our users know that we feel strongly about attaching provenance data to these images.


Promoting transparency in AI

Adobe is committed to promoting transparency around content generated with AI tools like Adobe Firefly.

When downloading or sharing content generated with Adobe Firefly:

- Content Credentials will be applied to let people know it was generated with AI.
- The Content Credentials will note when your creation used a reference image.

Learn more about Content Credentials

□ Don't show this again

IMAGE 3.
A screenshot depicting the notification users get when they create and download a Firefly image.

Another obstacle has been to fulfill the Framework's call to, "Aim to disclose in a manner that mitigates speculation about content, strives toward resilience." Since even cryptographically verifiable metadata can be removed by malicious parties, Firefly disclosure now leverages Adobe's Content Credentials Cloud to store the provenance data (not the image) so it can be looked up through fingerprinting. In the next iteration of Firefly image generation, a high-resilience steganographic watermark will also be used to ensure permanent binding between the metadata and the asset. This combination or "triad" of signed metadata, watermarking, and fingerprinting provide a combined utility that is far better than any of the techniques alone. Critically, this approach can be applied equally to images, video, and audio.

One area that we, and PAI's Framework, will have to address further is how to accurately create a meaningful and comprehensive disclosure, since the majority of content — even that which does not mislead or cause harm — will likely have some AI elements in it going forward. Simply labeling something as AI doesn't consider the many nuanced ways AI could be used.

Another area PAI should continue to address is the complex and diverse ecosystem and the value of interoperability in order to implement C2PA-based provenance standards in a uniform way across a diverse range of sectors and platforms.

We want to clearly let our users know that we feel strongly about attaching provenance data to these images.

6 Benefits

Identify the opportunities created for your organization by utilizing the Framework to address the challenge.

ADOBE'S RESPONSE

The Framework has provided guidance and clarity for our internal teams and leadership. As we further develop Firefly and push into other types of content generation, disclosure methods like Content Credentials will be required.

It has also led to transparency for Adobe customers. By providing C2PA-based transparency technologies, Firefly images will provide responsible users with a way to build trust while still allowing them to use AI to add quality and clarity to their work more easily and efficiently than ever before. By adding this level of transparency, people can make informed decisions about trusting any AI-edited content they are consuming.

Implementing the Framework has also benefited our creative community. As mentioned above, while Adobe primarily operates as a developer of these technologies and tools, we have a deep connection to the creators that use our products. In addition to maintaining transparency, attaching Content Credentials to generative AI content also allows artists and creatives the ability to retain attribution for their work. We have published several artists series on how they are incorporating generative AI and Content Credentials in their work. Additionally, they can use CAI provenance technology and attach "Do Not Train" credentials that travel with their content. With this universal standard now published in the latest C2PA technical specification, we believe this will be adopted quickly across industries.

We see this case study being instructive and informative for policymakers. As policy debates continue at the Federal, State, Local, and International levels, it will be a resource to educate policy makers on a practical, real-world deployment of attaching provenance data to generative AI images. For example, one of the policy issues that Adobe has engaged on has been the issue of disclosure of generative AI on election ads. Our case study could be instructive on how policymakers can consider Content Credentials.

Firefly images will provide responsible users with a way to build trust while still allowing them to use AI to add quality and clarity to their work.

7 Conclusion/Key Takeaways

A description of how implementing the Framework ended for your organization, including any lessons learned.

ADOBE'S RESPONSE

At Adobe, we have seen the benefit of PAI's Framework. In addition to having an internal AI ethics program and review process, it underscores the importance of having leaders and decision makers who care about the responsible use of AI technology.

While we have made incredible progress in four short years, a lot of work remains. For a solution like Content Credentials to work, we need this technology everywhere — from the devices used to capture content to social media platforms and other areas where many users share and consume content.

Currently, if an individual creates a piece of content and attaches Content Credentials, a social media platform might strip out the credential before it is uploaded to their site.

Broadly, we define success as ubiquity. The diverse community of CAI members can ensure that Content Credentials become the industry standard norm and are visible everywhere across the digital ecosystem. Their absence will become an indicator that something deserves a second look to determine its provenance. Alongside ubiquity, we have to prepare for the ways in which disclosure mechanisms might play a role in the liar's dividend, where the prevalence of disclosed deepfakes makes it more likely that bad actors can claim real and authentic content is synthetic. This actually helps reinforce the need for universal adoption of Content Credentials. For example, if platforms that distribute content can clearly identify that which is synthetic, it introduces a level of friction necessary to prevent malicious synthetic content from defrauding consumers, and if applied accurately, can validate that authentic content is indeed authentic. As a result, consumers will become more discerning and skeptical, which might drive more responsible actors to deploy and use technologies like Content Credentials in the long-term. User education is equally important as consumers will also need to become more aware of the provenance technologies that exist, how they work, and what their limitations are.