PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How OpenAI is building disclosure into every DALL·E image

OpenAI

This is OpenAI'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.

OPENAI'S RESPONSE

OpenAI is an AI research and deployment firm based in San Francisco, CA. We deploy general purpose AI tools for ChatGPT and API customers that serve 100M+ users in most countries throughout the world. Per Partnership on AI's (PAI) Synthetic Media Framework, we fall under the "**Builder of Technology and Infrastructure**" category and develop generative AI tools across text, image, and audio.

This case outlines OpenAI's initial exploration of image provenance.

OpenAI has long been a proponent of appropriate disclosure of the use of AI, as evidenced by the Sharing and Publication Policy we launched several years ago. OpenAI is researching and exploring a variety of provenance techniques, such as an experimental classifier that we released in January 2023, to help determine if text was generated by OpenAI models. In 2022, we hired a researcher to focus on researching text watermarking. In 2023, we commissioned a public opinion poll that surveyed 18,000 respondents in 9 countries to better understand their perspectives on disclosure and detection of AI-generated content across a range of mediums and contexts.

Our AI image generation tool, DALL·E, was first released in January 2021. The most recent version, DALL·E 3, was released in ChatGPT, our first-party product, in September 2023 and produces high-quality images. Several safety risks increase as image models produce more photorealistic imagery, including risks of visual misinformation, sexual content involving minors, and hateful imagery (though these risks are not exclusive to photorealistic content). OpenAI recognizes the need for image provenance to reduce misuse of image generation models and has undertaken work to explore and develop tools to that end.

Media provenance, particularly AI-generated audiovisual media provenance (note: what PAI has described as indirect disclosure), has been a topic of great discussion and debate across academia, media, industry, and policymakers. In July 2023, the White House released a set of voluntary commitments onto which several AI industry participants, including OpenAI, signed. One such commitment is that signatories "Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content." These commitments further underscore the importance of media provenance.

2 Challenge

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

OPENAI'S RESPONSE

We began seriously discussing image provenance in early 2023. The primary challenges we faced were consideration of:

1. Which provenance method(s) to explore and/or pursue initially, balancing various goals and trade-offs.
2. Whether to provide access to any provenance tools to actors outside of OpenAI, and if so, whom.
3. How best to communicate our initial image provenance approach, given significant societal and policymaker expectations — and some confusion — around the potential of image provenance.

Several companies are exploring various image provenance techniques, each with different objectives in mind. This area has garnered notable attention from both the media and policymakers. The effectiveness of different provenance techniques varies depending on the intended purpose. We have considered three provenance techniques: metadata, steganographic watermarks, and classifiers. For PAI's description of some of these techniques, see here.

Metadata approaches, such as IPTC and C2PA, have the benefit, ostensibly, of providing public or consumer visibility into content provenance as the metadata should

be accessible to anyone coming across these images. These approaches offer the added advantage of not requiring significant resources to implement. The IPTC and C2PA, individually, have been adopted by several industry participants. In practice, however, current metadata approaches are easily evaded, even by unsophisticated actors; metadata can be removed by simply taking a screenshot of an image or by downloading the image, erasing the metadata, and then re-uploading it. Further, the utility of metadata to an end consumer depends greatly on whether the browser or platform through which an individual is accessing the image has adopted the specification. Adoption of metadata approaches by browsers and platforms, while increasing, is still limited.

We see similar concerns with cryptographic watermarking today: it can be relatively easily evaded, particularly by motivated adversarial actors. We understand research in these areas continues to advance, and watermarking approaches may improve in the future.

Given the challenges associated with these provenance methods, we then considered building a classifier. This could potentially provide OpenAI with confidence as to whether an image was generated by DALL·E 3, marking an initial foray into image provenance. Such a classifier could provide the benefit of durability to the types of simple modifications — e.g., screenshotting, cropping — that can undermine metadata approaches. We announced the provenance classifier on October 19, 2023. In early internal evaluations, the classifier was over 99% accurate at identifying whether an image was generated by DALL·E when the image had not been modified. It maintained over 95% accuracy when the image had been subject to common types of modifications, such as cropping, resizing, JPEG compression, or when text or cutouts from real images were superimposed onto small portions of the generated image.

A notable downside to the provenance classifier, however, is that it does not inherently provide consumer or public visibility into image provenance. Rather, OpenAI would manage and determine access. Trade-offs associated with this are discussed further below.

3 Objective

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

OPENAI'S RESPONSE

Early on, OpenAI's Trust & Safety team articulated the following goals with regard to pursuing image provenance, several of which align with the harms identified in Appendix B of PAI's Framework:

- Minimize harms potentially caused by images generated by OpenAI's tools, including those around mis- and disinformation, sexual content involving minors, and hateful imagery.
- Empower the broader public with greater context on images they encounter that were generated by OpenAI's products, thus contributing to AI literacy broadly.
- Maintain parity with — if not exceed — industry partners' commitments in this space, furthering OpenAI's position as a leader in ethical deployment of AI.

OpenAI's DALL·E research team is primarily focused on a provenance tool that would be highly accurate and relatively durable to adversarial activity. A tool of low accuracy or one which could be easily evaded would ultimately not serve any of the broader goals effectively.

In January 2023, OpenAI released an AI classifier intended to distinguish between text written by a human and text written by AIs from a variety of providers (see blog post). The decision to release this classifier was motivated by many groups — including but not limited to educators — requesting clarity about methods for detection of AI-generated text. At the time, other text-based classifier solutions were circulating without a clear articulation of the limitations and accuracy of these tools for text specifically, such as unreliable performance on non-English languages, shorter texts, and susceptibility to inaccuracies due to minor edits in the text. The goal was to release a tool, though imperfect, that would contextualize the limitations of this approach.

In July 2023, OpenAI made the decision to take down the AI written text classifier, due to feedback about its low rate of accuracy and concerns about reliance on the tool to make consequential decisions despite the clearly stated limitations. While PAI's Framework is focused on audiovisual content, we believe that the learnings from

this case study of text provenance can still apply. A key lesson from this was that it is important to take into consideration how limitations of provenance approaches may not be considered by end users and may ultimately lead to false confidence or inaccurate conclusions.

As we have learned more about and further discussed the merits and drawbacks of various image provenance approaches, we found that meeting all of our goals was likely impossible given the state of current provenance techniques. We have opted to build, initially, a provenance tool that prioritizes accuracy and durability to editing or stripping out, as without these our other goals would be negatively impacted. We believe that a provenance classifier will help us progress in minimizing the potential harms caused by images generated using OpenAI's tools. It will facilitate a better understanding of how these images are used or disseminated, and for what purposes. This knowledge may allow us to better understand abusive behavior and determine if additional model- or system-level mitigations may be effective in reducing the potential for abusive activity.

Given the high expectations society and policymakers have of image provenance, along with the prominence of OpenAI's leadership in the AI research and deployment space, we have been highly cognizant that any provenance approach we undertake, even if an initial exploration not necessarily indicative of our long-term approach, is likely to be of great public interest. We believe there is significant risk in undertaking an approach that breeds a false sense of confidence on the part of policymakers, media, or consumers in the durability and/or utility of the provenance method. We believe the most responsible immediate path is to advance cautiously and deliberately into the image provenance space, until such time when provenance methods reach an inflection point where the benefits provided by the method in question are significantly greater than the risk of over-confidence. How we will know when we have hit that inflection point is an open question. At minimum, this will include decreased potential for provenance methods to be evaded.

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.

OPENAI'S RESPONSE

The entirety of this case study centers on the principle of disclosure, and how best to pursue disclosure as a builder of synthetic media tools. As we do not provide a platform for content distribution, all of our direct users are aware the content produced by our tools is AI-generated. The concern, however, is that content generated by our platforms may be later distributed on other platforms in potentially malign or covert ways, such as in disinformation campaigns, one of the harms described in Appendix B of PAI's Framework.

As laid out above, we chose to explore a provenance option that prioritized durability to adversarial activity, acknowledging that this trades off against openness and accessibility. We are still weighing the various access strategy options, and may soon provide early testers (including journalists, platforms, and researchers) access to the classifier for feedback.

We anticipate the classifier approach will allow us to better understand — and better mitigate — harms caused by the use or dissemination of images generated by OpenAI's tools. The degree to which the classifier fosters an ability to learn and improve our safety systems is a primary indicator of success.

We continue to invest in and stay abreast of research in the image provenance and broader provenance space, to effectively evolve and expand our approach in the future.

5 Obstacles

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

OPENAI'S RESPONSE

The greatest obstacles we have faced in considering various image provenance approaches have been around balancing trade-offs of various provenance techniques and managing external expectations, as noted above.

The PAI Framework could benefit from providing clear directions on managing conflicting goals and priorities within or related to the principle of disclosure. Should we prioritize accuracy and durability over accessibility and openness? Should we prioritize building momentum behind a single approach over concerns around reinforcing unrealistic societal and policymaker expectations? We understand these are areas in which PAI is working to develop further guidance.

After deciding to explore a provenance classifier, we found that PAI's Framework did not provide any guidance on determining which parties should have access to the provenance tool. However, we note that PAI has conducted related work on access protocols for synthetic media detection systems. Due to our concern of not over-promising the utility of the classifier, coupled with the fact that it's a new tool we are experimenting with, we believe it's prudent to limit access to the tool. However, to increase the likelihood of achieving our goal of using the classifier to mitigate harms, it may be appropriate to provide access to certain organizations or individuals working to combat synthetic image harms, such as social media companies and academic researchers. While we plan to grant access to trusted early testers, to date, this is an open question.

6 Benefits

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

OPENAI'S RESPONSE

The case study process itself has prompted a helpful post-mortem exercise on the last several months of internal research, discussion, and decision-making. Reviewing PAI's Framework anew with greater context on disclosure opportunities and limitations, and broadly greater experience in working to prevent abuse of synthetic media, has provided further clarity on what the PAI Framework offers and where it can be expanded.

The Practices for **Builders of Technology and Infrastructure** are broadly instructive, though as noted elsewhere, do not engage in discussion of potential trade-offs or downsides. For instance, there are two principles ("Take steps to provide disclosure mechanisms for those creating and distributing synthetic media." and "make best efforts to apply indirect disclosure elements (steganographic, media provenance, or otherwise) within respective assets and stages of synthetic media production") that offer guidance without addressing or discussing the associated downsides or risks, particularly given the state of current provenance methods. We understand PAI continues to refine and improve upon this guidance.

7 Conclusion/Key Takeaways

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

OPENAI'S RESPONSE

At the time of writing, the DALL·E 3 provenance classifier has not yet launched, but we intend to do so in the first half of 2024, likely to a group of trusted testers (e.g., journalists, platforms, and researchers) for feedback.

Key lessons learned:

- No current provenance approach is a silver bullet. No current approach checks all of the boxes: durability to adversarial activity and adaptation, broad consumer access or visibility, building momentum behind a shared industry approach, etc.
- Setting appropriate expectations is perhaps the greatest challenge we face in developing an image provenance approach. At minimum, this requires educating policymakers and the broader public on the trade-offs of various provenance approaches, especially on the limited robustness of current techniques.

Open questions:

- What kind of interest will we receive to access our provenance classifier? Would external parties derive significant benefit in having access to the classifier? If so, which?
- How will we assess whether to adopt additional or other provenance techniques in the future? At what point will we consider the durability of a given method to have increased to a degree that we are comfortable adopting it?