PAI's Responsible Practices for Synthetic Media

A Framework for Collective Action

FEBRUARY 27, 2023

PARTNERSHIP ON AI

The Partnership on AI's (PAI) Responsible Practices for Synthetic Media is a set of recommendations to support the responsible development and deployment of synthetic media.

These practices are the result of feedback from more than 100 global stakeholders. It builds on PAI's work over the past four years with representatives from industry, civil society, media/journalism, and academia.

With this framework, we seek to:

1. Advance understanding on how to realize synthetic media's benefits responsibly, building consensus and community around best practices for key stakeholders from industry, media/journalism, academia, and civil society
2. Both offer guidance for emerging players and larger players in the field of synthetic media
3. Align on norms/practices to reduce redundancy and help advance responsible practice broadly across industry and society, avoiding a race to the bottom
4. Ensure that there is a document and associated community that are both useful and can adapt to developments in a nascent and rapidly changing space
5. Serve as a complement to other standards and policy efforts around synthetic media, including internationally

Governance and Involvement

The intended stakeholder audiences are those building synthetic media technology and tools, or those creating, sharing, and publishing synthetic media.

Several of these stakeholders will launch PAI's Responsible Practices for Synthetic Media, formally joining this effort. These organizations will:

1. Participate in the PAI community of practice
2. Contribute a yearly case example or analysis that explores the framework in technology or product practice

PAI will not be auditing or certifying organizations. This framework includes suggested practices developed as guidance.

PAI's Responsible Practices for Synthetic Media is a living document. While it is grounded in existing norms and practices, it will evolve to reflect new technology developments, use cases, and stakeholders. Responsible synthetic media, infrastructure development, creation, and distribution are emerging areas with fast-moving changes, requiring flexibility and calibration over time. PAI plans to conduct a yearly review of the framework and also to enable a review trigger at any time as called for by the AI and Media Integrity Steering Committee.

The Framework's Focus

Synthetic media presents significant opportunities for responsible use, including for creative purposes. However, it can also cause harm. As synthetic media technology becomes more accessible and sophisticated, its potential impact also increases. This applies to both positive and negative possibilities — examples of which we only begin to explore in this Framework. The Framework focuses on how to best address the risks synthetic media can pose while ensuring its benefits are able to be realized in a responsible way.

Further, while the ethical implications of synthetic media are vast, implicating elements like copyright, the future of work, and even the meaning of art, the goal of this document is to target an initial set of stakeholder groups identified by the PAI AI and Media Integrity community that can play a meaningful role in: (a) reducing the potential harms associated with abuses of synthetic media and promoting responsible uses, (b) increasing transparency, and (c) enabling audiences to better identify and respond to synthetic media.

For more information on the creation, goals, and continued development of PAI's Responsible Practices for Synthetic Media, see this FAQ.

PAI's Responsible Practices for Synthetic Media

Those building technology and infrastructure for synthetic media, creating synthetic media, and distributing or publishing synthetic media will seek to advance ethical and responsible behavior.

Here, synthetic media, also referred to as generative media, is defined as 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. See Appendix A for more information on the framework's scope.

PAI offers **recommendations for different categories of stakeholders** with regard to their roles in developing, creating, and distributing synthetic media. **These categories are not mutually exclusive.** A given stakeholder could fit within several categories, as in the case of social media platforms. These categories include:

- Those building technology and infrastructure for synthetic media
- Those creating synthetic media
- Those distributing and publishing synthetic media

SECTION 1

Practices for Enabling Ethical and Responsible Use of Synthetic Media

1. Collaborate to advance research, technical solutions, media literacy initiatives, and policy proposals to help counter the harmful uses of synthetic media. We note that synthetic media can be deployed responsibly or can be harnessed to cause harm.

   Responsible categories of use may include, but are not limited to:

   - Entertainment
   - Art
   - Satire
   - Education
   - Research

2. Conduct research and share best practices to further develop categories of responsible and harmful uses of synthetic media.

   These uses often involve gray areas, and techniques for navigating these gray areas are described in the sections below.

3. When the techniques below are deployed to create and/or distribute synthetic media in order to cause harm (see examples of harm in Appendix B), pursue reasonable mitigation strategies, consistent with the methods described in Sections 2, 3, and 4.

The following techniques can be deployed responsibly or to cause harm:

- Representing any person or company, media organization, government body, or entity
- Creating realistic fake personas
- Representing a specific individual having acted, behaved, or made statements in a manner in which the real individual did not
- Representing events or interactions that did not occur
- Inserting synthetically generated artifacts or removing authentic ones from authentic media
- Generating wholly synthetic scenes or soundscapes

For examples of how these techniques can be deployed to cause harm and an explicit, nonexhaustive list of harmful impacts, see Appendix B.

SECTION 2
Practices for **Builders of Technology and Infrastructure**

Those building and providing technology and infrastructure for synthetic media can include: B2B and B2C toolmakers; open-source developers; academic researchers; synthetic media startups, including those providing the infrastructure for hobbyists to create synthetic media; social media platforms; and app stores.

4. **Be transparent to users about tools and technologies'** capabilities, functionality, limitations, and the potential risks of synthetic media.

5. Take steps to **provide disclosure mechanisms** for those creating and distributing synthetic media.

   Disclosure can be **direct and/or indirect**, depending on the use case and context:

   - Direct disclosure is viewer or listener-facing and includes, but is not limited to, content labels, context notes, watermarking, and disclaimers.
   - Indirect disclosure is embedded and includes, but is not limited to, applying cryptographic provenance to synthetic outputs (such as the C2PA standard), applying traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, and single-frame disclosure statements in videos.

6. **When developing code and datasets, training models, and applying software** for the production of synthetic media, **make best efforts to apply indirect disclosure elements** (steganographic, media provenance, or otherwise) within respective assets and stages of synthetic media production.

   Aim to disclose in a manner that mitigates speculation about content, strives toward resilience to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation. (Note: The ability to add durable disclosure to synthetic media is an open challenge where research is ongoing).

7. **Support additional research to shape future data-sharing initiatives** and determine what types of data would be most appropriate and beneficial to collect and report, while balancing considerations such as transparency and privacy preservation.

8. Take steps to **research, develop, and deploy** technologies that:

   - Are as forensically detectable as possible for manipulation, without stifling innovation in photorealism.
   - Retain durable disclosure of synthesis, such as watermarks or cryptographically bound provenance that are discoverable, preserve privacy, and are made readily available to the broader community and provided open source.

9. **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.

SECTION 3
Practices for **Creators**

Those creating synthetic media can range from large-scale producers (such as B2B content producers) to smaller-scale producers (such as hobbyists, artists, influencers and those in civil society, including activists and satirists). Those commissioning and creative-directing synthetic media also can fall within this category. Given the increasingly democratized nature of content creation tools, anyone can be a creator and have a chance for their content to reach a wide audience. Accordingly, these stakeholder examples are illustrative but not exhaustive.

10. **Be transparent** to content consumers about:

    - How you received **informed consent** from the subject(s) of a piece of manipulated content, appropriate to product and context, except for when used toward reasonable artistic, satirical, or expressive ends.
    - How you think about the ethical use of technology and use restrictions (e.g., through a **published**, accessible policy, on your website, or in posts about your work) and consult these guidelines before creating synthetic media.
    - The capabilities, limitations, and potential risks of synthetic content.

11. **Disclose** when the media you have created or introduced includes synthetic elements especially when failure to know about synthesis changes the way the content is perceived. Take advantage of any disclosure tools provided by those building technology and infrastructure for synthetic media.

    Disclosure can be **direct and/or indirect**, depending on the use case and context:

    - Direct disclosure is viewer or listener-facing and includes, but is not limited to, content labels, context notes, watermarking, and disclaimers.
    - Indirect disclosure is embedded and includes, but is not limited to, applying cryptographic provenance to synthetic outputs (such as the C2PA open standard), applying traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, and single-frame disclosure statements in videos.

    Aim to disclose in a manner that mitigates speculation about content, strives toward resilience to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation.

SECTION 4
Practices for **Distributors and Publishers**

Those distributing synthetic media include both institutions with active, editorial decision-making around content that mostly host first-party content and may distribute editorially created synthetic media and/or report on synthetic media created by others (i.e., media institutions, including broadcasters) and online platforms that have more passive displays of synthetic media and host user-generated or third-party content (i.e., social media platforms).

For both active and passive distribution channels

12. **Disclose** when you confidently detect third-party/user-generated synthetic content.

    Disclosure can be **direct and/or indirect**, depending on the use case and context:

    - Direct disclosure is viewer or listener-facing, and includes, but is not limited to, content labels, context notes, watermarking, and disclaimers.
    - Indirect disclosure is embedded and includes, but is not limited to, applying cryptographic provenance to synthetic outputs (such as the C2PA open standard), applying traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, and single-frame disclosure statements in videos.

    Aim to disclose in a manner that mitigates speculation about content, strives toward resilience to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation.

13. **Provide a** published, accessible **policy** outlining the organization's approach to synthetic media that you will adhere to and seek to enforce.

For active distribution channels

Channels (such as media institutions) that mostly host first-party content and may distribute editorially created synthetic media and/or report on synthetic media created by others.

14. **Make prompt adjustments** when you realize you have unknowingly distributed and/or represented harmful synthetic content.

15. **Avoid distributing unattributed** synthetic media **content** or reporting on harmful synthetic media created by others without clear labeling and context to ensure that no reasonable viewer or reader could take it to not be synthetic.

16. **Work towards** organizational **content provenance** infrastructure for both non-synthetic and synthetic media, while respecting privacy (for example, through the C2PA open standard).

17. **Ensure that transparent and informed consent** has been provided by **the creator and the subject(s) depicted** in the synthetic content that will be shared and distributed, even if you have already received consent for content creation.

For passive distribution channels

Channels (such as platforms) that mostly host third-party content.

18. **Identify** harmful synthetic media being distributed on platforms by implementing reasonable technical methods, user reporting, and staff measures for doing so.

19. **Make prompt adjustments** via labels, downranking, removal, or other interventions like those described here, when harmful synthetic media is known to be distributed on the platform.

20. **Clearly communicate** and **educate** platform users about synthetic media and what kinds of synthetic content are permissible to create and/or share on the platform.

Appendices

APPENDIX A
PAI's Responsible Practices for Synthetic Media Scope

While this framework focuses on highly realistic forms of synthetic media, it recognizes the threshold for what is deemed highly realistic may vary based on an audience's media literacy and across global contexts. We also recognize that harms can still be caused by synthetic media that is not highly realistic, such as in the context of intimate image abuse.

This framework has been created with a focus on audiovisual synthetic media, otherwise known as generative media, rather than synthetic text which provides other benefits and risks. However, it may still provide useful guidance for the creation and distribution of synthetic text.

Additionally, this framework only covers generative media, not the broader category of generative AI as a whole. We recognize that these terms are sometimes treated as interchangeable.

Synthetic media is not inherently harmful, but the technology is increasingly accessible and sophisticated, magnifying potential harms and opportunities. As the technology develops, we will seek to revisit this framework and adapt it to technological shifts (e.g., immersive media experiences).

APPENDIX B
Potential Harms of Synthetic Media

List of potential harms from synthetic media to seek to mitigate:

- Impersonating an individual to gain unauthorized information or privileges
- Making unsolicited phone calls, bulk communications, posts, or messages that deceive or harass
- Committing fraud for financial gain
- Disinformation about an individual, group, or organization
- Exploiting or manipulating children
- Bullying and harassment
- Espionage
- Manipulating democratic and political processes, including deceiving a voter into voting for or against a candidate, damaging a candidate's reputation by providing false statements or acts, influencing the outcome of an election via deception, or suppressing voters
- Market manipulation and corporate sabotage
- Creating or inciting hate speech, discrimination, defamation, terrorism, or acts of violence
- Defamation and reputational sabotage
- Non-consensual intimate or sexual content
- Extortion and blackmail
- Creating new identities and accounts at scale to represent unique people in order to "manufacture public opinion"