PARTNERSHIP ON AI

Synthetic Media Code of Conduct

V3 - DRAFT FOR PUBLIC COMMENT

The PAI Synthetic Media Code of Conduct is a suite of guidelines on how to ethically and responsibly develop, create, and share synthetic media. This draft is the result of several months of feedback from a broad array of stakeholders (learn more here). **We are now seeking public comment via this form by Friday, September 9, 2022.**

PAI Synthetic Media Code of Conduct

As those building technology and infrastructure for synthetic media, creating synthetic media, and distributing or hosting synthetic media, we will operate in an ethical and responsible way when inventing, making, and sharing synthetic media.

Here, synthetic media, also commonly referred to as generative media, is defined as content (visual, text, audio, multimodal) that has been generated or modified (often via artificial intelligence) to create highly realistic outputs that, to the naked eye or ear, are indistinguishable from authentically captured media. These outputs may simulate artifacts, persons, or events. See ***Appendix A*** *for more detail on the Code's scope.*

We offer normative recommendations for ***stakeholders*** *contributing to the societal impact of synthetic media.* ***These categories are not mutually exclusive****; a specific stakeholder could fit within several categories, as in the case of technology platforms. These categories include:*

(1) Those building technology and infrastructure for synthetic media

(2) Those creating synthetic media

(3) Those distributing and hosting synthetic media

Section 1- Joint Commitments for Enabling Ethical and Responsible Synthetic Media

We will collaborate to advance research, technical solutions, media literacy initiatives, and policy proposals to help counter the use of malicious and deceptive synthetic media. We note that similar mechanisms for deploying synthetic media can be found in both reasonable or malicious uses.

- **Reasonable** uses may include:
  - Entertainment
  - Art or satire
  - Education
  - Research
- These uses often involve gray areas, some of which are considered in the (forthcoming) accompanying vignettes.

We will also take concrete mitigation strategies, make best faith efforts, and hold end-users accountable to ensure synthetic media is not created, distributed, and hosted that **intentionally deceives and/or harms by:**

- Claiming to be any person or from any company, media organization, government body, or entity without explicit consent to make that representation.
- Creating realistic fake personas that are intended to deceive individuals into assuming they are viewing or engaging with a real person.
- Representing a specific individual having acted or behaved in a manner in which they did not.
- Representing events that did not occur.
- Inserting synthetically generated artifacts or removing authentic ones from authentic media.
- Generating wholly synthetic scenes or soundscapes.
- Replicating a specific individual's voice and/or visual appearance, to deceive someone into sharing information or taking actions as if they were interacting with that person.

For examples of how these techniques are often deployed to cause harm, see **Appendix B** and the accompanying vignettes (forthcoming).

Section 2 - Commitments by Those Building Technology and Infrastructure for Synthetic Media

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

- **Be transparent** to users about capabilities, limitations, and potential risks of synthetic media through disclosure.
- **Require disclosure** when the media we have created or introduced includes synthetic elements for which failure to know about synthesis implicates the impact of the artifact (example AI labels; questions to consider when labeling).
    - Disclosure can be either **direct or indirect**, depending on the use case and context.
        - Direct disclosure includes, but is not limited to, content labels, context notes, watermarking, disclaimers, and media manipulation training.
        - Indirect disclosure includes, but is not limited to, applying cryptographic provenance to synthetic outputs, applying traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, single frame disclosure statements in videos.

    - When developing code and datasets, training models, and applying software for the production of synthetic media, we will make best efforts to apply indirect disclosure elements (steganographic or otherwise) within respective assets and stages of synthetic media production, regardless of the output's intended context.
    - We will disclose in a manner that mitigates speculation about content, is resilient to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation.
- **Commit to sharing relevant data** from synthesis tools to synthetic media detection efforts, upholding privacy while also enabling more robust training and eventual usefulness of detection models.
- **Research and develop** technologies that:
    - are as forensically detectable as possible for manipulation, without stifling innovation in photorealism; and,
    - retain durable disclosure of synthesis, such as watermarks and cryptographically-bound provenance that are discoverable, that are made readily available to the broader community and provided open source, while preserving privacy.
- **Provide a** published, accessible **policy** outlining the ethical use of our technologies and use restrictions that we will adhere to, enforce, and report on.

Section 3 - Commitments by Those Creating Synthetic Media

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.

- **Be transparent** to users about capabilities, limitations, and potential risks of synthetic content through disclosure.
- **Require disclosure** when the media we have created or introduced includes synthetic elements for which failure to know about synthesis implicates the impact of the artifact (example AI labels; questions to consider when labeling).
    - Disclosure can be either **direct or indirect**, depending on the use case and context.
        - Direct disclosure includes, but is not limited to, content labels, context notes, watermarking, disclaimers, and media manipulation training.
        - Indirect disclosure includes, but is not limited to, applying cryptographic provenance to synthetic outputs, applying

traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, single frame disclosure statements in videos.
    - We will disclose in a manner that mitigates speculation about content, is resilient to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation.
- **Ensure transparent and informed consent** from targets of manipulation, appropriate to product and context.
- If an institution, we will **provide a** published, accessible **policy** outlining the ethical use of our technologies and use restrictions that we will adhere to, enforce, and report on; if hobbyists and individual creators, we will consider being **transparent** about how we think about the ethical use of technology and consult guidelines before creating (e.g., on our website or in posts about our work).

Section 4 - Commitments by Those Distributing Synthetic Media
Those distributing synthetic media include both institutions with active, editorial decision-making around content that mostly host first party content (like media institutions, including broadcasters) as well as online platforms that have more passive displays of synthetic media and host user-generated or third party content.

Both active and passive distribution channels

- **Disclose** when the media we have created or introduced includes synthetic elements.
    - Disclosure can be either **direct or indirect**, depending on the use case and context.
        - Direct disclosure includes, but is not limited to, content labels, context notes, watermarking, disclaimers, and media manipulation training.
        - Indirect disclosure includes, but is not limited to, applying cryptographic provenance to synthetic outputs, applying traceable elements to training data and outputs, synthetic media file metadata, synthetic media pixel composition, single frame disclosure statements in videos.
    - We will disclose in a manner that mitigates speculation about content, is resilient to manipulation or forgery, is accurately applied, and also, when necessary, communicates uncertainty without furthering speculation.

- **Ensure transparent and informed consent** for the specific nature of how synthetic content will be shared and distributed, even if we have already received consent for content creation.
- **Provide a** published, accessible **policy** outlining our approach to synthetic media that we will adhere to, enforce, and report on.

Active distribution channels *(like media institutions) that mostly host first-party content*

- **Make prompt adjustments** when we realize we have unknowingly distributed and/or represented deceptive synthetic content.
- **Avoid distributing unattributed** synthetic media **content** or misrepresenting sources with the intent to deceive.
- **Work towards** organizational **content provenance** infrastructure for both authentic and synthetic media, while respecting privacy.

Passive distribution channels *(like platforms) that mostly host third-party content*

- **Implement** reasonable technical, user, and human **reporting measures** to identify synthetic media being distributed on platforms.
- **Publicly report** on detection rates, what happens to content detected as synthetic, detection process effectiveness in terms of false positives and false negatives, statistically representative sample of items labeled true and false.
- **Make prompt adjustments** via labels, downranking, removal, or other interventions like those described here, when intentionally misleading and harmful synthetic media is distributed on platform.
- **Clearly communicate** and **educate** platform users about synthetic media and what kinds of synthetic content are permissible on the platform.

APPENDIX A

Code Scope

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

This code has been created with a focus on audio-visual synthetic media. However, it may still provide useful guidance for the creation and distribution of synthetic text.

APPENDIX B

List of potential harms from synthetic media that mitigation strategies should be put in place to limit:

- 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, 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 images or audio.
- Extortion and blackmail.
- Creating new identities and accounts at scale to represent unique people in order to 'manufacture public opinion.'