PARTNERSHIP ON AI

The Partnership on AI response to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Background

Partnership on AI (PAI) is a non-profit partnership of academic, civil society, industry, and media organizations creating solutions to ensure that AI advances positive outcomes for people and society. PAI studies and formulates sociotechnical approaches aimed at achieving the responsible development of artificial intelligence (AI) (including machine learning (ML)) technologies. Today, we connect over 100 partner organizations in 14 countries to be a uniting force for the responsible development and fielding of AI technologies.

PAI develops tools, recommendations, and other resources by inviting multistakeholder voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance. We then work to promote adoption in practice, inform public policy, and advance public understanding. We are not an industry or trade group nor an advocacy organization. We aim to change practice, inform policy, and advance understanding.

The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI. The comments provided herein are intended to provide evidence-based information, based on PAI's research, in response to NIST's RFI.

Executive Summary

This submission responds to NIST's RFI. It addresses NIST's call for input on:

- NIST's development of a companion resource to its AI Risk Management Framework (RMF) for Generative AI
- Guidance for red-teaming for foundation models
- Guidance for synthetic content detection and disclosure

PAI welcomes this call for input. Generative AI technologies, and the foundation models on which they are often built, are developing rapidly, bringing the potential for both immense benefits and harms. The development and promotion of best practices for identifying and mitigating risks for these technologies, and for synthetic content produced by them, is urgent. As outlined in this paper, PAI is working actively in this field and has produced resources that can usefully inform NIST's tasks under the AI Executive Order.

This submission draws on three PAI workstreams, which set out best practices for different aspects of AI risk management and provide a framework within which technical solutions to achieve AI safety and security must be situated. These resources and the evidence below have been informed through comprehensive consultation with PAI's global partners spanning industry, academia, and civil society.

Some key themes from these resources that should inform NIST's work under the RFI are:

- Risk management measures must be framed holistically, and should be tailored to model and system capabilities and release strategies
- Risk management should address all actors across the AI value chain and AI lifecycle
- Audience-aware documentation is essential for transparency, accountability, and sound risk management practices
- Guidance and processes will require ongoing review to reflect evolving best-practices and diverse stakeholder expertise. Mechanisms should be in place to ensure they are updated as required.

These are emerging topics, and best practices continue to evolve. To ensure NIST's work is impactful, it will be necessary to draw on a broad range of expertise and perspectives. Forums such as PAI, with our cross-sectoral, multistakeholder partnership network, are uniquely placed to contribute to this work and **PAI welcomes the opportunity to support NIST in its next steps**. PAI's work in this space is ongoing and we would be pleased to contribute to further iterations of NIST's work.

Summary of Recommendations

Developing a generative AI companion resource to the NIST AI RMF

In developing a companion resource to the AI RMF for generative AI, NIST should ensure the resource:

- Includes mapping of the risk landscape for generative AI, including societal, malicious, and other kinds of risk. Risks beyond those traditionally considered as 'safety' risks should be included, e.g. labor market risks.
- Considers both known and speculative risks – especially for advanced foundation models which may be GenAI or agentic systems.
- Provides guidance for foundation model providers that is tailored and appropriate to the risk attendant on different model capabilities and release types.
- Is designed to evolve as new capabilities and risks emerge.
- Integrates and responds to risks related to synthetic content, including provisions for disclosure of such content.
- Considers what documentation processes are adequate to ensure transparency and accountability.
- Integrates human rights impacts assessments.
- Supports third party inspection of models and training data.
- Enables feedback mechanisms across the AI value chain.
- Considers the need to measure and disclose anticipated severe labor market risks.

We propose that NIST work with PAI and other stakeholders — including civil society, industry, academia, and media organizations — to build in feedback loops from the ecosystem, including how to achieve updates in a timely and effective way.

Developing guidance for red-teaming

In developing its red-teaming guidance, NIST should consider the need for red-teaming to be conducted as one measure in a comprehensive risk management framework.

NIST's red-teaming guidance should:

- Address all foundation models, including but not limited to dual-use models.
- Provide guidance tailored to proportionately addressing risks arising from different model capabilities and different model release strategies.
- Identify when in the product development cycle red-teaming should be performed.
- Include guidance about the composition of red teams, including the expertise of team members.
- Address when internal and external red-teaming are appropriate. Both internal and external red-teaming should be conducted where feasible.
- Address security considerations for red-teaming including security for sensitive red-teaming findings.
- Address how results of red-teaming should be reported. Where possible, these results should be made publicly available.

Synthetic Content

NIST's report on synthetic content should reflect the following:

- Guidance for responsible synthetic content will need specificity to address the risks posed by different modalities of synthetic content.
- Guidance for responsible synthetic content should clearly address the different roles of AI developers, content creators, and distributors in creating and distributing synthetic content.
- Any guidance for responsible synthetic content should include a mechanism for it to be updated, and should reflect technical innovations and evolving best practices.

To promote future work in this space:

- NIST should map potential harms associated with synthetic content, provide guidance tailored to avoiding those harms, and provide guidance about identifying further potential categories of harm that may be associated with particular forms or uses of synthetic content.
- NIST should promote an agreed glossary and terminology for synthetic content disclosure and detection methods, and suggest optimal

combinations of methods. These can build upon PAI's guidance on synthetic media.

Response to the RFI

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

PAI has developed a number of relevant guidance resources setting out best practices for Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, drawing on the expertise of its partners. Four of these are particularly relevant to the RFI:

- PAI's Guidance for Safe Foundation Model Deployment (2023) ("Model Deployment Guidance") is a framework for model providers to responsibly develop and deploy a range of AI models, promote safety for society, and adapt to evolving capabilities and uses.
- PAI's ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine Learning) Reference Document (2021) ("ABOUT ML Reference Document") is a resource providing guidance on documentation practices for transparency in developing machine learning systems. These practices underpin and support safe and responsible AI development and deployment.
- PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action (2023) ("Synthetic Media Framework") is a guidance tool for policy makers, technology builders, creators and distributors.
- PAI's Glossary for Synthetic Media Transparency Methods: Indirect Disclosure (2023).

PAI urges NIST to reflect the guidance and insights in these resources in fulfilling its functions under the EO.

1.a (1) Developing a companion resource to the AI Risk Management Framework (AI RMF) for generative AI.

PAI welcomes the development of this resource for generative AI. Model capabilities are increasing rapidly, and specific guidance is required to address the unique risks that can flow from these capabilities. The four PAI resources referred to above contain best practices relevant to managing generative AI risks.

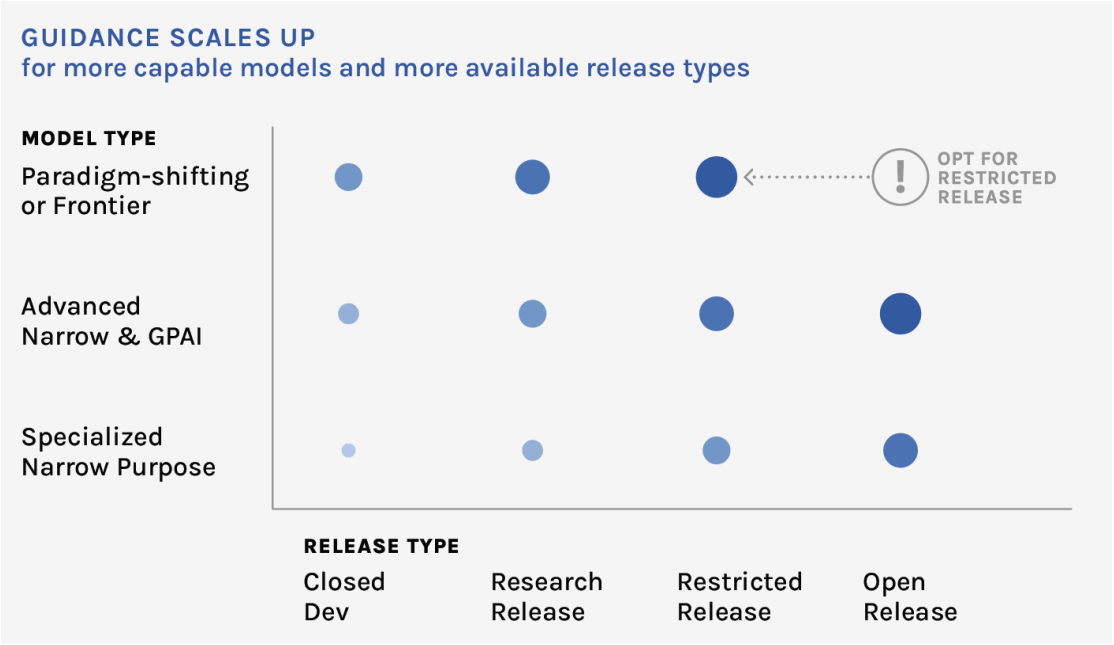
Risk management for generative AI must address the underlying technologies. Many of the powerful generative AI systems that have been spotlighted recently are built on foundation models. Addressing risks arising from foundation models, and particularly frontier models, can pose particular challenges and therefore

requires specific measures. PAI's Model Deployment Guidance offers a detailed and nuanced framework for identifying and mitigating these risks, including emergent risks. It recognises that the risks relating to foundation models depend on the model capabilities as well as release strategy. The Guidance is discussed in more detail below. **We propose that any Gen AI companion resource to the RMF:**

- **Ensures responsiveness to PAI's AI Risk Landscape,** including malicious uses, societal risks and other risks
- **Includes specificity of the type of generative AI being addressed in the companion framework** (see PAI's three different types of foundation models)
    - This includes providing guidance that is tailored to model release type (e.g. Open Release, Restricted API, Research, and Closed) and capability (e.g. frontier models, advanced models, and specialized narrow purpose models)
- **Ensures both known and speculative risks are considered** – especially for advanced foundation models which may be GenAI or agentic systems
- Is designed to **evolve as new capabilities and risks emerge**
- **Considers how best practices should be modified for open access model providers** who make model weights publicly available (see PAI's Model Deployment Guidance, especially the post-deployment guidelines in Annex A)
- **Considers best practices for models that are not released** (see the guidelines in PAI's Model Deployment Guidance for closed development release approach for a list of best practices)
- Integrates and responds to risks related to **synthetic content**
- Considers what **documentation processes are adequate to ensure transparency and accountability**
- **Integrates human rights impacts assessments**

Tailoring NIST's companion resource to model type, release approach and model capability

The existing NIST AI RMF is system and use-case agnostic, providing a single set of protocols to follow to identify risk for AI systems, but not tailoring those protocols to particular AI systems or contexts. **Building on PAI's Model Deployment Guidance, we propose that NIST categorizes risk according to system capabilities and release types for GenAI foundation models**. This will ensure that risk management guidance provided by NIST is tailored to specific attributes, and will be better positioned to address novel risks or technology-specific risks.



We propose specifically for NIST to consider:

- **Applicability across the landscape:** Ensure that NIST guidance is applicable across the entire spectrum of foundation models, encompassing existing models, frontier developments, and limited research releases.
- **Scaling guidelines and burden relative to model type and release:** PAI's guidance proposes a total of 22 guidelines. However, not all model types and releases are treated equally within the Guidance paradigm. **The suggested guidelines are more extensive for more capable models and more available release types.** This is important for:
  - **Scalability:** This approach accommodates the diversity of AI models and development scenarios.
  - **Accounting for openness**: Ensuring the need to adapt transparency and risk mitigation strategies specifically for open access models, providing guidance for both current and future open source model providers.

Release types in PAI's Model Deployment Guidance

- **Open Access:** Models released publicly with full access to key components, especially model weights. Can also include access to code, and data. Can be free or commercially licensed. Access can be downloadable or via cloud APIs and other hosted services.
- **Restricted API and Hosted Access:** Models available only through a controlled API, cloud platform, or hosted through a proprietary interface, with limits on use. Does not provide direct possession of the model. Allows restricting access and monitoring usage to reduce potential harms.
- **Closed Development:** Models developed confidentially within an organization first, with highly limited releases for internal evaluation or restricted external testing, before any potential public availability.
- **Research Release:** Models released in a restricted manner to demonstrate research concepts, techniques, demos, fine-tuned versions of existing

models. The release is meant to share knowledge and allow others to build upon it and excludes small-scale individual projects.

Model types/capabilities in PAI's Model Deployment Guidance

- **Specialized Narrow Purpose:** Models designed for narrowly defined tasks or purposes with limited general capabilities for which there is a lower potential for harm across contexts.
- **Advanced Narrow and General Purpose:** Models with generative capabilities for synthetic content like text, image, audio, video. Can be narrow purpose focused on specific tasks or modalities or general purpose. Also covers some narrow purpose models focused on scientific, biological or other high consequence domains.
  Encompasses general purpose models capable across diverse contexts, like chatbots/LLMs and multimodal models.
- **Paradigm-shifting or Frontier:** Cutting edge general purpose models that significantly advance capabilities across modalities compared to the current state of the art.

Build specificity into NIST's companion resource

The table below sets out the combination of model capability and release type. PAI's Model Deployment Guidance provides tailored safety guidance for each of these combinations.[1]

Narrow Purpose / Specialized Model Releases:	Advanced Narrow / Gen Purpose Releases:	Paradigm-Shifting or Frontier Model Releases:
• 1A - Specialized & Open • 1B - Specialized & Restricted • 1C - Specialized & Closed • 1D - Specialized & Research	• 2A - Advanced & Open • 2B - Advanced & Restricted • 2C - Advanced & Closed • 2D - Advanced & Research	• 3A - Frontier & Open • 3B - Frontier & Restricted • 3C - Frontier & Closed • 3D - Frontier & Research

Guidance for Significant Updates

NIST's companion resource for generative AI should contain guidance for foundation model providers that is appropriate to capabilities and release types.

[1] Tailored guidance for each combination of model capability and release type can be generated using the Custom Guidance tool.

The companion resource should be responsive to an AI risk landscape

Any framework for effectively managing risk for generative AI / foundation models, must identify and address relevant risks (this includes known, speculative, application and model risks). PAI's Model Deployment Guidance maps the categories of in-scope risks, capturing both the nature of those risks and the ways they can arise (see diagram on next page).

Building on this landscape mapping:

- NIST's companion resource for GenAI should **clearly map and maintain a risk landscape for generative AI and foundation models.**
- **There is a need to differentiate and address both known and speculative risks, and to ensure that the risk landscape used to inform NIST's work is applicable to all foundation models**, whether narrow, advanced or frontier. PAI's risk landscape addresses risks arising both from models and from potential applications of those models.
- **We propose a clear reflection of the risk landscape, including societal, malicious and other categories** including unknown risks in NIST's ongoing work.

The risk landscape for foundation model development[2]

SUB-CATEGORIES OF RISKS

- **Malicious uses:** Risks of intentional misuse or weaponization of models to cause harm
- **Societal risks:** Potential harms that negatively impact society, communities and groups
- **Other Risks:** Risks distinct from the above categories

Model risk refers to the potential risks associated with the foundation model itself. Includes biases in the training data, human-computer interaction harms resulting from interacting with the model, or vulnerabilities to adversarial attacks. Model risks focus on the inherent characteristics of the model and other negative impacts that model providers can address.

Application risk refers to potential risks that arise from downstream use-cases and applications built using foundation models or when these models are integrated into real-world products and services. Includes potential harms caused by incorrect or biased outputs and malicious uses.

Known risks are the risks that have been identified, acknowledged, and are reasonably well-understood. These risks are typically based on empirical evidence, research, or previous experiences with similar models or applications. Known risks are usually more predictable and quantifiable.

Speculative risks are the risks that are uncertain, hypothetical, or potential but have not been observed repeatedly or thoroughly studied. These risks may arise from emerging technologies, complex interactions, or unexpected consequences that are difficult to anticipate. Speculative risks are often more challenging to quantify or mitigate due to their uncertain nature.

[2] The Model Deployment Guidance addresses both risks from the foundation models themselves and risks that can arise downstream when others build applications using the models. While downstream developers have an important role in managing application risks, under the Guidance, model providers adopt accountability measures like providing synthetic media disclosures and supplying downstream use documentation, thereby addressing select application risks within the scope of the Guidance.

Integrating human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts

PAI's model deployment guidance calls for the implementation of comprehensive human rights due diligence methodologies to assess and address the impacts of models. Specifically for frontier models, we propose:

Example human rights actions, to be taken for frontier models

Example Baseline Practices

- Establish processes for conducting human rights impact assessments pre-deployment.
- Align with relevant guidance like the UN Guiding Principles on Business and Human Rights, and White House Blueprint for AI Bill of Rights. Proactively assess and address potential impacts on vulnerable communities.
- Continuously improve due diligence processes by collaborating with stakeholders and incorporating community feedback.

Example Recommended Practices

- Publicly disclose identified risks, due diligence methodologies, and measures to address impacts.

More widely, we encourage NIST to ensure that any new standards or accompanying tools for the RMF go beyond the traditional parameters of 'AI safety', to ensure that risks such as labor market impacts are factored in. The development of a GenAI companion resource for the RMF should demonstrate how it:

- Supports third party inspection of models and training data
- Incorporates the conduct of human rights due diligence
- Enables feedback mechanisms across the AI value chain
- Discloses synthetic content
- Measures and disclose anticipated severe labor market risks

Designing the resource to evolve as new capabilities and risks emerge

It is imperative that any GenAI companion resource is designed to evolve over time, and address emerging breakthroughs and use cases as they arise. This includes methods to incorporate feedback from a wide range of stakeholders as new risks emerge. **We propose that NIST work with PAI and other stakeholders to build these feedback loops from the ecosystem, including how to achieve updates in a timely and effective way.**

Any RMF companion resource for generative AI should also address risks posed by synthetic media.

NIST should ensure that any generative AI companion resource is integrated with its work on synthetic content. While synthetic content is addressed under a separate part of the RFI, this work should be integrated with the generative AI RMF. PAI's Synthetic Media Framework sets out best practices to mitigate risks associated with synthetic content, as discussed later in this paper.

Strong transparency and documentation are paramount

Proper documentation of model development is essential both to successful evaluation and to auditing of AI capabilities. PAI's ABOUT ML Reference Document provides principled guidance about how to approach the task of documentation for machine learning systems. See PAI's submission to the NTIA in 2023 for our guidance on effective documentation practices.

While the ABOUT ML resource shares specific best practices for designing strong documentation, PAI's Model Deployment Guidance sets out a number of protocols, and what they demand from a transparency perspective, **from the R&D stage through to post-deployment.**

Examples of protocols in PAI's Model Deployment Guidance: Transparency and documentation are required at various stages of the development lifecycle (See full list in Annex A)

Research and Development	**Produce a "Pre-Systems Card:** Disclose planned testing, evaluation, and risk management procedures for foundation/frontier models prior to development.
Pre-Deployment	**Publicly report model impacts and "key ingredient list"** Provide downstream use documentation
Post-Deployment	**Develop transparency reporting standards:** Collaboratively establish clear transparency reporting standards for disclosing foundation/frontier model usage and policy violations.
Societal Risks	Support third party inspection of models and training data

1.b Establishing guidelines to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests

Red-teaming is a critical tool for safe foundation model development and deployment. PAI's work with its partners in developing its Model Deployment Guidance indicated that it is a part of best-practice risk management for a range of foundation models – including, but not limited to, frontier models.

PAI therefore welcomes the development of best practice guidance for developers on this issue. This guidance should:

- **Address all foundation models**, including but not limited to dual-use models. Red-teaming is a part of best-practice risk risk management in developing foundation models with a wide range of capabilities and use-cases
- Be nuanced, addressing the different kinds and degrees of risk associated with **different model capabilities** and **different model release strategies**. Red-teaming practices must be tailored to address these factors. PAI's Model Deployment Guidance tailors risk management practices according to three categories of model capability (Narrow, Advanced, and Frontier) and four different release strategies (Open, Restricted, Closed, and Research)
- Identify **when in the product development cycle red-teaming should be performed**. Red-teaming should be performed iteratively through model development
- Include guidance about **the composition of red teams,** including the expertise of team members
- Address when **internal and external red-teaming** are appropriate. Both internal and external red-teaming should be conducted where feasible
- **Address security considerations** for red-teaming including security for sensitive red-teaming findings
- Address **how results of red-teaming should be reported**. Where possible, these results should be made publicly available, for instance as part of a "key ingredients list" published pre-deployment

Suggested red teaming practices for frontier and closed models in PAI's Model Deployment Guidance, that could be integrated into NIST's framework

Baseline practices

- Perform internal and external red teaming across model capabilities, use cases, and potential harms including dual-use risks using techniques such as adversarial testing, vulnerability scanning, and surfacing edge cases and failure modes.
- Conduct iterative red teaming throughout model development. Continuously evaluate results to identify areas for risk mitigation and improvements, including for planned safeguards.
- Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations.

Recommended Practices

- Commission external red teaming by independent experts such as domain experts and affected users to surface gaps.

Specificity remains important

- Any red teaming practices proposed by NIST should be reflective of the specific model type, capability and release approach.

While red-teaming has a key role in current best-practice for safe foundation model development/deployment, **it is only one part of robust safety practices**. PAI's Model Deployment Guidance situates red-teaming within a comprehensive risk management framework for foundation model developers. Guidance from NIST should highlight the fact that comprehensive risk management frameworks should be in place for all foundation model development.

Proper documentation of model development is essential in this context. Good documentation practices, as discussed in PAI's ABOUT ML Reference Document, will support robust red-teaming, and proper documentation of red-teaming practices will allow for appropriate reporting and scrutiny of findings.

What does NIST need to bear in mind as it drives forward its next steps?

In developing its red-teaming guidance, NIST should consider the need for red-teaming to be conducted as **one measure in a comprehensive risk management framework.** We are pleased to see that this approach (coupling red-teaming with wider practices such as documentation and reporting) appears to be a core part of NIST's model and framework ahead.

2. Reducing the Risk of Synthetic Content: Standards, Tools, Methods, and Practices

Summary

NIST has been tasked under the AI Executive Order with reporting on existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques for addressing certain risks associated with synthetic content.

In its reporting on reducing synthetic content risks, and in any further work to develop standards, tools, methods, and practices addressing those risks, NIST should give particular attention to:

- Mapping relevant harms, consistent with PAI's Synthetic Media Framework
- Guidance on technologies/best practices that addresses all key actors involved throughout the life cycle of synthetic media, including those building and developing, creating, and distributing synthetic content, with specific guidance tailored to each stakeholder
- Establishing a mechanism to ensure NIST guidance is kept up to date with evolving best practices, in particular for disclosure, detection, and harm assessment
- Promoting shared terminology, through a clear taxonomy for transparency and disclosure techniques (including their advantages and risk tradeoffs such as resilience and susceptibility to manipulation). PAI's Glossary for Synthetic Media Transparency Methods can be used as a model
- A set of questions included in this paper for NIST to consider when evaluating synthetic media transparency methods, including questions not reflected in NIST's RFI, for example, how can signals from a disclosure mechanism be interpreted differently by actors at various points in the life cycle of a piece of content (i.e. social media platforms and end users)?
- **The need to consider a context-based and multifaceted approach for disclosure** that involves permutations of metadata, watermarking, and fingerprinting will be critical, as well as acknowledging important limitations in detection and conveying those limitations to end users
- The need to understand that there is a distinction between content provenance and identification methods that reveal that a piece of content is synthetic to those who are not end users (e.g., developers and distributors), what we call indirect disclosure, and content disclosures methods that are directly end-user facing, like labels.

In February 2023, PAI released the first iteration of its Synthetic Media Framework. This AI governance resource sets out best practices for those building synthetic media technologies, those creating synthetic media, and those distributing synthetic media to ensure responsible use and avoid potential harms.

PAI's Synthetic Media Framework: Key Points

Through work with its Partner community of more than 100 organizations, PAI identified six key points for responsible creation, development, and distribution of synthetic media:

1. **Mechanisms to support responsible behavior should be targeted at actors across sectors, and reflect the roles played by different AI actors for managing risks and harms of generative AI**. Recommendations and mechanisms for accountability, including public policy, should address stakeholders throughout the synthetic media life cycle — specifically tech development, creation, and distribution.
2. **Consent is paramount.** Creators and Distributors of synthetic media should be transparent about whether they have received informed consent from their subject(s).
3. **Transparency is needed.** Builders and Creators should be transparent about the capabilities, functionality, limitations, and potential risks of synthetic media tools.
4. **Synthetic media should be clearly disclosed.** Builders, Creators, and Distributors are encouraged to enable and/or provide viewer or listener-facing labels as well as content-embedded provenance data.
5. **Synthetic media can be used responsibly or harmfully.** The Framework's Appendix B, which lists the potential harms of synthetic media, can be used to inform public policy (reproduced below). The Framework focuses on harm, and not intent.
6. **Effective governance requires adaptability.** Developments in generative AI are moving at a rapid and unprecedented pace. To maintain relevance and applicability, governance and policy must be similarly nimble.

Terminology and Scope

PAI's Synthetic Media Framework provides guidance for "synthetic media", defined to be "*visual, auditory, or multimodal content that has been generated or modified (commonly via artificial intelligence)*". This overlaps significantly with the definition of "synthetic content" in the AI Executive Order, although the latter term also includes generated text. NIST may need to consider specific issues relating to synthetic text that are not within scope for the Synthetic Media Framework.

Guidance for responsible synthetic content will need specificity to address the risks posed by different forms of synthetic content.

Reducing the risk of harmful uses of synthetic content will require accountability and measures for different stakeholders across the AI value chain

PAI's Synthetic Media Framework sets out best practices for three distinct groups involved in the life cycle of AI-generated content: (i) Technology Builders, (ii) Creators, and (iii) Distributors. This recognizes the different roles these actors play in creating and sharing synthetic media.

In developing responsible synthetic content guidance, NIST should clearly address the different roles of AI developers, content creators, and distributors in creating and distributing synthetic content. While we set out these different categories of stakeholders with regard to their roles in developing, creating, and distributing synthetic media it is important to note that these categories are not mutually exclusive. A given stakeholder could fit within several categories. For example, some social media platforms can be classified as a builder and distributor.

Identification of in-scope risks

Identifying in-scope categories of potential harms resulting from synthetic media is the starting point for risk management. PAI's Synthetic Media Framework includes the following non-exhaustive list of potential harms, notably developed with a global, multistakeholder group of over 100 stakeholders providing input:

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"

NIST should map potential harms associated with synthetic content, provide guidance tailored to avoiding those harms, and provide guidance about identifying further potential categories of harm that may be associated with particular forms or uses of synthetic content.

Guidance must also consider how harms tradeoff against potential benefits from synthetic content, including but not limited to those referenced as responsible uses in the Framework: entertainment, art, satire, education, and research.

Developing aligned terminology for transparency and disclosure methods for synthetic content

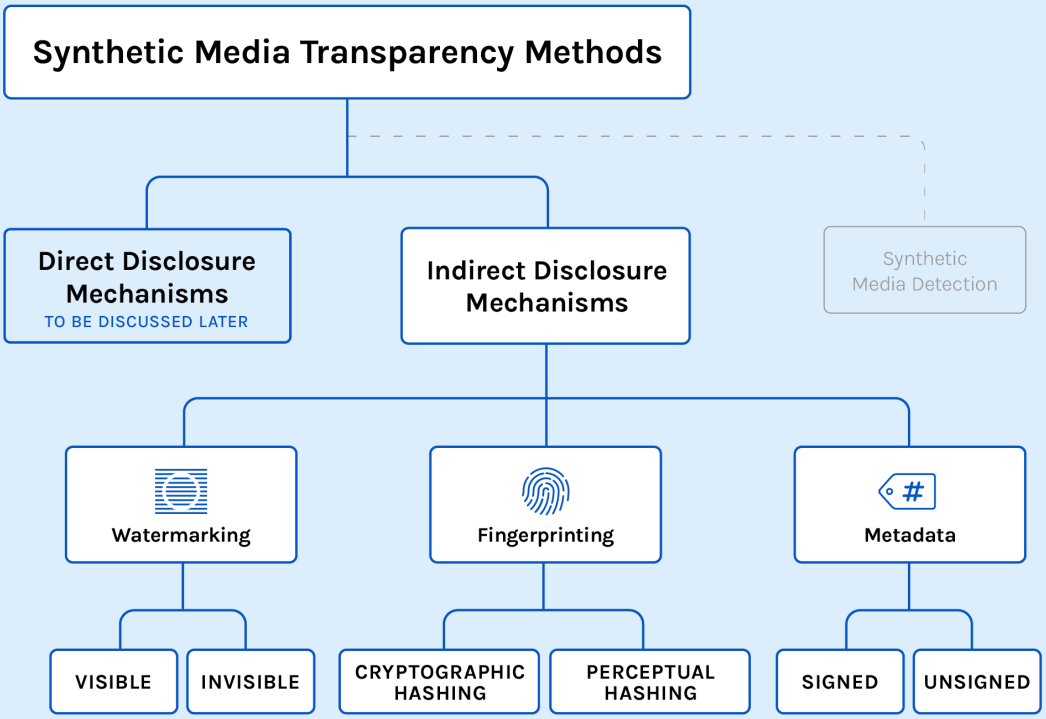
While there is increasing recognition that content transparency measures are needed to address the harms associated with synthetic content, key actors have not aligned on what combination of measures should be adopted. **The need for alignment is urgent.** To facilitate this, PAI has developed a series of resources. The first of these focuses on indirect disclosure practices, also known as content

provenance techniques, and includes a **glossary of key terms** — reflecting the fact that agreed language is a vital first step in reaching consensus on best practices (see Annex B for the Glossary). Note that indirect disclosure is a signal for conveying whether a piece of media is AI-generated or AI-modified, based on information about a piece of content's origin and/or evolution, but is *not* user facing.

We propose NIST drive forward this work on terminology, building on that of PAI, to establish a clear and consensus-based glossary related to disclosure methods. Two considerations for NIST:

- **A multifaceted approach:** PAI argues that only a multifaceted approach, one that involves permutations of metadata, watermarking, and fingerprinting (together or separately, depending on the use case) can respond to the challenge of synthetic media identification and transparency.
- **Limitations to note in synthetic media transparency:** While a context and content-specific, multifaceted approach to synthetic media transparency is ideal, each disclosure mechanism has its own set of limitations and it is important to convey them to end users. For example, non-pixel based invisible watermarking can be circumvented by taking a screenshot of the content. Unsigned metadata can easily be modified imperceptibly. Another mechanism, synthetic media detection (methods that rely on detecting unintentionally added patterns differentiating synthetic from non-synthetic content to determine the likelihood that a piece of content was generated or modified with AI), while in use at many institutions and a signal of content type that does not rely on good actors' proactivity, has proven limited in its applicability to real-world content. There is much work to be done about how institutions can make sense of such signals and, at times, best communicate the limits of such approaches to audiences without discouraging their implementation, which we will expand upon in upcoming efforts.

A proposed taxonomy for indirect disclosure, for NIST to build upon



NIST should promote an agreed glossary of terms for synthetic content disclosure methods, and can build upon PAI's guidance on synthetic media.

Questions for NIST to integrate and consider when evaluating indirect disclosure methods

The following questions for evaluating INDIRECT DISCLOSURE methods relate to the methods' impact on media transparency. They emerged from this MIT Tech Review piece and discussion with PAI's AI & Media Integrity Steering Committee.

- How resilient is *[disclosure mechanism]* to manipulation or forgery? How easily can it be removed by a bad actor?
- How accessible is *[disclosure mechanism]* to diverse audiences?
- How difficult would it be for organizations to adopt *[disclosure mechanism]* at scale?
- Is the *[disclosure mechanism]* associated with a piece of content maintained separately or is it embedded within the content's pixels or metadata?
- What organizations involved in the lifecycle of a piece of synthetic media (Builders, Creators, Distributors) need to opt-in in order for [disclosure

mechanism] to be successful and can it still be viable if it is adopted partially?
- How can *[disclosure mechanism]* signals be interpreted by those that interact with them through the lifecycle of a piece of synthetic media content, i.e. end users, internal decision makers, distributors, etc.?
- How resilient are associated detection tools to manipulation or adversarial attack?
- Can *[disclosure mechanism]* complement any other mechanisms to provide more robust disclosure?
- Does *[disclosure mechanism]* clash with any other existing disclosure mechanisms, i.e. are there other disclosure mechanisms that may render *[disclosure mechanism]* ineffective?

NIST will need a clear method to update guidance and standards on synthetic content

The proposed questions above for NIST to consider demonstrate that there are several existing limitations and that work is ongoing to develop techniques for reliably interpreting and disclosing synthetic content – best practices are still continuing to evolve. Further research is required to develop reliable techniques. PAI's Synthetic Media Framework is a living document, and will be updated consistently to reflect evolving capabilities, use-cases, and best practices for responsible development and use.

We propose that NIST adopt a similar approach, and recommend in its report that any synthetic content guidance should include a mechanism for it to be updated, and should reflect technical innovations and recognition of the technologies' social impact. While adapting and adjusting has traditionally been a challenge for standardization, this will be critically important to consider for achieving safety and security for synthetic content.

Conclusion

PAI welcomes this opportunity to provide input to NIST's work under the EO. Please contact us at policy@partnershiponai.org if further input would be of assistance, including about the PAI resources discussed in this submission.

Annexes

- Annex A: PAI's Guidance for Safe Foundation Model Deployment
- Annex B: Glossary - Synthetic Media Disclosure and Detection
- Annex C: Resources for designing, evaluating and auditing: limitations through which AI can cause harm (***Annex C** is set out within a separate document. Please follow the link to access these resources*).

ANNEX A

PAI's Guidance for Safe Foundation Model Deployment

PAI's Guidance for Safe Foundation Model Deployment contains 22 safety guidelines. Which guidelines should apply to a particular model varies according to model capability and release type. The Guidance contains Baseline practices and Recommended practices for each of the guidelines.

Research & Development		
1	**Scan for novel or emerging risks**	Proactively identify and address potential novel or emerging risks from foundation/ frontier models.
2	**Practice responsible iteration**	Practice responsible iteration to mitigate potential risks when developing and deploying foundation/frontier models, through both internal testing and limited external releases.
3	**Assess upstream security vulnerabilities**	Identify and address potential security vulnerabilities in foundation/frontier models to prevent unauthorized access or leaks.
4	**Produce a "Pre-Systems Card"**	Disclose planned testing, evaluation, and risk management procedures for foundation/frontier models prior to development.
5	**Establish risk management and responsible AI structures for foundation models**	Establish risk management oversight processes and continuously adapt to address real world impacts from foundation/frontier models.
Pre-Deployment		
6	**Internally evaluate models for safety**	Perform internal evaluations of models prior to release to assess and mitigate for potential societal risks, malicious uses, and other identified risks.

7	**Conduct external model evaluations to assess safety**	Complement internal testing through model access to third-party researchers to assess and mitigate potential societal risks, malicious uses, and other identified risks.
8	**Undertake red-teaming and share findings**	Implement red teaming that probes foundation/frontier models for potential malicious uses, societal risks and other identified risks prior to release. Address risks and responsibly disclose findings to advance collective knowledge.
9	**Publicly report model impacts and "key ingredient list"**	Provide public transparency into foundation/frontier models' "key ingredients" testing evaluations, limitations and potential risks to enable cross-stakeholder exploration of societal risks and malicious uses.
10	**Provide downstream use documentation**	Equip downstream developers with comprehensive documentation and guidance needed to build safe, ethical, and responsible applications using foundation/frontier models. (Note: It is well understood downstream developers play a crucial role in anticipating deployment-specific risks and unintended consequences. This guidance aims to support developers in fulfilling that responsibility.)
11	**Establish safeguards to restrict unsafe uses**	Implement necessary organizational, procedural and technical safeguards, guidelines and controls to restrict unsafe uses and mitigate risks from foundation/frontier models.
Post-Deployment		
12	**Monitor deployed systems**	Continuously monitor foundation/frontier models post-deployment to identify and address issues, misuse, and societal impacts.
13	**Implement incident reporting**	Enable timely and responsible reporting of safety incidents to improve collective learning.
14	**Establish decommissioning policies**	Responsibly retire foundation/frontier models from active use based on well-defined criteria and processes.
15	**Develop transparency reporting standards**	Collaboratively establish clear transparency reporting standards for disclosing foundation/frontier model usage and policy violations.

Societal Impact (cross-cutting through the model's lifecycle)		
16	**Support third party inspection of models and training data**	Support progress of third-party auditing capabilities for responsible foundation/frontier model development through collaboration, innovation and transparency.
17	**Responsibly source all labor including data enrichment**	Responsibly source all forms of labor, including for data enrichment tasks like data annotation and human verification of model outputs.
18	**Conduct human rights due diligence**	Implement comprehensive human rights due diligence methodologies to assess and address the impacts of foundation/frontier models.
19	**Enable feedback mechanisms across the AI value chain**	Implement inclusive feedback loops across the AI value chain to ethically identify potential harms.
20	**Measure and disclose environmental impacts**	Measure and disclose the environmental impacts resulting from developing and deploying foundation/frontier models.
21	**Disclose synthetic content**	Adopt responsible practices for disclosing synthetic media and advance solutions for identifying other synthetic content
22	**Measure and disclose anticipated severe labor market risks**	Measure and disclose potential severe labor market risks from deployment of foundation/frontier models.

ANNEX B

Glossary: Synthetic Media Disclosure and Detection

Synthetic Media (or "Generative Media")	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. **EXAMPLES** ● Ukraine: Deepfake of President Zelensky Asking Ukrainian Forces to Surrender ● Argentina: AI-Generated Image Campaign Showing What Abducted Children May Look Like as Adults ● Pakistan: Deepfake of Former Prime Minister Khan Addressing an Online Elections Rally
Synthetic Media Transparency Methods	The umbrella term used to describe signals for conveying whether a piece of media is AI-generated or AI-modified. Such signals can either be INDIRECT (not user facing) or DIRECT (user facing). INDIRECT DISCLOSURE signals can support understanding of whether content has been AI-generated or AI-modified and, when appropriate, guide development of DIRECT DISCLOSURES to audiences and end-users. **EXAMPLES** INDIRECT DISCLOSURES, DIRECT DISCLOSURES, SYNTHETIC MEDIA DETECTION
Indirect Disclosure	A signal for conveying whether a piece of media is AI-generated or AI-modified, based on information about a piece of content's origin and/or its evolution; is not user facing. The "disclosure" that takes place is typically to entities involved in content development, creation, and distribution—but it can be used to inform direct, or audience/user-facing, disclosure. **EXAMPLES** WATERMARKS, FINGERPRINTS, METADATA

Proactive Methods ● At generation ● Post-generation	SYNTHETIC MEDIA TRANSPARENCY METHODS that rely upon an actor purposefully applying a signal that can then be identified by a third-party who is able to detect or interpret the signal. Notably, in an adversarial setting, bad actors will not leverage such techniques and may attempt to alter existing signals. Can be classified further into two categories: **AT GENERATION** Signal is applied automatically by a media generation model at the moment of creation. **POST-GENERATION** Signal is applied after creation. **EXAMPLES** WATERMARKS, FINGERPRINTS, METADATA
Derived Methods	SYNTHETIC MEDIA TRANSPARENCY METHODS that determine the origin or evolution of media based on signals that do not rely upon a disclosure signal being applied by an actor. **EXAMPLES** SYNTHETIC MEDIA DETECTION
Direct Disclosure	A signal for conveying *to users* whether a piece of media is AI-generated or AI-modified; often informed by INDIRECT DISCLOSURES. **EXAMPLES** Labels, content overlays
Watermarking ● Invisible ● Visible	The PROACTIVE (at, or post, generation) insertion of modifications into a piece of content that can help support interpretations of how the content was generated and/or edited. Can come in two forms: **INVISIBLE** Modifications made to a piece of content that are imperceptible to the human eye or ear. Can only be identified by a WATERMARK DETECTOR (distinct from SYNTHETIC MEDIA DETECTION). **VISIBLE** Modifications made to a piece of synthetic content that are detectable to the human eye or ear and do not require the use of a DETECTOR to interpret them. **EXAMPLES** Google's SynthID (invisible), Meta AI's Imagine (visible)

Watermark Key/Detector	A digital tool, similar to a password, that is required for detect- ing an invisible/ hidden WATERMARK embedded in a piece of content. Can be shared broadly (open) or be restricted to select players (closed), or in between. This choice can affect the technical robustness of a watermark and its societal impact.
Fingerprinting • Cryptographic hashing • Perceptual hashing	The PROACTIVE (at, or post, generation) process by which a hash is generated for a piece of content for the purpose of identifying that content at a later date. Such hashes must be stored in a database in order to verify future content against the original. Unlike WATERMARKING, this hash is not embedded in the content file itself. Also known as "hashing and matching" or "hashing and logging." Can come in two forms: **CRYPTOGRAPHIC HASHING** An exact-match form of hashing where the hash for a piece of synthetic content will not match if the content has been modified in any way. **PERCEPTUAL HASHING** A probabilistic-match form of hashing where the hash for a piece of synthetic content is resilient to minor perturbations (i.e., will still match with minor changes). **EXAMPLE** YouTube's Content ID

Metadata ● Signed ● Unsigned	Information about the origin, structure, and/or editing history of a piece of content that is PROACTIVELY attached to the content itself. **SIGNED METADATA** Information that is PROACTIVELY attached to the content itself and stored using secure encryption; a trusted/validated signer certificate is added post generation. State of the art methods leverage cryptographic signatures. **UNSIGNED METADATA** Information that is PROACTIVELY attached to the content itself at generation but is *not* stored with secure encryption or validated with a trusted signer certificate, and potentially can be changed imperceptibly (weakening robustness). **NOTE** In policy and public discourse, metadata has sometimes been described as "media provenance" despite the fact that provenance can be a broader term describing *all* the methods under the umbrella of INDIRECT DISCLOSURE. For example, several PAI Partners refer to indirect disclosure methods as provenance methods. It is our hope that the terminology expressed in this glossary helps align the field on nomenclature. **EXAMPLES** C2PA Standard (signed), IPTC Standard (unsigned)
Synthetic Media Detection	Methods that rely on detecting unintentionally added patterns/forensic cues differentiating synthetic media from non-synthetic media to determine the likelihood that a piece of content was AI-generated or AI-modified; such methods do not rely on the PROACTIVE addition of artifacts such as WATERMARKS in content. Synthetic media detection is a DERIVED transparency method. **EXAMPLES** Intel's FakeCatcher, Google Jigsaw's Assembler (no longer active)

Detector Access ● Open ● Closed	Whether for WATERMARK DETECTION or SYNTHETIC MEDIA DETECTION, systems identifying synthetic media can either be shared broadly or restricted in their access, or somewhere in between. **CLOSED DETECTORS** Only available to a select number of organizations, minimizing the risk of adversarial exploitation at the expense of accessibility. **OPEN DETECTORS** Widely available, maximizing accessibility at the expense of increased risk of adversarial exploitation. **NOTE** Detector access does not necessarily have to be a binary choice between open and closed. PAI has conducted much work on the tradeoffs between open and closed access; see here, here, and here, all incorporating recommendations for "goldilocks" exposure between open and closed.