ITI Promoting Innovation Worldwide

Information Technology Laboratory
Attn: AI EO RFI
National Institute of Standards and Technology
100 Bureau Drive
Mail Stop 8900
Gaithersburg, MD 20899

via email to ai-inquiries@nist.gov

February 2, 2024

# ITI Feedback to Request for Information (RFI) Related to the National Institute of Standards and Technology's (NIST) Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Dear NIST Information Technology Laboratory Team,

The Information Technology Industry Council welcomes the opportunity to provide feedback to the National Institute of Standards and Technology **Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11).** We appreciate NIST's continued leadership in this space, and its ongoing commitment to collaborating with a diverse set of experts from industry, academia, civil society, international standards organizations, and other government agencies.

ITI represents the world's leading information and communications technology (ICT) companies. We promote innovation worldwide, serving as the ICT industry's premier advocate and thought leader in the United States and around the globe. ITI's membership comprises leading innovative companies from all corners of the technology sector, including hardware, software, digital services, semiconductor, network equipment, and other internet and technology-enabled companies that rely on ICT to evolve their businesses. Artificial Intelligence is a priority technology area for our member companies, who are both developing and using the technology to evolve their businesses.

ITI is committed to fostering the responsible development and deployment of AI. We have been actively engaged in shaping AI policy around the world. In 2021, we issued a set of *Global AI Policy Recommendations*, aimed at helping governments facilitate an environment that supports AI while simultaneously recognizing that there are challenges that need to be addressed as the uptake of AI grows around the world.[1] We also launched our AI Futures Initiative in 2023, which is an initiative comprised of technical and policy experts aimed at addressing challenging questions that are emerging in the global conversation on AI. We have published several policy papers via this Initiative, including on the *AI Value Chain and Foundation Models*[2], and on *Authenticating AI-*

[1] Our complete *Global AI Policy Recommendations* are available here: https://www.itic.org/documents/artificial-intelligence/ITI_GlobalAIPrinciples_032321_v3.pdf

[2] ITI's *Understanding Foundation Models & the AI Value Chain* paper is available here: https://www.itic.org/documents/artificial-intelligence/ITI_AIPolicyPrinciples_080323.pdf

*Generated Content*[3], both of which we think are particularly relevant to NIST's RFI. We have also actively worked to inform the efforts of the National Institute of Standards and Technology (NIST)[4] to create an AI Risk Management Framework (RMF) and have consistently contributed to the debate in the EU on its AI Act.

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

Use of the AI RMF & Developing an AI RMF for Generative AI

We understand that NIST has been directed to develop a companion document to the AI RMF focused on generative AI. Our companies are already leveraging NIST's AI RMF to address various AI-related risks. We recently outlined in a letter to Senators Hickenlooper and Blackburn ways in which ITI members are using the AI RMF, as well as ways in which NIST can further support organizations seeking to improve their AI risk management processes.[5] Additionally, many of our members are already undertaking efforts aimed at ensuring generative AI safety, including testing systems throughout their lifecycle, implementing content authentication and provenance techniques in order to enable end-users to identify AI-generated content, identifying and mitigating cybersecurity risks and designing AI systems in a transparent way to increase public trust.

In thinking through a generative AI RMF, NIST should outline key areas of risk that are unique to generative AI and include outcomes that organizations will need to achieve in order to manage those unique risks. It should also indicate where other existing Frameworks, such as the SSDF or the Cyber Framework, can help to support generative AI risk management. Finally, it should maintain a focus on cross-functional collaboration. The Govern function in the AI RMF is particularly useful in cultivating a culture of risk management across an organization and should be upheld in a Generative AI RMF as well.

We note that other jurisdictions are also in the process of developing frameworks or guidelines for Generative AI and/or advanced AI systems. With that in mind, **we encourage NIST to continue to work with international counterparts to increase alignment of approaches across borders, leverage international standards where possible**, and where not possible, work with stakeholders to produce crosswalks for the RMFs to relevant international standards, as it did for NIST AI RMF. Notably, the recently published ISO/IEC 42001 (artificial intelligence management system or AIMS) provides a certifiable AIMS framework in which AI systems can be developed and deployed as part of an AI assurance ecosystem. While we recognize that complete harmonization of approaches is likely not possible, we encourage NIST to consider these other Frameworks in the development of its own and align with them where appropriate. For example, Singapore recently issued a draft Model AI Governance Framework for Generative AI, aspects of which might be useful for NIST to consider when developing its own, which may also be relevant for NIST to consider.[6]

[3] ITI's paper on *Authenticating AI-Generated Content* is available here: https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf

[4] See ITI response to RFI on AI RMF Concept Paper here: ITI Comments on AI RMF Concept Paper FINAL.pdf

[5] See more here: https://www.itic.org/documents/artificial-intelligence/ITIJune2023ResponsetoSens.HickenlooperandBlackburnAIRMFLetter.pdf

[6] https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf

Roles of Stakeholders in the AI Value Chain

Regarding the roles of different stakeholders in the AI value chain in managing risks related to generative AI, we note that there is a role for multiple actors throughout the AI ecosystem. In particular, we highlight the important distinction between developers of generative AI -- and the foundation models upon which many generative AI applications are built upon -- and deployers of generative AI. Other stakeholders, such as integrators/intermediaries who incorporate AI into their products and services for use by others or end users, also have a role to play. In our paper on *Understanding the AI Value Chain & Foundation Models*, we outline the ways in which we believe developers and deployers can manage risk and communicate information.

- A developer (sometimes used interchangeably with producer) is in control of certain information and decisions, e.g., how the model's training data is selected and used, what kind of testing and validation is performed on the model, etc. Accordingly, developers are best positioned to manage model-level risks and understand the capabilities and limitations of a particular model. In many instances, an AI model can be built into other products that are then deployed by a different entity.
- A deployer decides the means by and purpose for which the AI system or model is ultimately being used and often has a direct relationship with the user or end user. While developers are best positioned to assess -- to the best of their ability -- and document the capabilities and limitations of a model or system, deployers, when equipped with certain information from developers, are best positioned to document and assess risks associated with a specific use case. This is true in the context of generative AI as well.

Transparency & Documentation for Generative AI

Transparency can play a key role in fostering trust in AI systems across the value chain and among different stakeholders, supporting good accountability practices. We published *AI Transparency Policy Principles*[7] in 2021 that explore overarching concepts related to transparency and offer perspectives on how potential transparency requirements can be most helpful. We also explore concepts related to transparency in our *Foundation Models* paper, which is directly relevant to the conversation around generative AI, given these models underpin many generative AI applications today.

- *Transparency in the context of the developer-deployer relationship.* Transparency is important in the context of the developer-deployer relationship. As we reference above, in order for a deployer to appropriately assess risks and determine whether a model is fit-for-purpose, they need to receive information from the developer about the model they are seeking to deploy. For example, information about the data used to train the system, including ways in which bias was mitigated or otherwise accounted for in the training dataset, how risks like training data integrity, sensitive data protection, and access controls were assessed and mitigated, limitations of the system, and intended uses might be especially useful for a deployer to have information about. It may also be useful to provide information to a deployer about the ways in which a systems' capabilities were evaluated, including the metrics used.
- *Transparency in the context of the deployer-end user relationship.* We believe that transparency can be useful for end users as well. This is especially true when those

[7] See our AI Transparency Policy Principles here: https://www.itic.org/documents/artificial-intelligence/ITIsPolicyPrinciplesforEnablingTransparencyofAISystems2022.pdf

individuals are interacting with a generative AI system and are consuming AI-generated content. We are supportive of disclosure to end users interacting with or using a generative AI system, where the use or interaction with the AI-generated content may mislead the user, such as in the case of photorealistic images. In this instance, basic information, potentially including information about how the system works, whether there is an opt-out option available, and any key limitations is important so that the end user is equipped with an understanding of how the content they are interacting with is being created. This can help to inform a user's decision as to whether and how to use the generative AI application.

With regard to documentation, there is not yet a consistent way to provide the above transparency information and it is important to note that documentation needs will likely vary depending on the audience. Several of our companies are using tools like model cards or system cards to disclose information about how an AI system works to users, including about the intended end use of the system, limitations of the model, and its expected or anticipated level of accuracy, among other things. While it may not be necessary or prudent to provide model cards in every instance, they can be a helpful way to inform both deployers and consumers about key characteristics of an AI system so that they are empowered to make decisions about if and how to use it.

It is also important to recognize that there are tradeoffs that come with transparency, especially when revealing information about a dataset. For example, there can be tension between providing transparency and explanation to affected parties and the safety and security of systems. Efforts to enhance AI system transparency must account for the potential that malicious actors will seek to exploit information about a system's underlying algorithms, data sources, and decision-making processes. A malicious actor could identify specific weaknesses in a system and then exploit said vulnerabilities, thus undermining safety and security. There are also tradeoffs related to privacy that are important to consider. Indeed, there is tension between protecting privacy and making information about datasets available, especially if such an action would potentially compromise privacy or reveal sensitive information about a person. In thinking through practices related to transparency then, we encourage NIST to reflect these tradeoffs and highlight that organizations should keep in mind the potential to undermine cybersecurity, reveal sensitive business or personal information, or otherwise compromise privacy in seeking to be transparent.

Techniques for Evaluating Generative AI Systems & Guidance and Benchmarks for Auditing AI Capabilities

We encourage NIST to review our submission to NTIA on AI Accountability[8], which discusses our initial perspectives on the evaluation and auditing of AI systems, including the role of impact and other assessments. Perhaps most importantly, **evaluation and auditing should be tailored based on the level of risk an AI system poses as well as relevant context**. This will ensure adequate protection for high-risk scenarios without impeding valuable innovations in low-risk scenarios. Internal audits, assessments, and certifications can play a role in facilitating trust, communicating information, and driving internal change. We believe all organizations in the AI value chain should adopt practices focused on driving accountability and fostering trust. At the same time, mechanisms that might be used to help support accountability, like audits, assessments, or certifications should be scoped based on the level of risk posed and relevant context.

[8] See ITI's comments to NTIA's AI Accountability RFC here: https://www.itic.org/documents/artificial-intelligence/ITICommentstoNTIARFConAIAccountabilityPolicy.pdf

In considering how best to advance work on evaluating generative AI systems, NIST should progress work on TEVV as outlined in the AI RMF and determine where and how those evaluation methods can apply to generative AI. Indeed, to the extent possible, we encourage NIST to consider how to structure generative AI evaluations in a way that leverages existing model risk management practices, particularly in highly regulated industries like financial services.

We understand that NIST is also seeking specific input on existing validation and verification techniques. We highlight several here. We would be happy to discuss perspectives on these techniques further.

- **Cross-validation:** This technique can act as a quality check on the AI model, ensuring that it not only performs well on the data it was trained on, but that it also performs accurately when it encounters new data. It helps to ensure that the models can make predictions on unseen data as well, demonstrating the robustness of the model. It involves dividing a dataset into training and testing data, training multiple models, and then evaluating the model performance on the testing data.
- **Holdout method:** This method also tests how well a model reacts when it encounters new data. Instead of dividing the data and training multiple models, this method simply requires separation of the data into two parts (testing and training). Once the model is trained on the selected data, it can be tested on the withheld data.
- **Confusion matrix analysis:** This type of approach can help to visualize the performance of a model. It visually represents where the model is performing well, and where there might be errors by demonstrating true positives, false positives, false negatives, and true negatives. This can help with analyzing a model's accuracy and recall.
- **Precision, recall and F1 Score:** These are metrics, as opposed to techniques in and of themselves, that can help to measure a model's performance. Precision measures the accuracy of the positive predictions a model provides, recall measures the completeness of the predictions, and the F1 score is a combined measure of both.
- **Receiver Operating Characteristic curve (ROC) and Area Under the Curve (AUC):** These are also metrics that are used to evaluate a binary classification model's performance. They are helpful when data is imbalanced, or in testing a binary classifier across different threshold values.
- **Bootstrapping:** This is a resampling technique that helps to estimate the variability of a model's performance. This technique uses resampling to create many "fake" datasets upon which a new model can then be trained and tested.
- **Sensitivity analysis:** A sensitivity analysis tests how a model's output changes when input data changes. It is helpful in understanding how input data impacts a model's performance, and how certain factors might be weighted in comparisons to others.
- **Adversarial testing:** Adversarial testing involves feeding the model adversarial examples (or inputs that are purposely meant to fool the model) in order to assess a model's robustness to attacks.
- **Explainability and interpretability tools:** Tools like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) can help with understanding a model's decision. There are also gradient-based methods, counterfactuals, as well as built-in tools like attention mechanisms, to help with explainability. Explainability can help with transparency and understanding how and why a model made a decision.

- **Monitoring in production:** This involves monitoring a model after deployment, which can include using the metrics above to track performance and accuracy. This can help to proactively identify and mitigate any issues that may emerge.

Generative AI Content Authentication

ITI recently published a paper on *Authenticating AI-Generated Content,* which explores ways in which AI-generated content can be validated. We discuss specific techniques further in the section on mitigating risks of synthetic content below.

AI Red-Teaming

Red-teaming generally refers to attempts to break or hack into a system in a way that reflects an actual attack. This type of activity can help an organization bolster its network defenses through strategically orchestrated security testing. In cybersecurity, a red-team is generally comprised of employees within an organization or otherwise hired by the organization to conduct red-teaming exercises.

The concept of red-teaming for AI models is based on red-teaming in a cybersecurity context, but in general, **refers to testing an AI model for a broader set of flaws or system failures.** The goal is the same – for an organization to identify possible vulnerabilities or potential for failures and address them prior to placing the model or system on the market. Such flaws can be related to security, privacy, and other types of abuse. For example, red-teaming an AI system might include attempting to feed the model malicious prompts so that it produces an incorrect or invalid output (known as prompt attacks), attempting to extract training data from a model, changing the behavior of the model by altering model weights or fine-tuning it in a way that causes it to behave maliciously (known as backdooring), feeding the model adversarial examples, launching data poisoning attacks, or otherwise trying to exfiltrate sensitive or personal data from the model.[9] In all of these instances, red-teaming can be useful in demonstrating where or how the model may need to be adjusted to better mitigate risks. The composition of a red-team in this context may vary from the more traditional cybersecurity red-teaming, especially because we cannot expect to simply task cybersecurity red teams with red-teaming AI models and there is not currently a mature cadre of AI red-teamers to conduct external testing. It might be reasonable to also comprise AI red-teams of employees from outside the development team.

**NIST should also recognize that cybersecurity red-teaming in an AI context may differ from red-teaming in a traditional environment**. Where a traditional red-team will halt its offensive once it has successfully exfiltrated data, this may not always be the case for testing the cybersecurity of AI systems. Once teams identify the root cause of the vulnerability, they may need to continue to exploit those vectors to identify where they might be able to change a model's behavior. Measuring and assessing these results with respect to AI system exploitation is critical to evaluate and prioritize cyber risks.

In developing standards and guidelines for AI red-teaming, **we encourage NIST to differentiate between red-teaming an AI model and red-teaming an AI system, since the terms are sometimes used interchangeably.** Red-teaming an AI model is focused on the model itself, where red-teams attempt to break the model itself or cause it to fail in some way. The activities described in the

[9] As an example, see Google's tactics, tests, and procedures for AI red-teaming: https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/#:~:text=One%20of%20the%20key%20responsibilities,how%20the%20technology%20is%20deployed.

paragraph above are specific to red-teaming the model itself. Red-teaming an AI system takes a more holistic view, where the objective is to test the entire system in which the AI model operates – this includes the data pipeline, user interface, infrastructure, and associated processes. In certain circumstances (e.g., particularly high-risk scenarios) this type of testing may help to evaluate a system's *overall* reliability and resilience (beyond just the individual model's robustness, to include, for instance, the validity or reliability of the data set used for training the model) and help to determine whether the system as a whole is fit-for-purpose.

At the same time, **AI red-teaming should not be construed as a silver bullet solution**. It is one technique that can help to pressure test AI models and expose their flaws but should be viewed as a part of a broader AI risk management approach. To be sure, while red-teaming can highlight flaws or vulnerabilities of a model or a system, it cannot, on its own, fully evaluate and mitigate societal risks that may arise when an AI system is deployed in the real world. As such, it is particularly important that red-teaming is undertaken in conjunction with other risk management practices, such as those outlined in the AI RMF. While we understand that NIST is directed to develop guidance on red-teaming for all AI developers, especially those developing dual-use foundation models, we believe it would be helpful for **NIST to articulate in its guidance that red-teaming activities should be scoped to the risk level of AI systems** including to **determine the extent to which red-teaming is appropriate.**

We also encourage NIST to consider that in the context of this EO, red-teaming will become de facto mandatory for certain companies given the directive under Section 4.2(C), which will require those organizations developing dual-use foundation models to provide the results of AI red-teaming performed to the Commerce Department. However, given AI red-teaming practices are rapidly evolving and because AI more generally is a dynamic space, **it is important that any guidance that NIST develops is future proof – it should be flexible enough to adapt to these changes and should not prematurely lock organizations into one set of practices that becomes quickly outdated**. In general, we have appreciated the way in which NIST's AI RMF offers companies flexibility to implement specific practices to achieve desired outcomes. In establishing these guidelines, NIST should keep in mind that in the future, it is also possible that some forms of AI red-teaming will be automated.

## 2. Reducing the Risk of Synthetic Content

While the introduction of generative AI has not necessarily created *new* risks, it has the potential to exacerbate or make more significant some of those risks given its ability to create content at scale. To be sure, mis and dis-information are well-documented risks that are also present with human-generated content. With that in mind, we recognize that the accessibility of generative AI applications has increased concerns across the community about the potential for such applications to spread false, untruthful, or inaccurate information.

As such, ITI recently released a paper on *Authenticating AI-Generated Content*[10]*,* which explores the risks discussed above, and offers perspectives on several of the themes NIST is interested in learning more about. The paper provides an overview of various techniques currently available to authenticate AI-generated content, including best practices and current limitations and/or

[10] See our Authenticating AI-Generated Content Paper here: https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf

tradeoffs related to the implementation of those techniques. In particular, the paper explores **watermarking**, **provenance tracking**, **metadata auditing**, and **human authentication**.

- **Provenance tracking** allows for tracing of the history and quality of a dataset. For AI-generated content, provenance refers to signals embedded in the dataset used to create the content as well as information about the content's source and history, and modifications to the content subsequent to its creation.
    - Examples of provenance tracking standards include:
        - ⇒ **C2PA Specification.** C2PA, which includes corporations of all sizes, leading journalism entities, not-for-profits, and academics, was formed in 2021 to address concerns related to misleading information and to develop technical standards for content provenance, called Content Credentials.[11] The C2PA's voluntary, open technical standard, or specification, works with visual, video and audio content by binding provenance information to a piece of media at its point of creation or when it is altered and attaches to the content with Content Credentials, which is a combination of secure metadata and watermarking. [12]
        - ⇒ **JPEG 2000.** The Joint Photographic Experts Group (JPEG), a joint committee formed by international standardization organizations ISO/IEC JTC 1/SC 29 and ITU-T Study Group 16, published the international standard "Secure JPEG 2000," which provides a syntax for security services such as source authentication and integrity verification to be applied to JPEG 2000 coded image data.[13] In 2020, JPEG began an exploration on "Fake Media" that resulted in a call for proposals for a standard that can facilitate the secure and reliable annotation of media asset creation and modifications, and intends to address use cases that are in good faith as well as those with malicious intent.[14]
- **Watermarking** involves embedding a signal in a piece of text or an image with information about its source or creator, or to identify whether it was AI-generated. There are different types of watermarking, including visible and invisible watermarks, and there are different places within the AI value chain where watermarks can be inserted.
- **Metadata auditing** builds upon techniques of altering or embedding signals in metadata and involves rigorously checking elements of content metadata, including editing history, timestamps, relevant device information, etc. to reveal inconsistencies about AI origins.
- **Human authentication** requires human involvement to verify whether content has been AI-generated at or across certain points of the AI value chain.

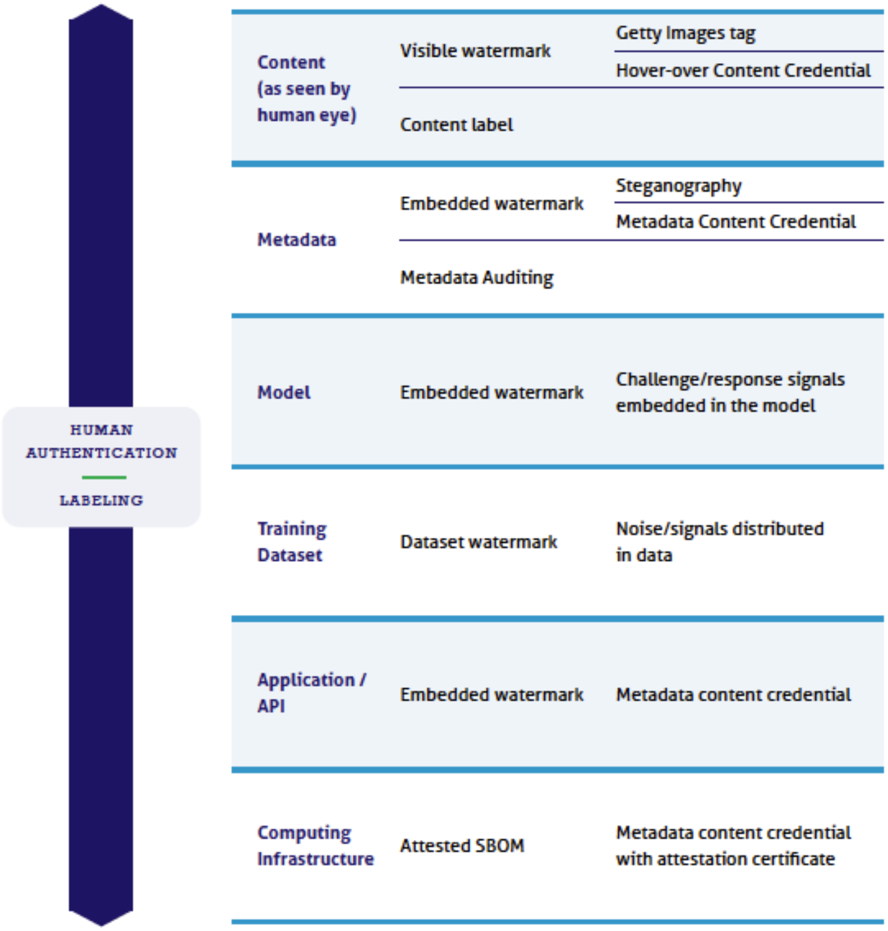
Several of these techniques overlap along the AI value chain or leverage similar processes. They also each have limitations and trade-offs that are discussed in more depth in ITI's paper. It is our view that a combination of authentication methods will be most effective at authenticating AI-generated content. Because of the evolving nature of these techniques as well as the dynamic nature of AI-generated content, **it is unlikely that any one technique will fully address all challenges, so we encourage NIST to consider a flexible and risk-based approach that allows for**

[11] Joint Development Foundation. (n.d.). Coalition for Content Provenance and Authenticity. Retrieved from https://c2pa.org/.
[12] Ibid.
[13] International Organization for Standardization. (n.d.). Information Technology. ISO/IEC 15444-8:2023 https://www.iso.org/standard/82566.html.
[14] Joint Photographic Experts Group (JPEG). (n.d.). JPEG Fake Media. Retrieved from https://jpeg.org/jpegfakemedia/

**multiple techniques appropriate to the level of risk and use case**. It may be useful to initially focus efforts on photorealistic material that may mislead the user into thinking it is human-generated.

We have pulled out a graphic from the paper that we believe is particularly helpful in demonstrating which techniques are applicable at different points in the value chain. The earlier in the AI value chain authentication techniques are applied, the more helpful such authentication is to a broader range of stakeholders and provides greater assurance about the integrity of the content.

Figure 1: AI Authentication Along the AI Value Chain


HUMAN AUTHENTICATION
LABELING

| | | |
|---|---|---|
| Content (as seen by human eye) | Visible watermark | Getty Images tag |
| | | Hover-over Content Credential |
| | Content label | |
| Metadata | Embedded watermark | Steganography |
| | | Metadata Content Credential |
| | Metadata Auditing | |
| Model | Embedded watermark | Challenge/response signals embedded in the model |
| Training Dataset | Dataset watermark | Noise/signals distributed in data |
| Application / API | Embedded watermark | Metadata content credential |
| Computing Infrastructure | Attested SBOM | Metadata content credential with attestation certificate |

While we have presented a high-level overview here, **we encourage NIST to review the paper in depth as it speaks to many of the questions NIST seeks answers to in the RFI.**

## 3. Advancing Responsible Global Technical Standards for AI Development

We appreciate that the EO recognizes the importance of developing and implementing a coordinated effort with key international allies and partners with standards development organizations by establishing a global engagement plan. It is important for the U.S. government to be a leader in AI policy discussions, including discussions around standards development. The U.S. should also encourage other nations to prioritize the development and use of international standards over unique national AI standards for areas of common global interest and to rely on and reference international standards in relevant policies and regulations. Below we offer more specific thoughts on several of the areas NIST references in its RFI.

Importance of Global Technical Standards & Global Trade Implications

We appreciate that the EO recognizes this and that NIST is considering how to encourage and support private sector participation in international standards development. The development and adoption of voluntary, consensus-based, industry-driven technical standards play a crucial role in driving AI innovation and facilitating international trade, reducing market access barriers, and contributing to U.S. economic growth. They ensure that products and services are compatible across borders, foster economies of scale, and create a common lexicon for trading partners. **Importantly, global technical standards serve as valuable tools for AI risk management and fostering the EO objectives for safe, secure, and trustworthy AI.** An example of this is the ongoing work of ISO-IEC/JTC 1/SC 42, which is developing multiple standards focused on enhancing transparency, taxonomy, objectives for explainability of ML models and AI systems, mitigating bias, risk management and impact assessment, and more. By establishing benchmarks, these standards help to facilitate innovation and technological advancements and structured governance and effective risk management.

Given the fact that many countries are beginning to consider how best to address AI risks and AI governance, **we encourage NIST to maintain international consistency to the extent possible,** especially as it progresses with the development of additional frameworks and guidelines pursuant to the Executive Order. This will promote interoperability, foster a common lexicon, and set a strong precedent for other countries and regions considering options for AI governance. It can also reduce the complexity and cost associated with complying with multiple risk management regimes around the world. As nearly all of ITI's members operate globally, we highly encourage NIST to contribute its deliverables to the development of international standards (such as ISO/IEC). Since the NIST convened efforts reflect consensus of broad and varied U.S stakeholders, contributing NIST deliverables to international standards efforts also supports the U.S. national standards strategy of increasing U.S. participation and leadership in international AI standardization. Furthermore, maintaining alignment with international standards progress and contributing NIST deliverables would reduce the need to continually develop crosswalks between NIST deliverables and relevant international standards (such as what happened for the NIST AI RMF), and best leverage the limited time and resources of U.S. experts to participate in NIST initiatives, industry initiatives, and international efforts.

AI Nomenclature & Terminology

ISO/IEC 22989: 2022 Information technology —Artificial intelligence concepts and terminology defines terms and stakeholders across the field of AI -- such as AI provider, user, customer, partner and subject and ISO/IEC 23053 establishes a framework for describing general AI systems using ML. With the growth of Generative AI and LLMs, amendments to both standards are in progress to add terminology for generative AI, foundation models and related concepts. NIST and U.S. stakeholders should contribute to these efforts to facilitate international consistency for AI terminology (including for defining new terms utilized in the U.S. EO AI etc. such as "dual use foundation models," terms included in the EU AI Act, such as "general purpose AI models with systemic risk," and the G7 definition of "advanced AI systems," all which presently have slightly different interpretations). SDOs should consider defining additional terms such as red-teaming and AI-generated or synthetic content to ensure international consistency. Indeed, there are presently multiple different interpretations of what a “frontier model” is, as well as what red-teaming means in an AI context.[15]

Future AI Standards Work

In the context of our comments to NTIA on AI Accountability Policy, we explored the important role that standards play in conducting various types of assessments and acknowledged the important work that ISO/IEC has done to create a standard that provides guidance on conducting AI impact assessments, ISO/IEC 42005:2023. This will be a useful starting point for organizations seeking to understand how best to undertake an impact assessment. However, regarding conformity assessment, development of additional globally-recognized standards on which to conduct such assessments as to a specific set of requirements would be useful.

Relatedly, there is an emerging field of frameworks, toolkits, and technical solutions to manage risks related to AI systems and to potentially test them in one way or another, but it is not clear how these tools are working in conjunction (or in comparison) with one another within and across industries. This can, in some cases, impact the uptake of assessments given it may be confusing to organizations as to how best to operationalize specific practices, or to consistently communicate the results of such an assessment. As such, we believe it would be helpful for NIST to actively contribute to and participate in international standardization work on broader testing and evaluation techniques in order to attempt to align these various solutions.

We are supportive of the NIST AI Risk Management Framework as it provides organizations a solid baseline from which to structure their risk management programs, and incorporates impact assessments as a best practice, but even in this regard there are not globally-recognized, consensus-based standards available to populate each of the categories in the RMF, as there are in other NIST frameworks, such as the Cybersecurity Framework. As such, the work that NIST is doing to support implementation of its AI Roadmap as well as the work it will do under the U.S. AI Safety Institute will be imperative to build upon the NIST AI RMF. To the extent possible, it would be beneficial to coordinate this work with other nations undertaking similar activities, especially related to the development of testing, evaluation, validation, and verification (TEVV) standards, and prioritizing the development of international standards as appropriate. In particular, guidelines to help organizations establish risk thresholds and determine their risk tolerance, as well as continued standardization work focused on helping organizations to determine risk level of a specific system (e.g., a frontier AI model) or use case will be critical.

Improving Stakeholder Engagement

We appreciate that NIST recognizes the importance of stakeholder participation in standards development bodies. Indeed, **international technical standards bodies, such as ISO/IEC, and technical experts, like those in industry and within government agencies like NIST, should be at the forefront of the development of AI technical standards and best practices.** These groups are well-positioned to provide perspectives on existing technical gaps which may need to be bridged with the creation of new standards that support technology innovation. Because the field of AI standards is nascent and continues to evolve, it is all the more important that the USG continues to support industry participation in international standards development efforts, specifically ISO/IEC JTC 1/SC 42, which can apply across various AI applications, and ISO/IEC JTC 1/SC 27 which is

[15] See additional discussion of how these terms are defined in various policy activities around the world here: https://www.atlanticcouncil.org/blogs/geotech-cues/ai-governance-on-a-global-stage-key-themes-from-the-biggest-week-in-ai-policy/

developing guidelines for AI security and privacy protection. **We strongly encourage NIST to prioritize and resource growing its engagement in AI standards development in ISO/IEC JTC 1, such that its participation is commensurate to the responsibilities being assigned to NIST, as this will ensure that the AI RMF is aligned with international standards that are in the process of being developed and vice versa**.

More generally, the U.S. government should seek to support increased U.S. industry participation in standards bodies on AI, through supporting industry-led bodies with transparent, rules-based processes, making the U.S. a more attractive meeting location for standards development organizations to host meetings, and ensuring that current and future policies do not unintentionally inhibit U.S. company participation in international standard bodies. In considering how to make the U.S. a more attractive meeting location for standards development organizations to host, NIST and the USG more broadly should recognize that attending standards meetings require significant investment in travel and time. Many international participants may not participate in U.S. based meetings due to visa processing delays. To the extent possible, NIST should highlight this issue in order to encourage participation from diverse international stakeholders by facilitating visa applications for foreign standards experts to routinely attend meetings in the U.S.

Mechanisms that could be leveraged to promote international standards collaboration

We encourage the U.S. to leverage existing bilateral and multilateral mechanisms to support international standards collaboration. For example, the U.S. Department of Commerce has ongoing Commercial Dialogues, many of which have working groups focused on standards and conformity assessment. Including industry in these dialogues when discussing ways in which standards collaboration on AI could take place in these venues could be useful. The U.S. Department of State also has standing ICT dialogues with various countries. We encourage industry participation in discussions about standards collaboration in these fora, with the ultimate objective of bringing conversations to international SDOs. Additionally, it may be useful to discuss ways in which different countries might be able to collaborate in multilateral fora, such as the Subcommittee on Standards and Conformance in APEC.

***

Once again, we appreciate the opportunity to provide feedback to NIST on its various taskings stemming from the AI Executive Order. We believe that a generative AI RMF or profile, standards and guidelines for AI red-teaming and model evaluation, and a plan to engage in and advance the development of international standards are all integral to advancing the development and deployment of trustworthy AI. We look forward to continuing to collaborate with NIST as they progress these efforts and encourage NIST to view ITI as a resource. Please feel free to reach out to Courtney Lang (Clang@itic.org) with any questions.