Convening stakeholders across industries to craft principles and concrete codes of practice for the development and use of artificial intelligence.

February 2nd, 2024

Dr. Laurie Locascio
Under Secretary of Commerce and Technology
Director of the National Institute of Standards and Technology (NIST)
Submitted electronically to www.regulations.gov

RE: NIST–2023–0309, Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

These comments are submitted on behalf of the Alliance for Trust in AI (the Alliance), a nonprofit association of companies using artificial intelligence (AI) representing diverse sectors. Members of the Alliance seek to ensure that AI can be a trusted tool by promoting effective policy and clear codes of practice for AI. We appreciate the opportunity to respond to the National Institute of Standards and Technology's (NIST) Request for Comment Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (AI EO).

The Alliance is appreciative of the work that NIST has done to introduce common language and practices to design, develop, use, and evaluate AI. We hope that NIST will continue to keep in mind the broad diversity of technologies, contexts, and uses of AI. In particular, the Alliance encourages NIST to ensure that guidelines, standards, and best practices are risk-based, broadly applicable across industries, and designed to be flexible over time and contexts. Additionally, the Alliance urges NIST to keep in mind the successful risk-based and technology-agnostic approaches already in regulation and industry best practices, and to complement them rather than seek to replace them as more comprehensive guidelines, standards, and best practices are established.

About the Alliance

The Alliance for Trust in AI brings together companies using advanced AI in many sectors to advocate for ways that we can build trust in all the kinds of AI that empower companies across the country and world. The Alliance works with companies developing foundational AI models, creating AI systems, and implementing these systems and models in their own work across industries.

We aim to give organizations concrete guidance around how to build AI responsibly, implement AI principles, support learning and information sharing across sectors, and establish a shared voice for the many users of AI. The Alliance is building on work done by technologists, policymakers, and academics to create a shared understanding of how to develop and use AI responsibly. Through multi-stakeholder partnership with members across industries and sectors, the Alliance is developing definitions, principles, and codes of practice that allow developers, implementers, and users of AI systems to demonstrate responsibility and accountability.

Realizing AI opportunity requires shared principles, developed across sectors

It is clear that AI offers incredible opportunities. Organizations in every sector are using AI in simple and profound ways, automating repetitive tasks, optimizing their work, making progress on social challenges, and rooting out sophisticated fraud and cyberthreats. Using data to train algorithms is not new, but the pace of technology's ability to harness massive datasets has accelerated dramatically, drawing questions around the impact on industry and society.

Organizations across sectors have established their own AI governance with shared goals and common approaches, building upon NIST's risk management practices. Companies using AI in manufacturing or agriculture are facing surprisingly similar questions as companies using AI in marketing or transportation. The high level principles they use to reason about AI governance and tools can help other organizations develop, deploy, and maintain AI systems in ways that increase trust.

Together, the Alliance's members have brought together a set of high-level principles around development, implementation, and use of AI based on these experiences. This document, also attached to these comments, compiles AI principles and best practices from across sectors that reflect existing national, international, and standards

frameworks. These principles are designed to evolve over time and be flexible across contexts, and include several categories of principles:[1]

- Our **governance principles** seek to reduce the potential harms of AI by implementing various guardrails and governance strategies for AI development and use. Accountability, transparency, explainability, and oversight are all core elements of effective governance for AI, regardless of whether it is simple or complex, narrow or broad.
- Our **data principles** focus on ensuring that data is used responsibly to empower AI models and systems. Quality, representative data is vital to the training of AI models, particularly for ensuring reasonable and appropriate outcomes for automated decision making impacting people. Privacy should be considered throughout AI model development and implementation to ensure appropriate data use.
- Our **society principles** call for alignment between our values regarding AI,the purposes of AI models, and their implementation. The principles discuss preventing unintentional bias, developing AI consistent with organizational and societal values, and the responsibility of those developing and deploying AI systems.
- Our **safety principles** focus on reducing risk and maintaining the resilience and security of AI models by anticipating, preventing, and mitigating incidents that they may cause. They also take measures to ensure that systems work as intended. Security, accuracy, robustness, and validity are all important components of safety.
- Our **implementation principles** consider the context in which an AI system is used, with ongoing development and risk assessment as appropriate. This includes ensuring appropriateness, considering scalability and adaptability, and creating support and feedback mechanisms.

We have existing tools to regulate and govern AI, including advanced and frontier AI. Technology-neutral laws and regulations apply, and many sectors (e.g., financial, health, and employment) have advanced programs to ensure compliance and manage potential unintended bias. As NIST works to further develop guidelines, standards, and best practices for AI, the Alliance hopes that you will look to the sectors that have already developed similar practices for governance by necessity. While each sector approaches these questions from different perspectives, there are many commonalities that can help inform NIST's work.

[1] Principles, Alliance for Trust in AI, https://alliancefortrustinai.org/principles/

NIST responsibilities under the EO

1. ***Developing cross-sectoral guidelines, standards, and best practices for AI safety and security is a core responsibility***

NIST is seeking information regarding topics related to generative AI risk management, AI evaluation, and red-teaming.

The Alliance deeply appreciates NIST's long focus on risk and risk management across all kinds of technologies, and we are supportive of its efforts to broaden AI guidelines, standards, and best practices. While AI feels novel, it has been developed and used in many contexts for decades. Even emerging AI technologies have been in use for years, including in voice assistants or for advanced research into pharmaceuticals. Being at the forefront of AI innovation, the organizations that developed these tools have significant expertise in responsible AI and AI governance. As NIST continues its AI-related initiatives, it stands to benefit from this expertise.

a. *Standards should reflect the real-world breadth of AI use and focus on real-world risk*

Artificial Intelligence (AI) includes all sorts of techniques, models, and use cases. For simple models, AI can help to detect fraud in financial transactions, suggest a route for your drive, and provide automated customer support. For more advanced models, AI can perform advanced drug discovery research, drive autonomous robots, and optimize global logistics. While these are all distinct uses for AI, they have a few things in common: they're driven by data, machine learning, and computing power. While AI is a term for algorithms performing these types of tasks, public discourse often conflates it with specific types of AI such as generative AI or Artificial General Intelligence (AGI). The Alliance recently published a post to help explain the breadth of use in plain language.[2]

As NIST works to develop additional guidelines, standards, and best practices, we suggest ensuring that this work reflects the broad set of AI tools, techniques, and uses. Different kinds of AI, and different use cases for AI, will require context and risk aware regulation and governance. Guidance should be consistent across AI techniques whenever possible and should not be biased against one type (i.e., biased against generative AI). Using illustrative examples for standards, and creating crosswalks

[2] Coming to Common Terms on Artificial Intelligence, Alliance for Trust in AI, https://alliancefortrustinai.org/coming-to-common-terms-on-artificial-intelligence/

comparing how they apply to different kinds of AI, may be useful as organizations look to implement them.

Although policymakers have paid much attention to the risk of frontier models, the UK AI Safety Summit *Roundtable on Risks from Loss of Control over Frontier AI* concluded that "current models do not present an existential risk."[3] Moreover, members of the discussion argued that "it is unclear whether we could ever develop systems that would substantially evade human oversight and control." Rather than focusing on hypothetical scenarios like these, NIST should work to address current and near-future considerations and risks for the AI systems we are developing today, with flexibility for uses we have not envisioned yet.

b. Effective standards for AI must be risk-based, context driven, and designed to evolve over time

In order to be effective while allowing innovation, AI guidelines, rules, and standards need to be risk-based, context driven, and designed to evolve over time. Different kinds of AI may require different safeguards, but AI standards should be generally consistent. Risk management approaches, like regulation, should be technology-agnostic and apply to a broad range of AI whenever possible. Different sectors can learn from each others' guidelines, standards, and best practices and apply them in their own context. NIST should ensure that guidance, best practices, and standards can be appropriately tailored to an organization's role and use of AI while adhering to a broader and shared set of guidelines.

Risk is not based on the size of the model, but rather how it is used, and its capability within that context. The amount of computing power or data used to develop an AI model or system does not necessarily correlate to the level of risk posed in the abstract or in practice. Risk evaluation should draw from existing practices and knowledge from sectors that are using AI with effective governance and risk management.

Self-governance, responsible practices, and diligence should always be a part of the designing, developing, deploying, and using AI. Different actors (i.e., developers, deployers, end users) have different responsibilities appropriate to their role in the ecosystem, and throughout the lifecycle of AI development and deployment. Safeguards and guardrails should be built in at appropriate points in the system's development and

[3] AI Safety Summit 2023: Roundtable Chairs' Summaries, 1 November, https://www.gov.uk/government/publications/ai-safety-summit-1-november-roundtable-chairs-summaries/ai-safety-summit-2023-roundtable-chairs-summaries-1-november--2

deployment; there is no one size fits all approach. Developers will have different responsibilities and will manage different risks than deployers or end users.

c. *Use existing consensus to support standards that work for diverse organizations and industries*

Best practices for the use of AI have common principles are pragmatic, mitigate risk, and implement pragmatic accountability mechanisms across many development, business, and deployment models. Standards should not unduly favor any of these approaches, but should be broadly applicable and achievable by organizations of many sizes and with diverse approaches to AI. By focusing on contextualized risks, explainability, and measurement, NIST can avoid picking winners and losers. Instead, it can enable all developers and users of AI to appropriately assess and manage risk while building products, services, and tools that benefit the ecosystem.

d. *Be clear about the goals and scope of red-teaming, evaluation, and testing standards*

Testing and evaluation are critical steps in the design, development, deployment, and use of AI systems. Developing testing and red-teaming standards is one of the most helpful things that NIST and other global standards organizations can do, especially with increased attention on both the positive benefits and negative risks of AI use.

However, the term “red-teaming” can be confusing in this context, given its existing definition in cybersecurity testing. As NIST develops guidance and best practices for testing and red-teaming as defined in EO 14110, it will be important to specifically and clearly define the risks, goals, and purpose of red-teaming and associated testing.

e. *Standards should caution against using AI in unintended contexts*

It is easy to assume that AI works the same way that our brains do, but they do not. AI systems are not as broadly capable as human thought are and typically suited to a particular set of tasks. More data does not necessarily create a better or more capable AI model, and broad models are not necessarily better suited to a given task. Capability, risk, and benefit must all be evaluated in context as AI is developed, deployed, and used; each context bears different considerations.

At its core, AI is computing: it does what you tell it to do - so you need to know how to tell it to do what you want it to do. As guidelines are developed, it will be important to include measures to help evaluate how, and when, AI is appropriately deployed and for what kinds of tasks. Sometimes a set of tasks is quite narrow, and sometimes broader,

but AI is likely to fail when attempting tasks for which they are not trained or suited. Using AI in high risk or high impact contexts requires additional diligence.

f. Explainability, documentation, and transparency are necessary elements as AI is developed and deployed

Explainable AI systems incorporate ways for people to understand and trust their outputs. Transparency and explainability measures allow for internal and external oversight, including the data used, the system's training, and its usage. This information should include the kinds of training data, source of training data, training methods, and what the system is trained to do. The information should also be tailored to the audience for which it is intended.

While transparency around the use of AI is important, it should not overwhelm end users, many of whom may not have deep technical backgrounds. To improve trust in AI, NIST should consider context, intended purpose, and outcome as thresholds for transparency.

Documents and descriptions supporting the explainability for AI needs to be tailored to the audience for which they are intended. Users of the AI system, regulators, and internal developers will all need descriptions that correspond to their varying levels of technical expertise. Information also needs to be appropriate for the kinds of knowledge required to carry out their specific task.

2. Reducing the risk of synthetic content requires nuanced and context-appropriate approaches

NIST is seeking information regarding topics related to synthetic content creation, detection, labeling, and auditing.

The Alliance cautions against the implication that “synthetic content” - content that has been modified or generated by AI - is inherently more harmful, non-credible, or otherwise pose more risk than other content. Tracking, detecting, labeling, and auditing synthetic content or content creation is a complex approach that may not be the most effective way to address the potential risks of synthetic content. Any approach proposed to manage this risk should be carefully defined and scoped to focus on contextual risks, rather than attempting to solve myriad social ills such as misinformation and election manipulation through technical standards.

As a society, we long ago left behind the idea that content springs, fully formed, from someone’s head to a paper. Instead, we use many tools to help create prose, images,

and video content. From spellcheck to image-tuning, from customer service assistants to weather predictions, these tools incorporate significant use of AI-generated content, and there's no inherent reason that they are not credible or that they pose risk.

Some synthetic content may well be entirely trustworthy, and other content may be misleading (either intentionally or unintentionally). Therefore, conflating "synthetic" and risk is unproductive. Instead, exploring authentication of content from credible sources and examination of contexts where synthetic content may pose particular risk may be a better approach than attempting to ensure that all synthetic content has been watermarked or otherwise labeled.

Authenticating content and tracking its provenance may prove a more effective way to reduce risk. We already have effective ways to authenticate content as it is distributed, but we do not have ways to ensure that detection or watermarking are effective. To attempt to manage this by detecting synthetic content invites a cat-and-mouse game with adversaries and bad actors who are well equipped to mask their creation of malicious content via AI. There are many legitimate concerns around synthetic content including those mentioned in the RFI (i.e., repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses), but it is important to remember that AI is a tool. We can build in safeguards, but we cannot stop misuse of tools. Labeling and watermarking, both visible and invisible, can be removed or altered, limiting effective use cases of these techniques. Since real labels can be removed and fake labels can be added, they are not good indicators of trustworthy content.

The goal of E.O. 14110 Section 4.5(a) is to reduce the risk of synthetic content. Authenticating credible content may help do that: if people know what organization stands behind a particular piece of content, then they can judge the trustworthiness of that content. This will be more effective than attempting to detect or watermark all synthetic content, or attempting to solve the entire ecosystem of risk related to synthetic content.

3. *NIST should partner to advance shared, responsible global technical standards for AI*

NIST is seeking information regarding topics related to the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing that should be considered in the design of standards.

International and national rules should be harmonized as much as possible in order to provide a unified framework for powerful AI that can provide global benefit and leadership. NIST should promote the AI RMF and other guidelines, standards, and best practices globally in ways that partner with peers and other governments early on in order to gain consensus and buy-in early in the process and ensure that standards are seen as international rather than U.S. specific. Involving the U.S. AI Safety Institute, and partnering with other national AI Safety Institutes, may be one way to accomplish this.

Promoting the principles of the NIST AI Risk Management Framework[4] and the U.S. Government National Standards Strategy for Critical and Emerging Technology[5] internationally is a key element in ensuring that AI can be trusted across borders and is not subject to fragmented regulation and enforcement.

Before work can begin on any international or global regulation on AI, we must begin to speak the same language - at least, in terms of how we talk about AI. AI nomenclature and terminology is unclear at best, and actively contradictory at worst. Any attempts at driving consensus standards and coordination must include explicit terminology definitions as a baseline to be efficient and impactful.

Conclusion

The Alliance for Trust in AI supports NIST's work in creating the best practices, standards, and guidance that will ensure that AI can be a safe, secure, and trustworthy tool across every kind of organization. We welcome additional work to build on the NIST AI RMF, and to continue building on the strong foundation of AI and cyber and safety work. We appreciate the opportunity to discuss and provide input during this process and look forward to engagement with NIST on these and other important topics. If you have questions, or believe that we can be helpful to your work in any way, please contact the Alliance's coordinator Heather West, at hewest@venable.com.

[4] AI Risk Management Framework, NIST, https://www.nist.gov/itl/ai-risk-management-framework
[5] U.S. Government National Standards Strategy for Critical and Emerging Technology, https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf

Appendix: ATAI Principles

The Alliance for Trust in AI has brought together a set of high-level principles around development, implementation, and use of AI. These principles bring together consensus ideas and framing for AI principles and best practices that reflect existing national, international, and standards frameworks that seek to build trust in AI. These principles will vary in weight depending on context, and will evolve over time.

Governance Principles

The governance principles seek to reduce the potential harms of AI by implementing various guardrails and governance for AI development and use.

- **Accountability and governance** measures ensure that AI is developed or used under appropriate consideration. Accurate and up to date record keeping, along with clear internal and external policies, can ensure appropriate responsibility is taken.
- **Transparency and explainability** measures allow for internal and external oversight, including the data used, the system's training, and its usage.
- **Human control and oversight** measures provide appropriate oversight for high risk or high impact systems and decisions. Levels of human involvement should match the context and impact of decisions.
- **Internal policies and guidance**, along with employee education, are necessary for the safe and appropriate implementation of AI models. Publicizing these policies can further improve transparency and trust in AI models.
- **Compliance** with existing and new laws, regulations, and best practices is necessary.
- **Remediation** measures ensure there is a mechanism to mitigate or remedy harms.

Data Principles

The data principles focus on ensuring that data is used responsibly to empower AI models/systems. Quality data is vital to quality training of AI models, particularly for

assuring reasonable and appropriate outcomes through automated decisions impacting people.

- **Privacy** should be considered throughout AI model development and implementation to ensure appropriate data use.
- **Confidentiality, sensitivity**, and integrity of data should be considered when used as inputs or outputs, or in contexts where AI-based determinations impact rights and eligibility for certain opportunities, products or services are made about individuals.
- **Training data** is vital for ensuring AI systems are representative of communities. Data should be well curated, representative, and high quality for robust, accurate, and fair AI systems.

Society Principles

The society principles call for alignment between our values and the purposes to which AI models and systems are applied, how they are implemented, and who has access.

- **Fairness and equity** call for preventing unintentional bias in AI models. It involves considering appropriateness for the context, special data categories, and improving algorithmic fairness.
- **Values** reflect how AI can be developed consistent with, and furthering, contextual, organizational, and societal norms and values.
- **Responsibility** ensures that those developing and deploying AI systems anticipate their short- and long-term impacts and consult the right stakeholders.

Safety Principles

The safety principles focus on reducing risk and maintaining the resilience and security of AI models by anticipating, preventing, and mitigating incidents that they may cause. They also take measures to ensure that systems work as intended.

- **Security** involves protecting AI systems from cyber-attacks, data breaches, and other threats to the confidentiality, integrity, and availability of information and systems. AI systems present novel attack surfaces.
- **Accuracy** is the percentage of correct outputs or predictions based on given inputs or data, which may or may not correlate with truthful output.
- **Robustness** is the ability of an AI system to remain accurate in different settings and conditions.
- **Validity** measures ensure that AI models perform as intended.

- **Quality assurance** measures test with different datasets, environments, and scenarios to evaluate their performance over time.

Implementation Principles

The implementation principles look to considerations for the context in which an AI system is used, with ongoing development and risk assessment as appropriate.

- **Appropriateness** should be verified to ensure that AI systems are suitable for a given task before use in that context.
- **Scalability and adaptability** should be considered to ensure that the system can scale to meet increasing demand and adapt to changing circumstances or requirements, and that they are appropriately resourced.
- **User support**, training, and documentation should be provided to assist and educate.
- **Feedback mechanisms** for users and stakeholders can inform product and functionality improvements.