SIIA

June 2, 2024

Via Email Submission

Re: Request for Comments on Draft Documents Responsive to NIST's Assignments Under Executive Order 14110 (Sections 4.1, 4.5, and 11) – Docket No. NIST-2024-0001

Dear National Institute of Standards and Technology:

The Software & Information Industry Association (SIIA) appreciates the opportunity to provide comments on the draft documents in this docket. Our comments focus on NIST AI 600-1, the initial public draft of *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, and NIST AI 100-5, *A Plan for Global Engagement on AI Standards*.[1]

SIIA is the principal trade association for companies in the business of information. Our members include nearly 400 companies reflecting the broad and diverse landscape of digital content providers and users in academic publishing, education technology, and financial information, along with creators of software and platforms used by millions worldwide, and companies specializing in data analytics and information services. Our membership includes upstream and downstream AI designers, developers, and deployers of AI systems in myriad environments.

Recommendations on NIST AI 600-1

Draft NIST AI 600-1 reflects close engagement with the many challenges raised by managing risks associated with generative artificial intelligence (GAI) and we believe it serves as a strong foundation for continued work to produce a GAI RMF profile that will have enduring value for organizations across the AI value chain and for governments around the world. As a member of the NIST AI Safety Institute Consortium, SIIA is pleased to participate in the development of that product.

[1] NIST, NIST AI 600-1 (Initial Public Draft): *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (Apr. 2024) https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf ("Draft AI 600-1"); NIST, AI 100-5 (Draft for Public Comment): *A Plan for Global Engagement on AI Standards* (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.100-5.Global-Plan.ipd.pdf ("Draft AI 100-5").

Recommendation 1: Clarify the Scope of GAI Covered by the Profile

The introduction to Draft AI 600-1 relies on the definition of GAI contained in EO 14110 and notes that the document uses GAI generally to apply to "dual-use foundation models" as that term is defined in EO 14110.[2] Given this, we recommend that NIST further clarify the scope of this profile in two ways.

First, we believe clarification is necessary to reflect the draft's acknowledgement that "not all GAI is based in foundation models." Many actions in Draft AI 600-1 are designed to address foundation model-based GAI and are less relevant for other forms of GAI. We discuss this issue in more depth in our recommendation that NIST incorporate a risk-based assessment of GAI into the profile.

Second, as we have explained in a recent submission to NTIA, we question whether EO 14110's definition of "dual-use foundation model" is intended to encompass all foundation models.[3] We respectfully refer you to our comments in that submission, which explains why there is value in focusing the scope of the terms as used in EO 14110.

Recommendation 2: Rely on the AI RMF Approach to Identifying Potential Risks

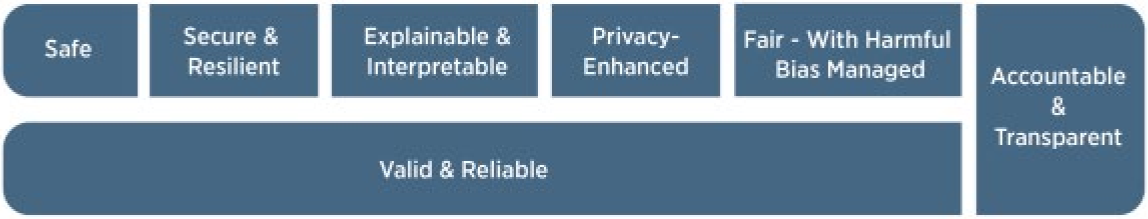
The AI RMF has emerged as a cornerstone for ongoing AI risk management and governance in the United States and globally. One reason it has resonated is because the framework was "designed to address new risks as they emerge," a "flexibility [that] is particularly important where impacts are not easily foreseeable and applications are evolving."[4]

[2] Draft NIST AI 600-1 at 1, note 1.

[3] SIIA, Response to NTIA's Request for Comment Regarding Dual Use Foundational Artificial Intelligence Models with Widely Available Model Weights (Mar. 27, 2024), https://www.siia.net/wp-content/uploads/2024/04/SIIA-Response-to-NTIA-on-AI-Open-Model-Weights.pdf ("SIIA Open Weights"), at 2-3.

[4] NIST, NIST AI 100-1: Artificial Intelligence Risk Management Framework (AI RMF 1.0) (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, at 4 (defining "risk" as "the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event"). The AI RMF further explains: "In the context of the AI RMF, risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet."

Following this approach, the AI RMF does not contain a comprehensive list of risks associated with AI systems. Rather, it adopts an innovative approach for categorizing risk based on the characteristics of trustworthy AI systems, or "managing risks in pursuit of AI trustworthiness." Those characteristics are reflected in this graphic from the AI RMF:


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

While we appreciate the careful discussion of risk categories in Draft NIST AI 600-1, we believe the guidance would be strengthened by embedding these risks within the same schema reflected in the AI RMF. This approach has several benefits. First, it will assist stakeholders in mapping risks to the core characteristics of trustworthy AI. These characteristics serve as a frame of reference to discuss AI systems of all types, including GAI. The list of twelve risks in Draft NIST AI 600-1 seems to recommend a different approach to managing AI risk, when in fact eleven of the twelve fit within one or more of the trustworthy AI characteristics.[5] For example:

Trustworthy AI Characteristic	**Risk Identified in Draft NIST 600-1**
Valid & Reliable	Information Integrity Confabulation
Safe	CBRN Information Dangerous or Violent Recommendations Information Security Obscene, Degrading, and/or Abusive Content
Secure & Resilient	Information Security Intellectual Property Value Chain and Component Integration
Explainable & Interpretable	Human-AI Configuration Information Integrity Confabulation
Privacy-Enhanced	Data Privacy

[5] The outlier is "Environmental." We recommend that NIST address this class of risks outside of the trustworthy AI rubric.

Fair - With Harmful Bias Managed	Human-AI Configuration Obscene, Degrading, and/or Abusive Content Toxicity, Bias, and Homogenization
Accountable & Transparent	Intellectual Property Dangerous or Violent Recommendations Obscene, Degrading, and/or Abusive Content Toxicity, Bias, and Homogenization Value Chain and Component Integration

Second, supplementing the AI RMF approach will avoid a need to regularly update the risk category rubric to account for new research and developments. This includes scientific and technical research into explainability, potential mitigations, and unanticipated uses of GAI. It also includes legal and regulatory developments that affect the environment in which GAI systems are developed and deployed. For example, federal courts and the U.S. Copyright Office are grappling with many of the issues raised in Section 2.9 of the framework, and the outcomes of this ongoing and iterative process will have an impact on GAI risk mitigation.

Third, aligning risks associated with GAI in the AI RMF scheme will facilitate coordination with AI Safety Institutes and governments across the world. For example, consider NIST's crosswalk of the AI RMF and Japan's AI Guidelines for Business. This crosswalk is one of many efforts to coordinate terminology and align approaches to advance interoperability and understanding of AI governance across jurisdictions. As the AI Safety Institutes build on the agreements reached at the A Seoul Summit, working towards concrete guidance ahead of the next summit in February 2025, we believe the work of NIST should be foundational to advance "interoperability between AI governance frameworks" as stated in the leaders' declaration.[6] Building out guidance within the AI RMF rather than offering a new approach for a class of AI is likely to help to achieve this goal.[7]

[6] https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-declaration-for-safe-innovative-and-inclusive-ai-by-participants-attending-the-leaders-session-ai-seoul-summit-21-may-2024.

[7] This approach may help in fostering a common framework to connect NIST's work with other important efforts in the global AI safety and security community, such as the recently issued *International Scientific Report on the Safety of Advanced AI* (May 2024), https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

Recommendation 3: Incorporate a Risk-Based Assessment of GAI Systems

We recommend revising Draft NIST AI 600-1 to incorporate an approach to assessing the risk of GAI systems based on characteristics of those systems, and then tailoring the risk management actions to the risk profile of the GAI system at issue.

In our recent submission to NTIA on its request for input regarding dual-use foundation models with widely available weights,[8] we addressed the fundamental need to conduct a risk assessment of these models across the level-of-access gradient, from fully closed models to those that provide one or more of the following: hosted access, API access, API access for fine tuning, access to weights, access to training data, access to code with use restrictions, access to features without restrictions, and so on.[9] We urged NTIA to defer to NIST on developing an AI RMF use-case profile for GAI to guide their assessment of openness, noting, among other things, that certain "risks can be mitigated through various measures, including staged release; less than fully open access; limitations on who can access the weights (e.g., through license and user restrictions); and limitations on how the assets can be used (E.g., use restrictions and contract terms)."[10] We cautioned "against a one-size-fits-all approach to mitigating risks for open models due to the gradient of openness, the differences among models, and differences around model training data. In addition, advances in foundation models, risk mitigation techniques (TEVV, auditing, red-teaming, and so on) and the capabilities of bad actors mean that any approach must be sufficiently flexible and agile to adapt."[11]

Our submission to NTIA considered just one set of characteristics of certain GAI systems – openness. There are many other characteristics of GAI systems that will bear on the risks associated with those systems, and factor into the actions of various actors involved in the design, development, and deployment of GAI systems.[12] These include, for example, the system's architecture (generative adversarial networks, variational autoencoders, autoregressive models, diffusion models, transformer-based models, and so forth), training mechanism, applications, training libraries, computational requirements, and other features.

[8] SIIA Open Weights, supra note 3.

[9] This analysis built in part on Rishi Bommasani and Sayash Kapoor, et.al, Stanford University Human-Centered Artificial Intelligence, Issue Brief: Considerations for Governing Open Foundation Models, (Dec. 13, 2023), https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

[10] SIIA Open Weights at 8.

[11] SIIA Open Weights at 9.

[12] https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A.

Moreover, GAI has been in use for many years, long before the general public began experimenting with chatbots. Several of the recommended actions appear designed to address GAI tools used by the general public rather than those custom-tailored to specific business applications.

We believe the usability and resilience of AI 600-1 would benefit by beginning with a profile of GAI systems based on key characteristics and how they may impact the pursuit of trustworthy AI characteristics. Aligning that profile to the AI RFM Core based on different categories of GAI models and their risk profiles would contribute to organizations' ability to apply NIST's guidance to mitigate risk and also provide the broader community, including international forums and foreign governments, with a robust roadmap for advancing trustworthy GAI.

Recommendation 4: Delineate AI Actor Responsibility

Although Draft AI 600-1 identifies relevant AI actors at the bottom of each action table, in most cases it does not distinguish between actors situated at different points across the GAI value chain. This will create a challenge for translating AI 600-1 for differently situated actors, such as developers of GAI models and third parties that use or adapt those models. Moreover, a deployer of a custom-built GAI model that is used internally at an organization does not confront the same set of risks as a deployer model intended to be used generally by the public.

Part of this challenge may stem from the way in which NIST has defined AI actors based on *task* rather than based on both task *and posture in the value chain*.[13] For example, "Governance and Oversight" tasks cover organizations involved in design, development, and deployment of AI systems. Several of the actions identified in the tables, such as those around end user disclosures and downstream monitoring, would be infeasible for system designers; and others, such as documenting data sources, would be infeasible for third-party deployers.

To address this, we recommend that NIST provide clarity about which action items should be undertaken by which actors, with particular attention to different responsibilities for actors positioned differently across the AI value chain. We also recommend that NIST consider augmenting the "Descriptions of AI Actor Tasks" to distinguish among different AI actors undertaking these tasks. These steps would help to align the action tables with the intent set out in the draft, which states that "not all actions apply to all AI actors. For example, not [all] actions relevant to GAI developers may be relevant to GAI deployers."[14]

[13] See https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A.

[14] Draft AI 600-1 at 11.

Recommendation 5: Calibrate Actions Based on the Risk Profile of Different GAI Systems and Provide Clarity that Not All Actions Are Necessary for All GAI Systems

We agree that "[o]rganizations should prioritize actions based on their unique situations and context for using GAI applications," as the draft states in preface to the action tables. Yet we are concerned that this approach runs counter to the line that follows: "Some subcategories in the action tables below are marked as 'foundational,' meaning they should be treated as fundamental tasks for GAI risk management and should be considered as the minimum set of actions to be taken."[15] Over 315 action items are marked in this way.

Not all of the over 315 action items marked as foundational should be required for all GAI systems and all GAI actors. In addition to distinguishing more clearly between those actions relevant to developers versus deployers, actions considered *foundational* best practices should also be calibrated to the risk profile of the GAI system, as discussed above.

We recommend that NIST provide further guidance to help organizations prioritize the action items both in general and as adjusted to the risk profiles of different GAI systems. In addition, NIST should consider whether certain of the "foundational" action items would pose an unreasonable burden on small- and medium-sized enterprises disproportionate to the intended impact of those items.

Recommendation 6: Balance Recommended Actions Against Countervailing Legal, Ethics, Security, and Technical Concerns, Limitations, and Trade-Offs

Achieving the characteristics of trustworthy AI in GAI systems is, as NIST notes, an art rather than a science. Part of the art requires balancing competing interests. Those competing objectives may involve legal compliance, ethical AI best practices, safety and security measures, and technological limitations.

As the action items are refined, we recommend that NIST consider the following:

- *Balancing transparency with restrictions on data sharing, intellectual property protection, and security needs.*

Several of the recommended actions are designed to increase transparency through disclosure of information and sharing of data. For example, MP-4.1-012 provides as follows:

> "Implement reproducibility techniques, including: share data publicly or privately using license and citation; develop code according to standard software practices; track and document experiments and results; manage the software

[15] Draft AI 600-1 at 11.

> environment and dependencies; utilize virtual environments, version control, and maintain a requirements document; manage models and artifacts; tracking AI model versions and documenting model details along with parameters and experimental results; document data management processes and establish a testing/validation process to maintain reliable results"

This action, described as "foundational" for a range of AI actors, raises potential concerns. Sharing data reflecting reproducibility techniques would create a security risk and in addition could expose valuable intellectual property that those with access to the data could exploit.

In the appendix to this submission, we flag some of the additional action items that raise concerns of this nature.

- *Balancing content provenance recommendations against limitations of current content provenance technologies and lack of standardization.*

Draft AI 100-4 detailed overview of technical approaches to synthetic content detection and provenance data tracking, including a discussion of "ongoing research and related research gaps," limitations in current technology and science, the lack of international standards, and more. It recognizes that "[t]here is no perfect solution to solve the issue of public trust and harms stemming from digital content" but recognizes that improvements in provenance, detection, labeling, and authentication can advance trust in GAI.[16]

We believe that these techniques hold significant promise for the integrity of the information ecosystem and building trust in AI tools, and as an association have long advocated for increased attention to content provenance.[17] Yet in the context of Draft AI 600-1, we have some concern that the many action items related to content provenance are somewhat at odds with the ongoing work, reflected in Draft AI 100-4, to improve provenance data tracking and synthetic content detection. We believe that the state of the technology requires additional time to improve technologies and develop standardized approaches before any one, imperfect approach is endorsed over potentially better solutions – or solutions that work better in particular contexts, for particular GAI risk profiles. We recommend that NIST reexamine the content provenance action items to reflect this uncertainty.

[16] Draft AI 100-4, at 1-2.

[17] See, e.g., https://cdt.org/event/cdt-siia-democracy-affirming-technology-restoring-trust-online/; https://www.siia.net/wp-content/uploads/2021/11/Bipartisan_Deepfake-Task-Force-Act-Inclusion-in-FY22-NDAA_SupportLetter.pdf; https://www.siia.net/wp-content/uploads/2022/03/SIIA-Submission-for-National-AI-Strategic-Plan.pdf.

- *Balancing audit recommendations against limitations of current auditing capabilities and procedures and lack of standardization.*

Several action items recommend undertaking third-party audits. Auditing of GAI systems (and AI systems more broadly) remains an emerging area, without formal standardization of tools and methods. We recommend that NIST carefully examine these items to ensure that they are feasible, and narrowly tailored to avoid creating risks that may arise, for example, from disclosure of intellectual property in datasets or AI models, sensitive data included in training datasets. These risks, coupled with still-developing art for undertaking audits, are among the reasons that third-party audits are not yet required for most AI systems. To that end, we further recommend that NIST caveat its recommended action items regarding audits as suggested actions rather than as foundational, “must have” actions.

Recommendation 7: Include Guidance on Best Practices for Small- and Medium-Sized AI Actors

As noted above, the full set of recommended action items will be burdensome on some small- and medium-sized AI actors. This may be most pronounced in recommended actions around content provenance and synthetic content detection and tracking. We recommend that NIST consider clarifying its guidance for these AI actors in a way that advances the goals of trustworthy AI while also mitigating the risks related to their role in the AI value chain.

Comments on NIST AI 100-5

SIIA strongly supports NIST’s plans for global engagement on AI standards. We also recognize that NIST has already undertaken significant efforts towards implementation of portions of this plan and applaud its proactive approach.

We appreciate how NIST has organized the priority topics for standardization work in Section 4 of the NIST AI 100-5 draft. In particular, the set of urgent priority topics in Section 4.1 aligns with our understanding based on feedback from our membership and discussions in the broader policy and research communities.

One area that would benefit from increased attention involves methods to increase domestic capacity-building. We provided recommendations on this topic in our submission to the NSSCET, and would reprise those here.[18] We recognize that certain suggestions, such as creating a grant program to allow for small- and medium-sized enterprises to participate directly in international standards meetings, may require appropriations and/or authorizations beyond those currently available to NIST.

[18] https://www.siia.net/wp-content/uploads/2023/11/SIIA-Comments-on-NSSCET-RFI.pdf

*   *   *

Thank you for the opportunity to engage in this consultation. Please reach out to me with any questions or comments. SIIA looks forward to continuing to work with NIST to develop risk management guidance for GAI and on other critical issues as part of the AI Safety Institute Consortium.

Respectfully,

Paul Lekas
Senior Vice President
Head of Public Policy and Government Affairs
Software & Information Industry Association
plekas@siia.net

APPENDIX: Specific Action Items in Draft AI 600-1

Following are specific action items in Draft AI 600-1 that we would recommend further refining to address comments raised in this submission. These action items, to varying degrees, would benefit from revision to:

- ensue technical feasibility;
- address ambiguities in the legal framework governing training data, output data, and AI models, such as around data privacy and application of copyright law;
- reflect the lack of standardization in emerging areas of transparency such as content provenance and third-party audits;
- balance risk with implementation burden; and
- acknowledge the inherent ambiguity in defining certain categories of problematic content and rendering actions for data and models.

GV-1.1-001	*GV-1.2-005*	*GV-1.2-006*
GV-1.2-007	*GV-1.3-001*	*GV-1.5-001*
GV-1.5-003	*GV-1.5-006*	*GV-1.5-007*
GOVERN 1.7	*GV-2.1-003*	*GV-3.2-001*
GV-3.2-006	*GV-4.3-001*	*GV-4.3-004*
GV-5.1-003	*GV-6.1-001*	*GV-6.1-003*
GV-6.1-012	*GV-6.1-013*	*GV-6.1-014*
GV-6.1-015	*MP-1.1-006*	*MP-2.3-001*
MP-2.3-002	*MP-2.3-004*	*MP-2.3-005*
MP-2.3-008	*MP-4.1-004*	*MP-4.1-007*
MP-4.1-009	*MP-4.1-012*	*MS-1.1-017*
MS-1.1-018	*MS-1.1-019*	*MS-1.3-010*
MS-2.2-006	*MS-2.5-009*	*MS-2.6-002*
MS-2.6-003	*MS-2.6-010*	*MS-2.7-021*
MS-2.8-001	*MS-2.8-011*	*MS-2.8-013*
MS-2.9-003	*MS-2.10-006*	*MS-2.10-014*
MS-2.11-003	*MS-2.11-006*	*MS-2.11-007*
MS-2.12-005	*MANAGE 2.4*	*MG-3.1-007*

This is not a comprehensive list of action items we believe would benefit from further tailoring. It also does not reflect the recommendation to more precisely align action items with specific actors in the GAI value chain, which applies across the action tables.