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

Point of Contact: Vince Minerva, Member of the Public

February 1, 2024

Recommendation:

Leveraging proven federal government strategies for information security provides a credible path for establishing consensus guidelines and best practices for development and deployment of safe, secure, and trustworthy Artificial Intelligence (AI) systems. Developing a companion resource to the AI Risk Management Framework (AI RMF) applicable to generative AI and AI systems in general is recommended.

Creating actionable safety, security, and trustworthiness requirements for AI systems analogous to NIST Special Publication (SP) 800-171r2, Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations (Ref. 1), provides necessary technical leadership for the AI community. NIST SP 800-171 provides information security requirements applicable to all nonfederal systems and organizations that process, store, and/or transmit controlled unclassified information. NIST SP 800-171 is intended for use by federal agencies in contractual vehicles or other agreements established between those agencies and nonfederal organizations (Ref. 1, p. iii). For example, the Department of Defense (DoD) requires a significant percentage of the 100,000 + (Ref. 2) Defense Industrial Base (DIB) companies to comply with NIST SP 800-171 via contractual requirements which benefit our national security by safeguarding sensitive information. The DoD is in the process of further building on NIST SP 800-171r2 with the proposed Cybersecurity Maturity Model Certification (CMMC) Program as stated in the Federal Register: "DoD is proposing to establish requirements for a comprehensive and scalable assessment mechanism to ensure defense contractors and subcontractors have, as part of the Cybersecurity Maturity Model Certification (CMMC) Program, implemented required security measures to expand application of existing security requirements for Federal Contract Information (FCI) and add new Controlled Unclassified Information (CUI) security requirements for certain priority programs." (Ref. 3, p. 89058). The expected number of entities affected by the CMMC program is shown in Figure 1 below.

Table 3 - Estimated Number of Entities by Type and Level

Assessment Level	Small	Other than Small	Total	Percent
Level 1 Self-Assessment	103,010	36,191	139,201	63%
Level 2 Self-Assessment	2,961	1,039	4,000	2%
Level 2 Certification Assessment	56,689	19,909	76,598	35%
Level 3 Certification Assessment	1,327	160	1,487	1%
Total	**163,987**	**57,299**	**221,286**	100%
Percent	74%	26%	100%	

Figure 1 Entities Affected by DoD CMMC Program (Ref. 3, p. 89085)

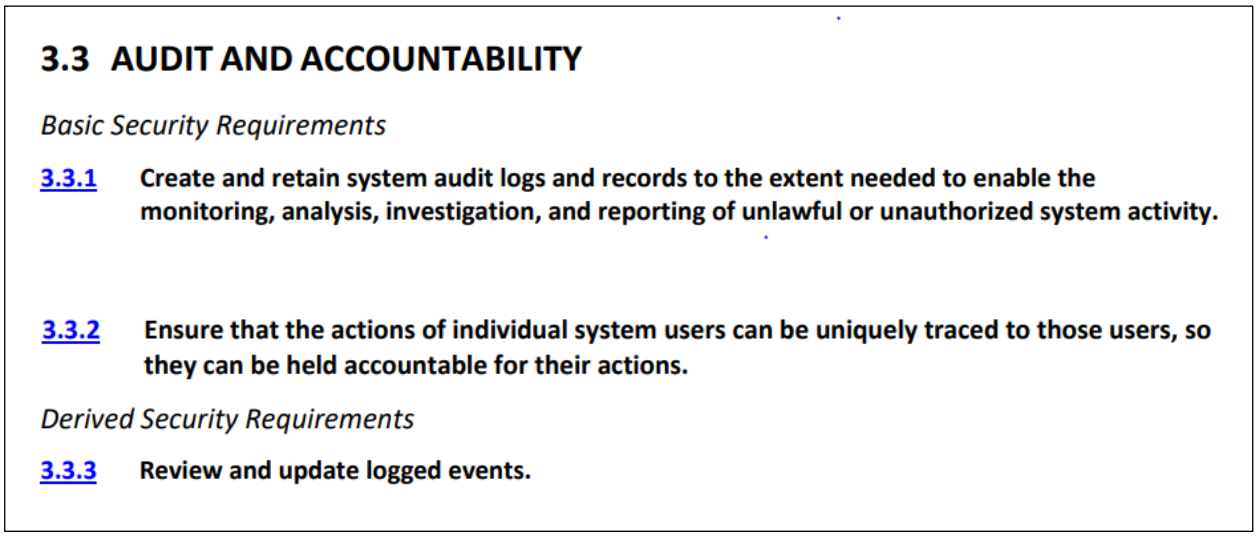
Similarly, creating a NIST publication of requirements for safe, secure, and trustworthy AI systems should strengthen our national AI capabilities. Many of the use cases for Generative AI require large scale computing resources, which will likely result in most of the public and private industry users acquiring services from commercial providers. Many, if not most, of the cloud and enterprise Information Technology (IT) providers with financial and IT system resources sufficient to run large language models are contractors to the federal government. The federal government's ability to influence the safety, security, and trustworthiness of AI systems should be increased similarly to the demonstrated successes in cybersecurity.

Developing Guidance for Safe, Secure, and Trustworthy AI Systems

NIST SP 800-171 security requirements are organized into fourteen (14) security families containing basic and derived requirements as shown in Figures 2 and 3 below. Each of the requirements has a discussion section providing additional information to facilitate requirement implementation and assessment.

FAMILY	FAMILY
Access Control	Media Protection
Awareness and Training	Personnel Security
Audit and Accountability	Physical Protection
Configuration Management	Risk Assessment
Identification and Authentication	Security Assessment
Incident Response	System and Communications Protection
Maintenance	System and Information Integrity

Figure 2: NIST SP 800-171 Security Requirements Families (Ref. 1, p. 7)



3.3 AUDIT AND ACCOUNTABILITY

Basic Security Requirements

3.3.1 **Create and retain system audit logs and records to the extent needed to enable the monitoring, analysis, investigation, and reporting of unlawful or unauthorized system activity.**

3.3.2 **Ensure that the actions of individual system users can be uniquely traced to those users, so they can be held accountable for their actions.**

Derived Security Requirements

3.3.3 **Review and update logged events.**

Figure 3: NIST SP 800-171 Basic and Derived Requirements Examples (Ref. 1, pp. 17-18)

Applying this approach to AI Systems leads to candidate requirement families as shown in Table 1 below. The seventeen (17) candidate requirement families were primarily derived from the NIST AI RMF 1.0 and the AI Bill of Rights. However, the process of formalizing the requirement families enables incorporation of concepts from a diverse set of resources, such as, the Organization for Economic Co-operation and Development (OECD) AI Principles (Ref. 4). Once the requirement families are defined through proven review and comment processes, they can be defined by basic and derived requirements. Note: The Initial Public Draft (IPD) for NIST SP 800-171r3 eliminates the distinction between basic and derived requirements which may be appropriate for the proposed document.

The candidate requirement families are likely contributors to the breadth of NIST AI responsibilities beyond the specific request of this RFI. Table 2 highlights the likely contributions of the proposed document to the specific topics in this RFI.

Safe, Secure, and Trustworthy Requirement Family	**Summary Objective**
Accountability	"This is the fundamental need: to ensure that machines remain subject to effective oversight by people and the people who design and operate machines remain accountable to everyone else. In short, we must always ensure that AI remains under human control." Ref. 5, p. 4
Accuracy	"Accuracy is defined by ISO/IEC TS 5723:2022 as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." Measures of accuracy should consider computational-centric measures (e.g., false positive

	and false negative rates), human-AI teaming, and demonstrate external validity (generalizable beyond the training conditions).” Ref. 6, p. 14; AI systems should provide a confidence level for predictions.
Explainability and Interpretability	*“Explainability* refers to a representation of the mechanisms underlying AI systems’ operation, whereas *interpretability* refers to the meaning of AI systems’ output in the context of their designed functional purposes.” Ref. 6, p. 16
Fairness	Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Ref. 6, p. 17
Ongoing Monitoring	Automated systems should have ongoing monitoring procedures, including recalibration procedures, in place to ensure that their performance does not fall below an acceptable level over time, based on changing real-world conditions or deployment contexts, post-deployment modification, or unexpected conditions.” Ref. 7
Planning	Develop, document, and disseminate policies, plans, and procedures necessary to implement trustworthiness requirements.
Privacy	“Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity.” Ref. 6, p. 17; “Automated systems should be designed and built with privacy protected by default”. Ref. 8
Reliable	“Reliability is defined in the same standard as the “ability of an item to perform as required, without failure, for a given time interval, under given conditions” (Source: ISO/IEC TS 5723:2022).” Ref. 6, p.13
Resiliency	“AI systems, as well as the ecosystems in which they are deployed, may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from: ISO/IEC TS 5723:2022).” Ref. 6, p. 15
Risk Management	“AI risk management offers a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, while also providing opportunities to maximize positive impacts. Addressing, documenting, and managing AI risks and potential negative impacts effectively can lead to more trustworthy AI systems.” Ref. 6, p. 4
Robustness	“Robustness or generalizability is defined as the “ability of a system to maintain its level of performance under a variety of circumstances” (Source: ISO/IEC TS 5723:2022). Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated.” Ref. 6, p. 14

Safety	"AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022)." Ref. 6, p. 14
Security	"…concerns related to the confidentiality, integrity, and availability of the system and its training and output data…" Ref. 6, p. 8
Supply Chain Risk Management	"Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues." Ref. 6, p. 24
Test, Evaluation, Verification, and Validation (TEVV)	Tasks are performed throughout the AI lifecycle that are carried out by AI actors who examine the AI system or its components or detect and remediate problems. Ref. 6, p. 35
Transparency	"Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so." Ref. 6, p. 15
Valid	"Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015)." Ref. 6, P. 13

Table 1: Candidate Safe, Secure, and Trustworthy Requirement Families

Safe, Secure, and Trustworthy Requirement Family	**Contributors to RFI Objectives**
Accountability	"Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users);" Ref. 9, RFI Topic 1
Fairness	"Human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts;" Ref. 9, RFI Topic 1
Privacy	"Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data;" Ref. 9, RFI Topic 3
Resiliency	"Resilience of techniques for labeling synthetic content to content manipulation;" Ref. 9, RFI Topic 2
Risk Management	"AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems;" Ref. 9, RFI Topic 3
Safety	"Developing Guidelines, Standards, and Best Practices for AI Safety and Security" Ref. 9, Topic 1

Security	"Developing Guidelines, Standards, and Best Practices for AI Safety and Security" Ref. 9, Topic 1
Supply Chain Risk Management	"Risks arising from AI value chains in which one developer further refines a model developed by another, especially in safety- and rights-affecting systems;" Ref. 9, RFI Topic 2
Test, Evaluation, Verification, and Validation (TEVV)	"Model validation and verification, including AI red-teaming;" Ref. RFI Topic 1; "Guidelines and standards for trustworthiness, verification, and assurance of AI systems;" Ref. 9, RFI Topic 3
Transparency	"Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data;" Ref. 9, RFI Topic 3

Table 2: Candidate Safe, Secure, and Trustworthy Requirement Families Mapped to RFI

Additional support can be provided to the AI Systems community by creating a document that provides procedures for self-attestation or third-party audits of safe, secure, and trustworthy requirements similar to NIST SP 800-171A, Assessing Security Requirements for Controlled Unclassified Information (Ref. 10). NIST SP 800-171A enables organizations to generate evidence to support assertion of requirements satisfaction.

Summary

- Establishing foundational guidance documents for developing and using safe, secure, and trustworthy AI systems strengthens federal government leadership in the AI community.
- Leveraging federal government demonstrated successes with information security provides confidence in a high value outcome.
- Processes used to develop safe, secure, and trustworthy requirements facilitate collaboration amongst a diverse set of stakeholders.
- The voluntary commitments to manage risk posed by AI secured from leading AI companies by the Biden-Harris Administration (Ref. 11) are easily incorporated into the safe, secure, and trustworthy requirements.
- Safe, secure, and trustworthy requirements support development of Generative AI systems aligned with federal government priorities such as the AI Bill of Rights.
- NIST-led effort ensures that the necessary cybersecurity principles for secure AI systems are incorporated into the requirements.

References:

1. NIST Special Publication (SP) 800-171 Revision 2, Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations
2. https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/defense-industrial-base-sector
3. Federal Register / Vol. 88, No. 246 / Tuesday, December 26, 2023 / Proposed Rules, 89058
4. https://oecd.ai/en/ai-principles
5. Governing AI: A Blueprint for the Future, Microsoft
6. NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0)
7. https://www.whitehouse.gov/ostp/ai-bill-of-rights/safe-and-effective-systems-3/
8. https://www.whitehouse.gov/ostp/ai-bill-of-rights/data-privacy-2/
9. Federal Register / Vol. 88, No. 244 / Thursday, December 21, 2023 / Notices, 88368
10. NIST Special Publication (SP) 800-171A, Assessing Security Requirements for Controlled Unclassified Information
11. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI | The White House