Published on *The National Law Review* https://natlawreview.com

NIST Issues AI Risk-Management Guidance

Article By:

Nicholas Martin

Andrew (A.J.) Tibbetts

Go-To Guide:

- Organizational AI development and deployment within runtime systems
- Risk-based AI identification and mitigation
- AI-related consensus standards to provide guidance and cooperation across the world

On July 26, the U.S. government's National Institute of Standards and Technology (NIST) issued four guidance documents – three final versions and one draft open to public comment – related to artificial intelligence development and implementation. Each was issued pursuant to instructions under the Biden administration's AI Executive Order 14110, which directed several U.S. government agencies to promulgate guidance and regulations with respect to safe, secure, and trustworthy AI. The first guidance document, titled "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" (RMF GAI), describes and defines risks associated with generative AI (GAI) and how organizations can govern, manage, and mitigate such risks. The second document, titled "Secure Software Development Practices for Generative AI and Dual-Use Foundation Models" (SSDF), updates prior NIST software development guidance to add recommendations with respect to a framework for implementing secure development practices specifically tailored to generative AI systems. The third document, titled "A Plan for Global Engagement on AI Standards" (AI Plan), provides directives to drive worldwide development and implementation of AI-related consensus standards, cooperation, and information sharing. The fourth document, open to public comment through Sept. 9, 2024, titled "Managing Misuse Risk for Dual-Use Foundation Models" (MMRD), offers comprehensive guidelines for identifying, measuring, and mitigating misuse risks associated with powerful AI models.

Together, the documents attempt to define best practices to reduce risks that arise when developing and deploying AI models. While the NIST guidance is not legally binding, those developing or deploying AI models might take note, as deviation from prevailing practices or recommendations could introduce insurance or liability risks, particularly for those operating in accordance with federal information systems.

The RMF GAI profiles the functions and categories of NIST's 2023 AI Risk Management Framework (AI RMF) applications as they relate to specific GAI settings and usage. It covers the requirements, risk tolerance, and resources of framework users, as well as how AI RMF profiles can help

organizations decide how to manage AI risks in alignment with their goals, considering legal and regulatory requirements, best practices, and risk management priorities. The document offers insights into managing risk across various stages of the AI lifecycle, specifically for GAI technology. The document provides a cross-sectoral profile of how to govern, map, measure, and manage risks related to activities or business processes that implement GAI. Moreover, the document's GAI insights can be applied across different sectors, such as the use of large language models, cloud-based services, or acquisitions. And such insights are prospective in that they address current concerns while looking forward to venues where GAI may be harmful.

The SSDF notes risks that may arise in AI model development and emphasizes secure practices for GAI and dual-use models. The scope covers the entire AI model development lifecycle, including data sourcing, model training, and integration with other software. High-level practices to secure AI elements, such as model weights and training data, are highlighted as a way of mitigating risks from malicious tampering during AI development and training to ensure AI model integrity and confidentiality.

The SSDF contains several detailed recommendations, best understood by reviewing the table set out in the document. Generally speaking, the SSDF leverages NIST's 2023 AI RMF, designed for the flexible yet secure development of AI, which encourages a risk-based approach tailored to each organization's needs. For example, the recommendations stress review of all source code during AI development and training, whether human-written or AI-generated, to detect, evaluate, and address any identified and/or potential vulnerabilities. The document also emphasizes collaboration among AI model producers, AI system producers, and acquirers, advocating for clear agreements on security responsibilities. The SSDF additionally highlights challenges in tracking AI model lineage and versioning, acknowledging the difficulty in securing all aspects of model development. The profile encourages organizations to document security gaps transparently and integrate secure practices where feasible, particularly in machine learning operations (MLOps) and continuous integration/continuous delivery (CI/CD) pipelines. The SSDF thus provides for comprehensive risk management, inclusive of data privacy and bias concerns, across the AI development lifecycle by providing a baseline via AI-specific recommendations.

NIST's AI Plan further leverages the 2023 AI RMF, as the document is guided by principles set forth in the framework to manage AI-associated risks for individuals, organizations, and society. The AI Plan warns that for AI standards to be successful, they must be context-sensitive, performance-based, human-centered (e.g., accounting for people's need and how they interact with AI), and responsive to societal considerations. The document further indicates that AI standards should be developed in an open, transparent, and consensus-driven way to ensure standards are not only effective but also applicable to the dynamic real-world scenarios for which they will be needed. The document, however, notes that while striving for standardization, "more work is needed on most or all sub-topics before standard can be developed." The document advises that some paths to certain standards may take longer than others, but measured approaches to such foundational work can reap a curated system that converges on governance and accountability across users, systems, and the very AI systems being implemented. Indeed, actionable guidance can be achieved for developers, project managers, senior leaders, and other AI actors, securing how such systems are developed and deployed in the world.

In the MMRD, NIST outlines guidelines for managing misuse of dual-use foundation models. Like SSDF, it contains several proposals. Some key recommendations include the following:

- Identifying and maintaining a list of threat profiles covering significant potential misuses;

- Assessing the impact of each identified threat profile if a misuse occurs;
- Estimating model capabilities before development by comparing them to existing models;
- Determining acceptable levels of misuse risk and aligning development plans accordingly;
- Implementing security practices to protect against model theft;
- Using cybersecurity “red teams” (e.g., ethical hackers) to assess whether threat actors could bypass safeguards;
- Monitoring distribution channels for evidence of misuse;
- Maintaining a process to respond to incidents of actual and/or potential model misuse;
- Providing safe harbors for third-party safety research; and
- Publishing regular transparency reports about misuse risks and management practices.

Given the breadth of its recommendations, the MMRD advises that organizations prioritize their AI development, implementation, and safeguards based on a dynamic set of factors, including context of use and resources used, collaboration between departments and/or third parties, and transparency. Accordingly, organizations should tailor the MMRD profile’s recommended framework to fit their specific capabilities, risk profile, and/or current and evolving needs. Comments on the MMRD are due Sept. 9.

The guidance documents outline potential AI security, economic, health, and safety risks to users and the world at large, and provide prospective measures that may help ward off such concerns.

National Law Review, Volume XIV, Number 215