KPMG

Regulatory Alert

Regulatory Insights

May 2024

NIST Draft AI Guidance, Report, and Global Plan

KPMG Insights:

- ***Flurry of Releases:*** *NIST releases part of a continuing series of principle-based frameworks/guidance under the AI EO.*
- ***Span of Principles:*** *AI issuances span GenAI-related risk management and development frameworks, transparency/explainability approaches for synthetic content, and plans for global alignment.*
- ***Quick Turn:*** *Public comments due back in a month, demonstrating swiftness for which agencies are working to establish AI regulatory guidance in advance of elections.*

The U.S. Department of Commerce's National Institute of Standards and Technology (NIST) releases four draft items, including two guidance documents, a report, and a global plan, covering the following:

1. Guidance: Artificial Intelligence Risk Management Framework: Generative artificial intelligence (GenAI) Profile (NIST AI 600-1)
2. Guidance: Secure Software Development Practices for GenAI and Dual-Use Foundation Models (NIST Special Publication (SP) 800-218A)
3. Report: Reducing Risks Posed by Synthetic Content (NIST AI 100-4)
4. Plan: Global Collaboration on AI Standards (NIST AI 100-5)

The releases respond to directives in the Executive Order (EO) on the Safe, Secure and Trustworthy Development of AI (see KPMG's Regulatory Alert, here), and are intended to help improve the safety, security and trustworthiness of AI and GenAI systems.

Highlights from each of the publications are detailed below.

1. Guidance: AI Risk Management Framework: GenAI Profile

In January 2023, NIST published its AI Risk Management Framework 1.0 (AI RMF 1.0) which is intended for voluntary use to assist companies' incorporation of trustworthiness considerations into design, development, use, and evaluation of AI products, services, and systems.

The draft AI RMF: GenAI Profile (NIST AI 600-1) is designed as a "companion resource" for users of the AI RMF 1.0. It is similarly voluntary and serves as both a use-case and cross-sectoral profile of the AI RMF 1.0 related to GenAI risk management. It is intended to assist companies in considering legal and regulatory requirements, as well as industry best practices for managing GenAI-specific risks. In particular, the draft profile defines a group of risks that are unique to, or exacerbated by, the use of GenAI, and provides key actions to help govern, map, measure, and manage them. These risks include:

- Chemical, biological, radiological, or nuclear (CBRN) weapons information.

- Confabulation (e.g., "hallucinations" or "fabrications").
- Dangerous or violent recommendations.
- Data privacy, particularly biometric, health, location, personally identifiable, or other sensitive data.
- Environmental impacts due to resource utilization in training GenAI models.
- Human-AI configurations (arrangement or interaction of humans and AI systems which may result in such things as "algorithmic aversion", automation bias, or misalignment of goals).
- Information integrity.
- Information security.
- Intellectual property.
- Obscene, degrading, and/or abusive content.
- Toxicity, bias, and homogenization.
- Value chain and component integration (non-transparent/ untraceable integration of upstream third-party components (e.g., data acquisition and cleaning, supplier vetting across the AI lifecycle).

2. Guidance: Secure Software Development Practices for GenAI and Dual-Use Foundation Models

The draft Secure Software Development Practices for GenAI and Dual-Use Foundation Models (NIST SP 800-218A) is designed as a "Community Profile" companion resource to supplement NIST's existing Secure Software Development Framework (SSDF) (SP 800-218) and is intended to be useful to producers of AI models, producers of AI systems that use those models, and acquirers of those AI systems.

While the existing SSDF focuses on assisting companies to secure software's lines of code, the draft profile expands on that focus to help address concerns around malicious training data adversely affecting GenAI systems. The draft guidance adds practices, tasks, recommendations, considerations, notes, and other information specific to GenAI and dual-use foundation model development throughout the software development lifecycle, including potential risk factors (e.g., signs of data poisoning, bias, homogeneity, or) and strategies to address them.

3. Report: Reducing Risks Posed by Synthetic Content

NIST's draft report, *Reducing Risks Posed by Synthetic Content* (NIST AI 100-4), provides an overview of technical approaches to promoting digital content transparency based on use case and specific context, including:

- Current methods and approaches for provenance data tracking, authentication, and labeling synthetic content (e.g., digital watermarking, metadata recording, content labels).
- Testing and evaluating techniques for both:
  - Provenance data tracking.
  - Synthetic content detection.
- Preventing and reducing harms from explicit/non-consensual intimate synthetic content (e.g., filtering various data (training, input, image outputs, etc.), hashing content that is confirmed to be harmful, and red-teaming and testing).

NIST notes that this report informs, and is complementary to, a separate report required under the AI EO Section 4.5(a) on monitoring the provenance and detection of synthetic content that will be submitted to the White House.

4. Plan: Global Collaboration on AI Standards

NIST's draft *Plan for Global Engagement on AI Standards* (NIST AI 100-5) calls for a coordinated effort to work with key international allies and partners and standards developing organizations to drive development and implementation of AI-related standards, cooperation and coordination, and information sharing. The plan outlines recommendations in the areas of:

- Standardization: Priority topics (e.g., terminology/taxonomy, risk measurements/mitigations, shared practices for testing, evaluation, verification, and validation (TEVV) of AI systems, mechanisms for enhancing awareness of the origin of digital content (authentic or synthetic), etc.), risk-based management frameworks, and other topics that may require more scientific research and development to understanding about critical components of a potential standard (e.g., energy consumption of AI models, incident response and recovery plans).

— Collaboration: Prioritize engagement with international standards developers, particularly on research and related technical activities; facilitate diverse multi-stakeholder engagement, including private sector leadership/efforts both domestically and more broadly; promote international exchange and alignment on standards and frameworks.

Comment Periods. NIST is soliciting public comments on all four releases (NIST AI 600-1, NIST SP 800-218A, NIST AI 100-4, and NIST AI 100-5), with a deadline to submit by June 2, 2024.

For more information, please contact Matt Miller or Bryan McGowan.

Contact the author:

Amy Matsuo
Principal and National Leader
Regulatory Insights
amatsuo@kpmg.com