KPMG

Regulatory Alert

Regulatory Insights

December 2023

AI Actions: NIST RFI

KPMG Insights:

- ***AI Guidelines Coming:*** *NIST seeks comments to develop guidelines for evaluation and red-teaming; consensus-based standards; "and more".*
- ***Whole of Government:*** *RFI is part of responses to the Executive Order on Safe, Secure, and Trustworthy Development and Use of AI.*
- ***Quick Turn:*** *RFI responses are due February 2; expect a continued fast pace across US agencies.*

The National Institute of Standards and Technology (NIST), an agency within the U.S. Department of Commerce, issues a Request for Information (RFI) to assist with implementation of its responsibilities under the recent Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (see KPMG Regulatory Alert, here). The Executive Order directs NIST to develop guidelines for evaluating AI technologies and for conducting red-teaming tests; facilitate development of consensus-based standards; and provide testing environments for the evaluation of AI systems.

The RFI specifically calls for information related to:

1. AI risk management and evaluation
2. AI red-teaming
3. Reducing the risk of synthetic content
4. Advancing responsible global technical standards for AI development.

NIST does not pose any specific questions but rather provides "non-exhaustive" and extensive lists of topics that may be covered by commenters. Comments are requested no later than February 2, 2024.

1. AI risk management and evaluation

NIST is seeking information to establish guidelines and best practices to promote industry standards in the development and deployment of safe, secure, and trustworthy AI systems. The effort consists of two parts: 1) a companion resource to the AI Risk Management Framework, and 2) guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on which AI could be used to cause harm.

Companion resource. Topics to be considered include:

- Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics
- Current standards or industry norms or practices for implementing core functions for generative AI (govern, map, measure, manage), or related gaps
- The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve
- Current techniques and implementations, including their feasibility, validity, fitness for purpose, and scalability, for: model validation and verification;

human rights impact assessments; content authentication

- Efficacy, validity, and long-term stability of watermarking techniques and content authentication tools for provenance of materials, including in derivative work
- Criteria for defining an error, incident, or negative impact
- Governance policies and technical requirements for tracing and disclosing errors, incidents, or negative impacts.

Guidance and benchmarks for evaluating and auditing. Topics to be considered include:

- Definitions, types, and design of test environments, scenarios, and tools for evaluating the capabilities, limitations, and safety of AI technologies
- Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness
- Generalizability of standards and methods of evaluating AI over time, across sectors, and across use cases
- Applicability of testing paradigms for AI system functionality, effectiveness, safety, and trustworthiness (including security, and transparency), as well as paradigms for comparing AI systems against each other, baseline system performance, and existing practice

2. AI red-teaming

NIST will establish guidelines, including procedures and processes, to enable developers of AI, especially developers of dual-use foundational modes, to conduct AI red-teaming tests. The RFI seeks information on possible topics, including:

- Use cases for AI risk assessment and management
- Capabilities, limitations, risks, and harms that AI red-teaming can help identify
- Current best practices for AI safety, including identifying threat models and associated limitations, or harmful or dangerous capabilities
- Sequence of actions for AI red-teaming exercises and documentation practices
- Limitations of red-teaming and practices to fill identified gaps
- Optimal composition of AI red teams, including backgrounds, skills, expertise

3. Reducing the risk of synthetic content.

NIST is seeking information on topics related to reducing the risk of synthetic content in both closed and open source models, noting that it recognizes "the most promising approaches will require multistakeholder input, including scientists and researchers, civil society, and the private sector." Information may be provided on existing tools and the potential development of future tools, measurement methods, best practices, active standards work, exploratory approaches, challenges and gaps for the topics of interest, including:

- Authenticating content and tracking its provenance
- Techniques for labeling synthetic content, such as using watermarking
- Detecting synthetic content
- Resilience of techniques for labeling synthetic content to content manipulation
- Preventing generative AI from producing harmful synthetic content
- Ability for malign actors to circumvent such techniques
- Different risk profiles and considerations for synthetic content for models with widely available model weights
- Approaches that are applicable across different parts of the AI development and deployment lifecycle, at different levels of the AI system, and in different modes of model deployment
- Testing software
- Auditing and maintaining tools for analyzing synthetic content labeling and authentication.

4. Advancing responsible global technical standards for AI development.

NIST is seeking information regarding topics related to the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing. Working with other federal agencies, NIST seeks to establish a plan for global engagement on promoting and developing AI standards that is guided by principles set out in the NIST AI Risk Management Framework. Topics include:

- AI nomenclature and terminology

- Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness in the collection and use of data
- Examples and typologies of AI systems for which standards would be particularly impactful (e.g., because they are especially likely to be deployed or distributed across jurisdictional lines, or to need special governance practices)
- Best practices for AI model training
- Guidelines and standards for trustworthiness, verification, and assurance of AI systems
- AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems
- Human-computer interface design for AI systems
- Application specific standards (e.g., for computer vision, facial recognition technology)
- Potential mechanisms, venues, and partners for promoting international collaboration, coordination, and information sharing on standards development

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

Contact the author:

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

kpmg.com/socialme