AI Risk: Evaluating and Managing It Using the NIST Framework

AInsights

05 / 18 / 23

If you have any questions regarding the matters discussed in this memorandum, please contact the following attorneys or call your regular Skadden contact.

Stuart D. Levi
Partner / New York
212.735.2750
stuart.levi@skadden.com

William Ridgway
Partner / Chicago
312.407.0449
william.ridgway@skadden.com

Lilia Jimenez
Law Clerk / New York
212.735.3662
lilia.jimenez@skadden.com

One Manhattan West
New York, NY 10001
212.735.3000

155 N. Wacker Drive
Chicago, IL 60606
312.407.0700

The rapid adoption of artificial intelligence (AI) technology into corporate environments has left many organizations understandably struggling with how to identify, measure and manage the unique risks of these nascent systems. Organizations are also trying to determine a pathway to build trustworthy AI systems in order to avoid the significant business and reputational risks that can arise from implementing AI systems that do not function as intended. One approach to address these issues is to adopt, in whole or in part, an AI risk framework released by the National Institute of Standards and Technology (NIST), an agency of the U.S. Department of Commerce that promotes U.S. innovation, typically through establishing standards and frameworks.

NIST designed the Artificial Intelligence Risk Management Framework (AI RMF) to help organizations better identify, manage and mitigate AI risks and create more trustworthy AI systems. Along with the AI RMF, NIST has released a companion "playbook" with further implementation guidelines for organizations, a roadmap of its plans regarding AI developments and a "crosswalk" explaining how the AI RMF matches up to the OECD Recommendation on AI, the proposed EU AI Act, U.S. Executive Order 13960 on promoting the use of trustworthy artificial intelligence in the federal government, and the Biden administration's Blueprint for an AI Bill of Rights. NIST has also launched the Trustworthy and Responsible AI Resource Center to facilitate implementation of, and international alignment with, the AI RMF.

We provide below an overview of the AI RMF and the related materials NIST has issued.

Background

NIST has explained that it released an AI-specific framework in addition to the other standards and frameworks that already exist for information technology systems, privacy and cybersecurity because the risks posed by AI are unique. For example:

- AI systems may be trained on data that can be biased or taken out of context or that can change in significant and unexpected directions, which can then affect the functionality and trustworthiness of an AI system in ways that are difficult to understand.
- AI systems are also often deployed in complex contexts, making it difficult for developers to detect and respond to failures, and requiring more frequent updates and maintenance than other software projects.
- AI systems — unlike most other software systems — are influenced by societal dynamics and human behavior.
- Given the unique nature of AI systems, testing them or to knowing what to test is difficult, resulting in fewer "best practices" and the release of AI systems that may not have undergone the same rigorous testing standards as other software projects do.

The premise of the AI RMF is that AI risk management (namely, minimizing negative impacts, such as threats to civil liberties and rights, while maximizing positive outcomes of using the software) is a key component of responsible development and use of AI systems. Such an approach will help "AI actors" (*i.e.*, primarily those who design, develop, deploy, evaluate and manage risks of AI systems) consider potential negative impacts, and thereby enhance the reliability of and cultivate public trust in AI systems. NIST characterizes the AI RMF as "voluntary, rights-preserving, non-sector-specific, use-case agnostic" guidance that is intended to be readily adaptable throughout the AI life cycle.

AI Risk: Evaluating and Managing It Using the NIST Framework

The first part of the AI RMF outlines the various risks presented by AI, and the second part provides a framework for considering and managing those risks. One of the key focus areas of the AI RMF is those involved in testing, evaluation, verification and validation (TEVV) processes throughout the AI lifecycle. However, NIST emphasizes that also critical to AI risk management are groups not normally involved in technology development, such as advocacy groups that can assist primary AI actors by providing context and understanding of potential and actual impacts of AI usage.

Overall AI Risks

According to NIST, AI risk is unique because of the different sectors it can impact. This includes:

- Harm to people (*e.g.*, harm to an individual's civil liberties, rights, physical or psychological safety or economic opportunity).
- Harm to organizations (*e.g.*, harm to an organization's reputation and business operations).
- Harm to an ecosystem (*e.g.*, harm to the global financial system or supply chain).

The AI RMF seeks to take into account each of these, and encourages stakeholders to do the same.

Unique Challenges in Managing AI Risks

The AI RMF sets forth some unique challenges in AI risk management:

- **Risk measurement.** Organizations developing AI systems may not be transparent about the risk metrics or methodologies they used, and there is a lack of consensus on robust and verifiable measurement methods for assessing risks for different AI use cases. This risk is compounded by the facts that measuring risk at earlier stages of the AI lifecycle can yield different results than measuring risk at later stages; developers may have different risk perspectives than those deploying the models; and risks presented when AI systems are tested in a controlled environment may differ from the risks posed when that same system is deployed in the real world.
- **Risk tolerance.** Risk tolerance refers to the risks an AI actor is willing to bear to achieve its objectives. The level of risk tolerance of an AI actor may vary depending on the circumstances, and may differ for a developer, an organization deploying an AI tool and the individual impacted by that tool.
- **Risk prioritization.** The AI RMF notes that not all AI risks can be eliminated. Instead, organizations need to prioritize which risks they want to eliminate or mitigate. While AI systems that interact directly with humans typically present higher risks, NIST cautions that AI systems intended not to be "human-facing" may still have downstream safety or social implications that merit consideration.
- **Organizational integration and management of risk.** An AI risk management programs needs to be integrated into other organizational risk programs — such as those governing privacy and cybersecurity, and may require training and even cultural changes within an organization in order to dictate how to assess and appreciate the risks of AI.

AI Trustworthiness

The AI RMF also provides a framework for assessing whether an AI system is "trustworthy" — a key aspect of a risk assessment. In most cases, the AI RMF draws on standards from the International Standard of Organization (ISO):

- **Validity and reliability.**
  - Validation means an actor has confirmed through objective evidence that the requirements for a specific intended use or application of AI have been fulfilled.
  - Reliability means overall consistency of the AI system given expected use.
  - These criteria include ensuring the AI system is accurate and robust (*i.e.*, able to maintain its level of performance under a variety of circumstances).
- **Safety.** Safety means an AI system does not endanger human life, health, property or the environment.
- **Resilience and security.**
  - Resilience means the ability of an AI system to return to normal functioning after an unexpected adverse event.
  - Security includes protocols to avoid, protect against, respond to or recover from attacks.
- **Transparency and accountability.** Transparency means information about an AI system and its outputs is available to those interacting with the system; *e.g.*, maintaining the provenance of training data would help create a transparent and accountable AI system.
- **Explainability and interpretability.** Explainable and interpretable AI systems provide information that enable end users to understand the purposes and potential impact of an AI system. AI systems that are explainable can be efficiently debugged, monitored and audited.
- **Privacy.** In the context of AI systems, privacy means, in part, freedom from intrusion or observation. The AI RMF notes that AI systems can promote or reduce privacy.
- **Fairness.** A trustworthy AI system addresses issues such as harmful bias and discrimination, and includes concerns for equality and equity. The AI RMF identifies three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive.

- Systemic bias can be present in AI datasets.
- Computational and statistical biases often arise from the use of nonrepresentative samples.
- Human-cognitive biases can arise from the ways individuals or groups perceive and use AI system information.

NIST notes that AI systems might increase the "speed and scale of biases" and perpetuate and amplify resultant harms.

Managing AI Risks

The AI RMF Core consists of four functions — governing, mapping, measuring and managing — which are broken down into subcategories and provide organizations and individuals with specific recommended actions and outcomes to manage AI risks. NIST notes that these four functions should not be seen as a checklist or an ordered and complete set of oversight actions.

- **Governing:** The governing function relates to how AI is managed within an organization. This includes creating a culture of risk management; outlining processes, documents and organizational schemes that anticipate, identify and manage AI risks; and providing a structure to align with overall organizational principles, policies and strategic priorities. Specific categories within this function include:
  - Creating and effectively implementing transparent policies, processes and practices across the organization related to the mapping, measuring and managing of AI risks.
  - Maintaining policies and procedures to address AI risks and benefits arising from using third-party software and data.
  - Establishing accountability structures so that the appropriate teams and individuals are empowered, responsible and trained for mapping, measuring and managing AI risks.
- **Mapping:** Mapping establishes the context within which to identify and frame the risks of an AI system (such as who the users will be and what their expectations are). This can include missions, goals and risk tolerance. After completing this function, organizations should have sufficient contextual knowledge about the impact of an AI system in order to decide whether to design, develop and deploy that system. Outcomes of the mapping function should form the basis for the measuring and managing functions. Specific categories within this function include:
  - Assessing AI capabilities, targeted usage, goals and expected benefits and costs.
  - Mapping risks and benefits for all components of the AI system, including third-party software and data.
  - Determining the impact of the system on individuals, groups, communities, organizations and society.
- **Measuring:** The measuring function uses information gathered from the mapping process as well as other tools and techniques to analyze and monitor AI risks. Organizations can address this function by implementing software testing and performance assessment methodologies. Measuring risks includes tracking metrics for trustworthy characteristics and impacts of the AI system, and should also provide management with a basis for making decisions when trade-offs of using AI arise. Specific categories within this function include:
  - Identifying and applying appropriate methodologies and metrics.
  - Evaluating AI systems for trustworthiness.
  - Maintaining mechanisms for tracking AI risks.
  - Gathering feedback about the efficacy of measurements being used.
- **Managing:** The managing function consists of allocating risk management resources to address the risks identified through the mapping and measuring functions on a regular basis: Organizations should use the information generated from the first two functions to manage and decrease the risk of AI system failures by identifying risks early and controlling for them. Organizations can implement this function by regularly monitoring and prioritizing AI risks based on assessments from the mapping and measuring functions. Specific categories within this function include:
  - Planning, preparing, implementing and documenting strategies to maximize AI benefits and minimize negative impacts, including input from relevant AI actors in the design of these strategies.
  - Ensuring that risks that arise, including the resulting responses and recovery actions, are documented and monitored regularly.

AI RMF Profiles

NIST suggests establishing use-case profiles as a means to evaluate how risk can be managed at various stages of the AI lifecycle or in a specific sector, technology or end-use application. For example, an organization might create a "hiring profile" where AI is used for hiring, while a comparison of a "current profile" and "target profile" might help an organization conduct a risk gap analysis. In NIST's view, profiles will help organizations manage AI risk in a manner that aligns with their organizational goals, takes into account legal/regulatory requirements and best practices, and reflects an organization's risk management priorities.

Additional NIST Resources

The NIST playbook released with the AI RMF provides additional recommendations and actionable steps for organizations, including further details on the AI RMF Core functions (governing,

measuring, mapping and managing). NIST also plans to release *The Language of Trustworthy AI: An In-Depth Glossary of Terms* to provide organizations and individuals with a shared understanding of AI terms and improve communication among those interested in trustworthy and responsible AI.

In March 2023, NIST established a Trustworthy and Responsible AI Resource Center (AIRC) that hosts the AI RMF, and will feature related resources to facilitate implementation of, and international alignment with, the AI RMF. The resource center is expected to include technical documents and AI toolkits, stakeholder-produced content, case studies and educational materials, and to serve as a repository hub for standards, measurement methods and metrics.

Key Takeaways

With the use of AI expanding in ways for which most companies were not prepared, the AI RMF provides companies with a comprehensive tool to understand and evaluate the risks posed by AI and understand how to build trustworthy systems. The AIRC will also be a resource for companies to reference new documents related to AI regulation. NIST has emphasized that AI technology is rapidly evolving and the institute expects to continuously update its frameworks and resources, including ways to measure improvements in the trustworthiness of AI systems. Finally, NIST has encouraged those who use the AI RMF to periodically evaluate whether the framework has improved their ability to manage AI risks, including through their policies, processes and expected outcomes.