ML STRATEGIES

NIST Invites AI Developers to Submit Models for Risk Assessment Testing — AI: The Washington Report

May 31, 2024 | Article | By **Bruce Sokler**, **Alexander Hecht**, **Christian Tamotsu Fjeld**, Raj Gambhir

VIEWPOINT TOPICS

- Artificial Intelligence

1. The National Institute of Standards and Technology (NIST) has launched the **Assessing Risks and Impacts of AI (ARIA) program**, an initiative designed to allow AI developers to submit their models to NIST for risk evaluation and testing. Through this program, NIST hopes to refine AI testing methodologies and provide AI developers with feedback to make their systems safer and more robust.
2. On May 28, 2024, ARIA announced the commencement of its inaugural program, **ARIA 0.1**. This program invites developers of large language models to present their models to NIST for three layers of evaluation: general model testing, red teaming,**[1]** and large-scale field testing.
3. NIST is encouraging those interested in learning more about or participating in the ARIA 0.1 pilot to join the ARIA mailing list by signing up on the **ARIA website** or **emailing ARIA**.

In April 2024, the National Institute of Standards and Technology's (NIST) Information Technology Laboratory (ITL) launched a new initiative to "assess the societal risks and impacts of artificial intelligence systems." Building off of the NIST **AI Risk Management Framework**, the recently announced **Assessing Risks and Impacts of AI (ARIA) program** invites developers from around the world to submit their AI models and systems to rigorous evaluation by NIST. Through these evaluations, NIST hopes to develop further "guidelines, tools, evaluation methodologies, and measurement methods" related to AI risk assessment.

NIST AI Risk Management Frameworks

ARIA represents just the latest development in NIST's yearslong effort to develop publicly available AI risk management methodologies.

The **National Artificial Intelligence Initiative Act of 2020** directed NIST to publish and periodically update "a voluntary risk management framework for trustworthy artificial intelligence systems." After over a year of deliberation, in January 2023, NIST released the Artificial Intelligence Risk Management Framework (AI RMF), a resource to provide designers of AI systems with "approaches that increase the trustworthiness of AI systems…to help foster the responsible design, development, deployment, and use of AI systems over time."

Despite the approbation that the AI RMF has received since its publication, some worried that rapid advances in generative AI that had occurred since the document's publication had rendered it outdated. To respond to these concerns, in June 2023, NIST announced the launch of a **Public Working Group on Generative AI** to "develop key guidance to help organizations address the special risks associated with generative AI technologies." On April 29, 2024, NIST released a draft companion to the AI RMF entitled "***Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile***."**[2]**

ARIA: Operationalizing the AI Risk Management Frameworks

To further develop and operationalize these risk management strategies, NIST ITL launched ARIA, a program designed to allow developers to submit their AI models to NIST for assessment. Through successive rounds of evaluation, model developers participating in ARIA will receive feedback on the safety and security of their models. Through these evaluations, NIST hopes to develop and refine its AI risk management methodologies.

Models submitted to ARIA will undergo **three levels of evaluation**, each intended to ascertain whether they will be "valid, reliable, safe, secure, private or fair once deployed."

1. **Model testing**: Examining the AI model's "functionality and capabilities." AI developers often subject their models to this sort of baseline evaluation in which the model's outputs are compared to "expected or known outcomes…to determine how well the model can perform a given set of tasks." NIST asserts that this type of baseline testing is necessary but insufficient since "model testing cannot account for what humans expect from or how they interact with AI technology or make sense of AI-generated output."
2. **Red teaming**: Designating a team of engineers to emulate the behavior of an adversary attempting to undermine or exploit an AI system in order to proactively identify vulnerabilities in said system. Red team testers seek to identify how volatile or unexpected outcomes can be elicited out of a system so that safeguards can be put in place to prevent such outcomes. NIST notes that, like model testing, red teaming "cannot provide direct insights about whether…functionality is realized when people interact with AI systems in regular use," and so is not a sufficient testing approach.
3. **Large-scale field testing**: Designating hundreds or even thousands of participants to interact "with AI applications in realistic settings across multiple sessions under test or control conditions." According to NIST, this mode of evaluation addresses some of the shortcomings of red teaming and model testing as it enables the "evaluation of AI's negative and positive impacts in the systems' native context of use" and can "help reveal what happens in people's regular interactions with technology."

NIST Launches Inaugural Evaluation Program

On May 28, 2024, NIST ITL launched ARIA's inaugural program, the **ARIA 0.1 Pilot Evaluation** (ARIA 0.1). This program "will focus on risks and impacts associated with large language models (LLMs)." Each submission to ARIA 0.1 will undergo model, red team, and field testing. According to the draft ARIA 0.1 evaluation guidelines, submissions will "be evaluated on ARIA scenarios using a suite of metrics focused on technical and societal robustness; these new metrics will be developed in collaborative engagement with the ARIA research community."

According to the ARIA 0.1 draft evaluation guidelines, submissions must conform to seven design constraints.

1. "The application MUST be a textual dialogue system between a user and the system with a prompt length of at least 512 characters to enable user flexibility."
2. "The application MUST implement a user session paradigm where the system may self-adapt within a user session but MUST be resettable to the same session-initial state that does not change for the duration of ARIA testing."
3. "The application MAY model the user and dialogue within a user session only."
4. "The application MUST accept parameterization through user dialogue."
5. "The underlying technology may be any combination of automated computing technologies (e.g., LLMs of any design or implementation including agents and assistants)."
6. "Responses generated by the application MUST be generated by software and not involve human input from the submitter side of the interaction."
7. "The application must implement the ARIA System Interaction API so that NIST can capture logs for further analysis."

At time of writing, NIST has not published the ARIA 0.1 schedule or submission documentation. Those interested in participating in ARIA 0.1 or learning more about the ARIA program can sign-up for the ARIA email distribution list by signing up on the **ARIA website**, or **emailing ARIA**.

Conclusion

Since the release of the AI RMF in January 2023, NIST's AI risk management methodologies have occupied a central position in both public and private sector deliberations regarding the mitigation of AI's potential and actual harms. The centrality of NIST's AI risk management efforts became even clearer in November 2023 when a bipartisan and bicameral group of legislators **introduced legislation** that would codify elements of the AI RMF.

Given the interest shown by a wide swath of regulators in the AI RMF, interested stakeholders should closely track AI related developments coming out of NIST, including the ARIA program. We will continue to monitor, analyze, and issue reports on these developments. Please feel free to contact us if you have questions as to current practices or how to proceed.

Endnotes

---

[1] As explained later in this newsletter, red teaming is a risk mitigation strategy in which a team of engineers emulates the behavior of an adversary attempting to undermine or exploit an AI system in order to proactively identify vulnerabilities in said system.

[2] Stakeholders interested in utilizing the GAI RMF are encouraged to review the draft and **submit comments**.

Authors

Bruce Sokler

Bruce D. Sokler is a Mintz antitrust attorney. His antitrust experience includes litigation, class actions, government merger reviews and investigations, and cartel-related issues. Bruce focuses on the health care, communications, and retail industries, from start-ups to Fortune 100 companies.

Alexander Hecht, Executive Vice President & Director of Operations

Alexander Hecht is Executive Vice President & Director of Operations of ML Strategies, Washington, DC. He's an attorney with over a decade of senior-level experience in Congress and trade associations. Alex helps clients with regulatory and legislative issues, including health care and technology.

Christian Tamotsu Fjeld, Senior Vice President

Christian Tamotsu Fjeld is a Senior Vice President of ML Strategies in the firm's Washington, DC office. He assists a variety of clients in their interactions with the federal government.

Raj Gambhir

Raj Gambhir is a Project Analyst in Washington, DC.