AI Evaluations: Assessing Risks and Impacts of AI

For Release May 9, 2024

Overview

ARIA (Assessing Risks and Impacts of AI) is a NIST evaluation program to advance measurement science for safe and trustworthy AI. Launched in spring 2024, ARIA aims to:

- address gaps in AI evaluation that make it difficult to generalize AI functionality to the real world
- improve understanding of AI's impacts to individuals and society, and
- provide participating organizations with crucial information about whether AI systems will be valid, reliable, safe, secure, private or fair once deployed.

NIST will engage the public through evaluations and related activities in a variety of domains under the ARIA umbrella. ARIA evaluations will include model testing, red-teaming, and field testing. Tasks and related activities will be customized for each evaluation.

Models and systems made available to NIST will be evaluated on ARIA tasks using a suite of metrics focused on technical and societal robustness; these new metrics will be developed in collaborative engagement with the ARIA participant community.

Expected program outcomes include scalable guidelines, tools, methodologies, and metrics for organizations to use for evaluating the safety of their AI systems in their specific use cases, and as part of their governance and decision making processes to design, develop, release or use AI technology.

ARIA 0.1 Pilot Evaluation

The initial evaluation (ARIA 0.1) will be conducted as a pilot effort to fully exercise the NIST ARIA test environment. ARIA 0.1 will focus on risks and impacts associated with large language models (LLMs). Future iterations of ARIA may consider other types of generative AI technologies such as text-to-image models, or other forms of AI such as recommender systems or decision support tools.

View the ARIA 0.1 Pilot Evaluation Plan at https://ai-challenges.nist.gov/aria/docs/evaluation_plan.pdf

Those interested in learning more about ARIA can join the ARIA email distribution list by signing-up at **https://ai-challenges.nist.gov/aria** or emailing **aria_inquiries@nist.gov.**

ARIA Evaluation Levels

ARIA will incorporate societal impacts alongside functional testing of the system. Estimating a given technology's impact in society requires a better understanding of what individuals and the broader society can and will do with – or how they adapt and react to – an AI model or system functionality. To this end, NIST will establish three measurement and evaluation levels for a more comprehensive approach. These are introduced below.

1. **Model testing** to examine the AI model or system components' functionality and capabilities.

   Model testing is the most common practice for evaluating the models and datasets underlying AI technology. Typical model testing involves comparing system outputs to expected or known outcomes (sometimes referred to as ground-truth) to determine how well the model can perform a given set of tasks. Demonstrating whether the model functions on these tasks as designed can shed light on how helpful or harmful the technology may be once deployed. This type of testing is easier to scale than red-teaming or field-testing but has limitations. Particularly when performed in laboratory settings, model testing cannot account for what humans expect from or how they interact with AI technology or make sense of AI-generated output. For estimating societal impact, static benchmark datasets serve only as loose proxies for dynamic human interactive behavior. These limitations make it difficult to understand or anticipate impacts once a model or system is deployed.

2. **Red-teaming** to identify potential adverse outcomes of the AI model or system and how they could occur, and to stress test model safeguards.

   For AI, red-teaming is a structured testing effort to find flaws and vulnerabilities in an AI system such as false, toxic, or discriminatory outputs in an AI system. Red teaming can be performed before or after AI models or systems are made available to the broader public. Complementary groups with diverse expertise can elicit different types of harms in AI red-teaming activities. Experts can emulate malicious behavior and surface narrow or targeted harms. Members of the general public can help to gather data en-masse for identifying systemic or diffuse harms. Red-teaming results can lead to remedies for *harmful* model functionality. However – similar to model testing – red-teaming cannot provide direct insights about whether such functionality is realized when people interact with AI systems in regular use.

3. **Large-scale field testing** to help reveal how the public consumes and makes sense of AI-generated information in their regular interactions with technology, including subsequent actions and effects.

   ARIA field testing may entail several thousands of human participants interacting with AI applications in realistic settings across multiple sessions under test or control conditions. This approach will enable evaluation of AI's negative and positive impacts in the systems' native context of use from a human perspective and enhance understanding of AI capabilities and impacts in post-deployment contexts. ARIA's field testing is designed to help reveal what happens in people's

regular interactions with technology. When conducted alongside model testing and red-teaming, results from the large number of human interactions in field testing can reveal:

- types of content and model functionality individuals were actually exposed to when interacting with the system;
- whether, how often, and for whom the interaction contributed to a positive or negative impact;

ARIA Metrics

Starting with the ARIA 0.1 pilot, evaluation output from all three levels will be annotated by professional assessors. Submitted models and systems will be evaluated using a suite of metrics focused on technical and societal robustness. Metrics will be developed in collaborative engagement with the ARIA research and participant community.

NIST will also run a mini challenge within ARIA for additional development and refinement of societal impact metrics. NIST will provide output data from all three evaluation levels after the completion of the ARIA 0.1 pilot for the challenge, so the broader measurement community can:

- pursue valid and generalizable societal impact metrics for the field of AI safety and trustworthiness;
- inform other AI safety evaluation efforts; and
- establish a diverse measurement community that brings new perspectives to the development of innovative metrics in the field of safe and trustworthy AI.