Frequently Asked Questions about NIST’s ARIA Program (Assessing Risks and Impacts of AI)

For Release May 9, 2024

1. **How does ARIA fit in with NIST's other evaluation programs?**

   As the Nation’s oldest measurement laboratory, NIST routinely employs evaluation-driven research to advance measurement science, inform and accelerate the development of emerging technologies, and drive innovation. ARIA (Assessing Risks and Impacts of AI) is a research effort to assist AI evaluators in improving their assessment methods. ARIA evaluations will fill in measurement gaps related to how technology integrates with society and creates impacts.

2. **How does ARIA connect to the US AI Safety Institute and consortium?**

   ARIA is a new internal NIST effort to develop a testing environment for advancing measurement science in Trustworthy AI. The ARIA effort will inform the work of the U.S. AI Safety Institute and over time, the U.S. AI Safety Institute Consortium may assist in enhancing and producing ARIA-style evaluations at scale, useful for all industries.

3. **Does ARIA fulfill one of NIST's assignments in the October 2023 AI Executive Order?**

   Yes. ARIA is one of several NIST evaluation initiatives that partially addresses NIST’s assignment under Section 4.1 (a)(i)(C) of the President’s Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (14110) to launch an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities.

4. **Does NIST select which systems to test?**

   No. For decades, NIST’s Information Technology Laboratory (ITL) has been conducting evaluation-driven research of algorithms and other system components in technology fields such as biometrics, multimedia, and information retrieval. As a neutral party, NIST does not select which systems to test, conduct product testing, or test any technology that has not been submitted by the owning entity. ITL evaluations remain open to any researcher, team, or interested party who finds it of interest, are able to submit their technology applications for measurement, and can comply with the evaluation rules.

   NIST-run evaluations are designed to be widely accessible and utilize a set of common tasks, data, metrics, and measurement methods to reduce the total overhead necessary to conduct research, assess current state of the art, and identify the most promising research directions. For more information about NIST AI technology evaluations, see https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/nist-ai-measurement-and-evaluation-projects

5. **Is NIST evaluating models or approaches in ARIA?**

   Both. The first ARIA activities will focus on risks and impacts associated with large language models (LLM), including the use of AI agents. The risks and impacts of LLMs will be evaluated across three levels – model testing, red-teaming, and field testing.

   As an evaluation of safe and trustworthy AI, submitting organizations will be required to provide documentation about their models, approaches, mitigations, and guardrails, along with information about their governance processes. In future evaluations, documentation requirements may be expanded and constitute part of the final score.

6. **Will all ARIA evaluations be limited to generative AI?**
   While the first set of ARIA activities will focus on risks related to the generative AI technology of LLMs, the ARIA evaluation environment is flexible and future iterations will broaden beyond generative AI. ARIA participant community and other researchers can provide input on future evaluation topics, domains, and technologies. For example, subsequent ARIA evaluations may consider other generative AI technologies such as text-to-image models, or other forms of AI such as recommender systems or decision support tools.

7. **What metrics will NIST use in the ARIA evaluation?**

   ARIA will originate a suite of qualitative, quantitative, and mixed methods to measure risks, impacts, trustworthy characteristics, and technical and societal robustness of models within the specified context of use. NIST will develop these metrics in close collaboration with ARIA participants. Selected ARIA evaluation output data will be made available as a rich corpus for research purposes, including the development of novel metrics for use in ARIA.

8. **Why would vendors participate in ARIA?**

   NIST evaluations provide all participants with the opportunity to obtain vital information about their submitted technology components, make adjustments based on what they learned, and resubmit for further testing. While many organizations evaluate their technology internally, involvement in NIST evaluations allows all participants to determine what is working with their models, often in comparison to other organizations on the same tests, with the same data, and under the same conditions. While the ARIA evaluations are open to all who wish to participate, the evaluation cycle typically concludes with a participant-only workshop to discuss information about new and promising approaches that may assist submitters' understanding about how they might improve their models. Teams participating in ARIA can expect to glean information during testing and the workshop(s) that will help deliver safe and trustworthy AI.

9. **Will results be made public? Are results anonymous?**

NIST evaluation results are made publicly available. The level of information to be made public is predetermined for each evaluation. ARIA participants may decide to anonymize their submissions so that each team only knows how they performed in comparison to others. Even when results are not tied to a particular participating organization, the public will have access to the specific results of all technologies which have been evaluated. That information is valuable in gaining an understanding of how these technologies perform in a real-world context.