SUMMARY - NIST Assessing Risks and Impacts of AI (ARIA) program

The **NIST Assessing Risks and Impacts of AI (ARIA) program** is an evaluation-driven research initiative to develop and refine methodologies, tools, and metrics for measuring AI safety and trustworthiness in real-world contexts. It expands upon the "Measure" function of the broader NIST AI Risk Management Framework (AI RMF).

Core Focus and Objectives

The primary goal of ARIA is to move beyond laboratory settings and understand what happens when people interact with AI in realistic scenarios. It aims to quantify how well a system maintains safe functionality within societal contexts once deployed.

Key objectives include:

- **Filling Evaluation Gaps:** Addressing shortcomings in current methods that fail to capture the societal consequences and real-world impacts of AI systems.
- **Developing New Metrics:** Creating a new suite of quantitative and qualitative metrics focused on "technical and societal robustness".
- **Supporting the U.S. AI Safety Institute:** Providing empirical data and evaluation methods to help build the foundation for safe, secure, and trustworthy AI.

Methodologies and Evaluation Levels

ARIA uses a multi-tier evaluation process that includes three distinct levels of testing to provide a holistic view of AI's effects:

- **Model Testing:** Evaluates the basic technical capabilities and functions of the AI model, often through automated or scripted prompts to confirm claimed capabilities.
- **Red-Teaming:** Involves stress-testing the AI system with adversarial inputs to identify vulnerabilities, biases, and potential harmful outputs.
- **Field Testing:** The most extensive level, which studies human-AI interactions under practical, real-world conditions to observe actual positive and negative impacts on users and society.

Tools, Metrics, and Measurements

ARIA is actively developing the tools and metrics necessary for these evaluations in collaboration with the research community.

- **Contextual Robustness Index (CoRIx):** A key measurement instrument and suite of metrics currently in development for scoring AI systems' ability to maintain robust and trustworthy functionality across deployment conditions.
- **Annotation Schema:** Trained human assessors use specific schema to annotate dialogue interactions captured during testing, characterizing dialogue dynamics, content quality, interaction style, and utility to identify if risks materialize in context.
- **Proxy Scenarios:** Specific, controlled scenarios (e.g., "TV Spoilers" as a proxy for privileged information, "Meal Planner" for health/safety guardrails) are used to elicit and measure specific types of risks in a repeatable task environment.

Ultimately, the results and data from the ARIA program will be made publicly available to foster broader research and help organizations make informed decisions about AI deployment and governance.