Can Your AI Model Handle This? Real World Evaluation

NIST's Assessing Risks and Impacts of AI (ARIA)

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

ARIA provides an evaluation environment to improve AI technology.

Image generated with ChatGPT 4o

Expected long-term outcomes:

- guidelines
- tools
- evaluation methods
- metrics

ARIA is <u>not</u> designed to test for operational, oversight, reporting or certification purposes.

ARIA will start with a pilot evaluation focused on the risks and impacts of large language models (LLMs).

Like other NIST evaluations, ARIA:

- Focuses on long-term scientific exploration.
- Is open to all, teams opt-in to participate.
- Releases evaluation output for research purposes.

It’s difficult to estimate real world impacts from performance-based approaches -- and to know the extent of AI’s opportunities and threats.

Current:
Is the **model output** correct and in line with organizationally established thresholds?

Future:
Does the model withstand deployment across real world contexts?

Image generated with ChatGPT 4o

Context is necessary to know whether a risk will lead to impacts - and what type, for whom, etc.

Risk

The composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event.

The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats.

NIST AI Risk Management Framework

ARIA expands the scope of evaluations to include **people and how they use AI** in real world conditions.

NIST

Model Testing	Red Teaming	Field Testing
Does the application demonstrate required capabilities and associated guardrails ?	Can the application be induced to produce violative outcomes? Under what conditions do violative outcomes occur?	Are people exposed to positively or negatively impactful information during regular use? Do they perceive that exposure? Do they act on that exposure?
how do AI capabilities…	connect to risks…	…and positive and negative impacts

Evaluation environment output can shed light on:

- functionality across risks and contexts
- effectiveness of guardrails and mitigations
- applicability of tests for given risks

AI is more than its data, models, architectures and algorithms.

Most AI risks relate to people, including how they interact with the AI system in context.

Image Generated by ChatGPT 4o

Image: Shutterstock AI Image Generator

ARIA will assist organizations by:

- providing deeper insights about **the conditions under which positive and negative impacts may arise** in context.
- informing decisions about **whether and how to deploy AI**.

ARIA makes use of proxies to “unlock” the structure of the risk and apply it across many use cases.

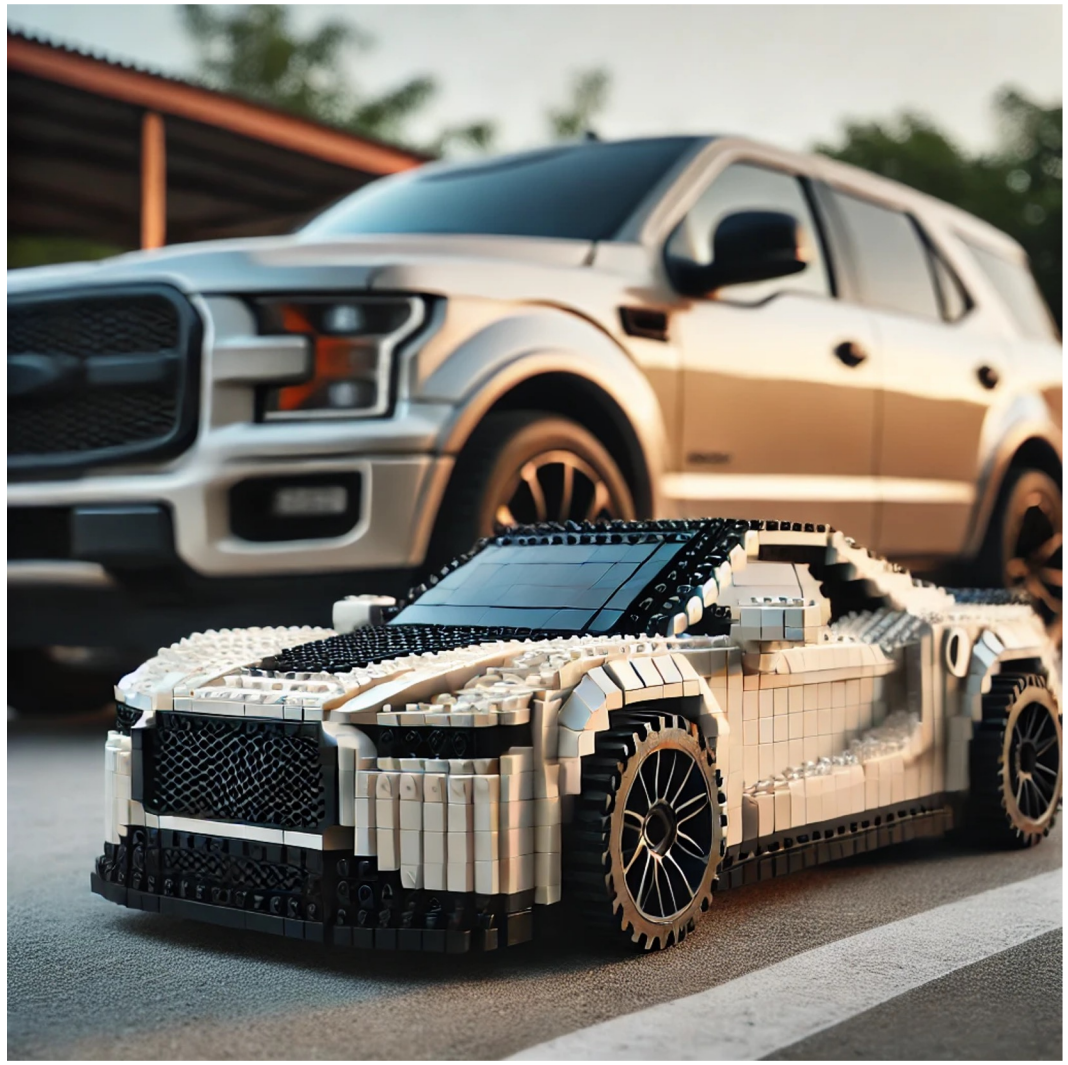

Image Generated by ChatGPT 4o

ARIA translates risks into “proxy” scenarios to enable:

- Focus on the HOW - the *structure of the impacts and associated conditions* - rather than the WHAT.
- *Reuse and adaptation* across contexts and for future evaluations that endure over time as risks and interest in them continues to change.

ARIA's three proxy scenarios enable investigation of how risks arise.

All images: Shutterstock AI Image Generator.

TV Spoilers

Use Case: Information seeking.

Risk: Access to privileged or nefarious information.

LLM Task: Shield protected information about tv series.

Meal Planner

Use Case: Personalization.

Risk: Harmful bias.

LLM Task: Generate meal plans tailored to different audiences.

Pathfinder

Use Case: Information synthesis.

Risk: Confabulation.

LLM Task: Synthesize factual geographical information into travel-related content.

ARIA will create a new space for AI metrology.

Evaluation output will be converted to scalable processes.

- **Guidelines**
- **Metrics**
- **Methods**
- **Tools**
- **Practices**

NIST Assessing Risks and Impacts of AI (ARIA) Team

Team Leads

Reva Schwartz

ARIA Program Lead

Linguistics/Phonetics, Human-Language Technology, AI Risk Management, Experimental Methods

Jonathan Fiscus

Evaluation Lead
Model Testing Lead

Computer Science, Natural Language Processing, Computer Vision, AI Measurement and Evaluation

Rumman Chowdhury

Red Teaming Lead

Data Science, Social Science, Quantitative Methods

Kristen Greene

Field Testing Lead

Cognitive Science, Cognitive Psychology, Human-Computer Interaction, Research Ethics

Gabriella Waters*

Cross-team Harmonization & Field Testing Team Member

Human-Centered Computing, Psychology, Neuroscience, Biology, Genetics

Individual Contributors

Experts drawn from NIST AI Innovation Lab

- Afzal Godil: Model Testing
- Craig Greenberg: Model Testing
- Theodore Jensen: Field Testing
- Patrick Hall*: Red Teaming
- Razvan Amironesei: Labeling Design
- Shomik Jain: Infrastructure

*NIST Associates

https://ai-challenges.nist.gov/aria

aria_inquiries@nist.gov