



CDAO

Chief Digital & Artificial
Intelligence Office



Operationalizing AI Assurance

Dr. Matthew K. Johnson
Chief of Responsible AI
US Department of Defense

Distribution Statement A

Approved for Public Release; Distribution Unlimited



Overview of the Policy Landscape for Trustworthy AI

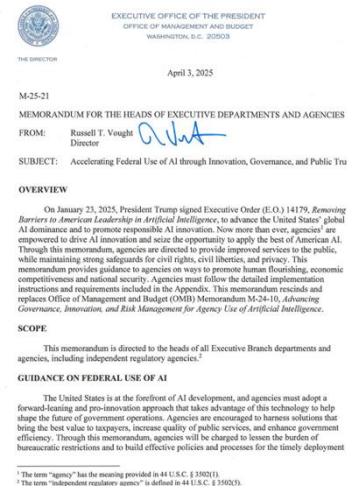


Promote Secure-By-Design AI Technologies and Applications

AI systems are susceptible to some classes of adversarial inputs (e.g., data poisoning and privacy attacks), which puts their performance at risk. The U.S. government has a responsibility to ensure the AI systems it relies on—particularly for national security applications—are protected against spurious or malicious inputs. While much work has been done to advance the field of AI Assurance, promoting resilient and secure AI development and deployment should be a core activity of the U.S. government.

Recommended Policy Actions

- Led by DOD in collaboration with NIST at DOC and ODNI, continue to refine DOD's Responsible AI and Generative AI Frameworks, Roadmaps, and Toolkits.



Sec. 3. Unbiased AI Principles. It is the policy of the United States to promote the innovation and use of trustworthy AI. To advance that policy, agency heads shall, consistent with applicable law and in consideration of guidance issued pursuant to section 4 of this order, procure only those LLMs developed in accordance with the following two principles (Unbiased AI Principles):

Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence



Three Conditions & Three Approaches for AI Assurance

THREE CONDITIONS

- Rapid pace of technological advancement
- AI Assurance Workforce Gaps
- Nascency of Alignment Tradecraft & Testing

THREE APPROACHES

- Crowdsource
- Automate
- Leverage Interdisciplinarity



Operationalizing AI Assurance with the RAI Toolkit

Home Executive Summary Background & Guide RAI Toolkit Appendix

Contact Us Sign in

Overview of RAI Activities Throughout the Product Life Cycle

Intake Ideation Assessment Development/Acquisition TEVV Integration & Deployment Use

SHIELD Navigation

Export Import

Clear Responses

View PDF

1. Intake

1.1 SET: Consider Previously I

1.2 SET: Determine Relevant

1.3 SET: Identify and Engage

1.4 SET: Concretize the Use

1.5 SET: Decide to Proceed to

2. Ideation

2.1 HONE: Define Requireme

2.2 HONE: Identify Risks & O

2.3 HONE: Write Statements

2.4 Design to Reduce Ethical

2.5 Accountability, Responsit

3. Assessment

3.1 HONE: Assess Requireme

3.2 HONE: Exploratory Data

3.3 Conduct AI Suitability As

3.4 Update Documentation

4. Development/Acquisition

4.1 Improve & Innovate: Instr

4.2 Update Documentation

5. TEVV

5.1 EVALUATE: Test System f

7. Have tools for explainability, uncertainty quantification, or competence estimation been used to increase assurance and reduce human error? How are you tracking that these metrics are understood correctly?

EQUI(NE2) XAI Toolkit - Saliency Python Outlier Detection (PyOD)

EQUI(NE2)

Library for uncertainty quantification.

Tool Link

Coding Level High

Tool Class Confidence Metrics

Principles Apply EQUI(NE2) to evaluate the AI capability's uncertainty on in-distribution inputs to help measure and demonstrate reliability (effectiveness of AI capabilities) and governability (fulfill intended functions).

RAI Activities Instrument AI to promote Assurance, Test Components for Robustness and Resilience, Operational Testing, Perform Continuous Monitoring of the System and its Use, Context, & Ecosystem

8. Have you established a cadence and procedure through which new data will be collected, models will be retrained, and the system will be updated?

Response...

9. Describe the periodicity and procedure through which new data will be collected, models will be retrained, and the system will be updated?

Response...

10. Revisit 2.4. How are you designing the system to reduce the ethical/risk burden on leaders, developers, deployment or existence of the system?

Response...

11. Have you established a cadence and procedure through which new data will be collected, models will be retrained, and the system will be updated?

Response...

4.2 Update Documentation

Filters & RASCI

1. Update SOC's and data/model cards, as necessary. Have team consult and update DAGR to support continuous risk identification - as new risks (or opportunities) are identified.

Response...

Export/Import function to save and share progress

Export as PDF

Navigation by Type of RAI Activity

SHIELD Assessment identifies risk and opportunities

Links to tools to address identified risks and opportunities

Navigation by AI Product Lifecycle Stage

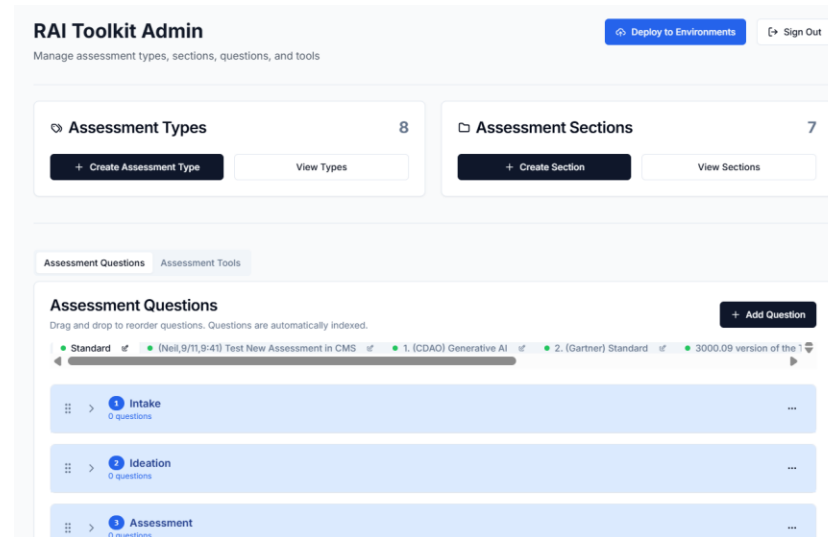
Rapid Deployment Toolkit displays most essential assessment questions

Filters assessment questions by Persona/Project Role, Discipline, etc.



Additional RAI Toolkit Features

- Mapping to other requirements, frameworks, & standards (MIL-STD-882E, NIST AI RMF, IEEE 7000, etc.)
- Additional use-case focused versions of the Toolkit
- Automation Enhancements
- Dashboard Enhancements
- Content Management System



DoD Responsible AI (RAI)

Build Trust, Drive Adoption, Field First



Operationalizing AI Assurance (Build Trust)

- RAI Toolkit + Assurance Portal
- AI Assurance Tools & Best Practices
- Partner Validation & Technical Support
- Frontier AI Red Teaming & Assurance Capability Development

Outward Focus:

- Risk-informed AI investment decision-making, delivering reliable, mission-ready AI for warfighters

Drive AI Assurance Policy & Strategy (Drive Adoption)

- Identify & Mitigate Policy Barriers
- AI Assurance Policies
- AI Acquisition Resources

Inward Focus:

- Strategy and Policy
- DoD Leadership, Authorizing Officials, & Risk Owners

Interoperability with Partners (Field First)

- Chair of White House's CAIO AI Assurance Working Group
- International Partnerships
- Industry Partnerships

Lateral Focus:

- Interagency
- Industry
- Allies & Partners

RAI Contacts

Chief of RAI

Dr. Matthew Johnson
matthew.k.johnson94.civ@mail.mil

Contact the RAI Team:

osd.pentagon.cdao.list.rai@mail.mil

Your Questions Answered

- **Why is RAI important for the DoD?**
 - RAI ensures that AI-enabled capabilities (1) Do what they're supposed to do, (2) Don't do what they're not supposed to do, and (3) Warfighters trust them. Baking in RAI ensures the system will be reliable and avoid downstream issues. Ensuring warfighter trust catalyzes adoption.
- **What do the RAI Toolkit & AI Assurance Portal do?**
 - The RAI Toolkit provides technical tools and processes for ensuring the performance of RAI capabilities and documenting an assurance case. The Portal enables sharing of tools and artifacts, to enable reusability.

