CDAO

Chief Digital & Artificial
Intelligence Office

The DoD’s Approach to Responsible AI & The Responsible AI Toolkit

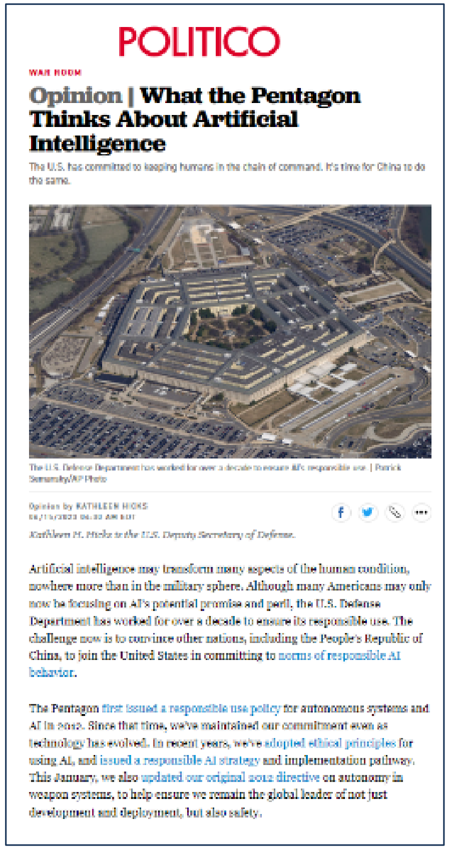
Drew Brooks
Lead Scientist for Responsible AI
Tools U.S. Department of Defense

Distribution Statement A.
Approved for public release: distribution is unlimited.

RESPONSIBLE AI | ToolKit

Winning Because of Our Values

> *“America and China are competing to shape the future of the 21st century, technologically and otherwise. That competition is one which we intend to win-not in spite of our values, but because of them.”*
>
> –Deputy Secretary of Defense Kathleen Hicks



POLITICO

WAR ROOM

Opinion | What the Pentagon Thinks About Artificial Intelligence

The U.S. has committed to keeping humans in the chain of command. It's time for China to do the same.

The U.S. Defense Department has worked for over a decade to ensure AI's responsible use. | Patrick Semansky/AP Photo

Opinion by KATHLEEN HICKS
06/15/2023 06:00 AM EDT

Kathleen H. Hicks is the U.S. Deputy Secretary of Defense.

Artificial intelligence may transform many aspects of the human condition, nowhere more than in the military sphere. Although many Americans may only now be focusing on AI's potential promise and peril, the U.S. Defense Department has worked for over a decade to ensure its responsible use. The challenge now is to convince other nations, including the People's Republic of China, to join the United States in committing to norms of responsible AI behavior.

The Pentagon first issued a responsible use policy for autonomous systems and AI in 2012. Since that time, we've maintained our commitment even as technology has evolved. In recent years, we've adopted ethical principles for using AI, and issued a responsible AI strategy and implementation pathway. This January, we also updated our original 2012 directive on autonomy in weapon systems, to help ensure we remain the global leader of not just development and deployment, but also safety.

What does it mean to "win because of our values?"

CDAO

What Is Responsible Artificial Intelligence (RAI)?

- RAI translates high-level values and Artificial Intelligence (AI) ethical principles into concrete actions, processes, metrics, and benchmarks to fit the use case at hand–and navigates any tradeoffs.

- RAI removes barriers to innovation and adoption through risk identification and reduction.

- RAI contributes to mission success through decision advantage and assurance.

CDAO

Value Proposition of RAI: Assurance

RAI increases assurance, thereby sustaining our tactical edge:

- **Assurance for the Warfighter, Operational Commanders, and DoD Personnel to Reduce Cognitive Load:**
  - Provides assurance that technology has been developed to reduce risks of failure, unintended consequences, and dangerous or difficult ethical situations and choices for operational users.
  - Reduces cognitive load, allowing greater focus on contributors to mission success.

- **Assurance for the Department to Aid Adoption/Innovation:**
  - Provides assurance process to remove barriers to adoption and support effective innovation.

- **Assurance for Industry to Maintain Competitive Advantage:**
  - Ensures industry’s trust that the U.S. Department of Defense (DoD) will responsibly steward its technologies.

- **Assurance for American Public:**
  - Ensures public’s trust that AI-enabled capabilities employed by the DoD are aligned with our values.

- **Assurance for Allies to Increase Interoperability:**
  - Provides systems, tools, and processes grounded in shared values.
  - Is crucial, given the increasing need for interoperability.

Background on the RAI Division

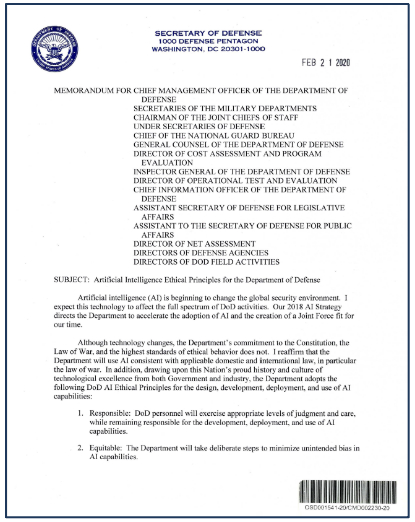
RAI at DoD

- The DoD defines RAI as a dynamic approach to the design, development, deployment, and use of AI systems that implements the DoD AI ethical principles to advance the trustworthiness of such systems.
- RAI at the DoD emphasizes technical maturity, organizational change, modernized governance structures, and an understanding of sociotechnical risk.

RAI Division's Role

- Is the primary technical advisor to the DoD on RAI.
- Oversees execution of the RAI Strategy and Implementation Pathway.
- Coordinates development and implementation of RAI tools, guidance, and other resources.
- Convenes DoD components to develop and recommend RAI best practices governing the creation, development, and use of AI within the DoD.

DoD AI Ethical Principles



SECRETARY OF DEFENSE
1000 DEFENSE PENTAGON
WASHINGTON, DC 20301-1000

FEB 2 1 2020

MEMORANDUM FOR CHIEF MANAGEMENT OFFICER OF THE DEPARTMENT OF DEFENSE
SECRETARIES OF THE MILITARY DEPARTMENTS
CHAIRMAN OF THE JOINT CHIEFS OF STAFF
UNDER SECRETARIES OF DEFENSE
CHIEF OF THE NATIONAL GUARD BUREAU
GENERAL COUNSEL OF THE DEPARTMENT OF DEFENSE
DIRECTOR OF COST ASSESSMENT AND PROGRAM EVALUATION
INSPECTOR GENERAL OF THE DEPARTMENT OF DEFENSE
DIRECTOR OF OPERATIONAL TEST AND EVALUATION
CHIEF INFORMATION OFFICER OF THE DEPARTMENT OF DEFENSE
ASSISTANT SECRETARY OF DEFENSE FOR LEGISLATIVE AFFAIRS
ASSISTANT TO THE SECRETARY OF DEFENSE FOR PUBLIC AFFAIRS
DIRECTOR OF NET ASSESSMENT
DIRECTORS OF DEFENSE AGENCIES
DIRECTORS OF DOD FIELD ACTIVITIES

SUBJECT: Artificial Intelligence Ethical Principles for the Department of Defense

Artificial intelligence (AI) is beginning to change the global security environment. I expect this technology to affect the full spectrum of DoD activities. Our 2018 AI Strategy directs the Department to accelerate the adoption of AI and the creation of a Joint Force fit for our time.

Although technology changes, the Department's commitment to the Constitution, the Law of War, and the highest standards of ethical behavior does not. I reaffirm that the Department will use AI consistent with applicable domestic and international law, in particular the law of war. In addition, drawing upon this Nation's proud history and culture of technological excellence from both Government and industry, the Department adopts the following DoD AI Ethical Principles for the design, development, deployment, and use of AI capabilities:

1. Responsible: DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

2. Equitable: The Department will take deliberate steps to minimize unintended bias in AI capabilities.

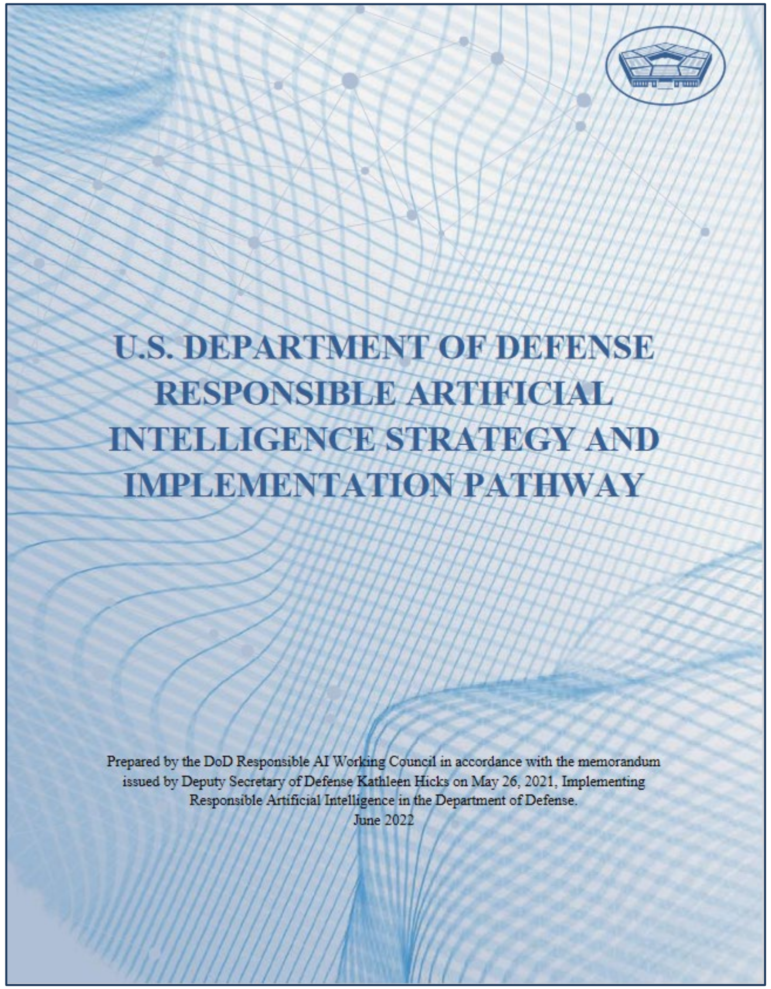
OSD001541-20/CMD002230-20

February 2020: AI Ethical Principles Memorandum
The DoD formally adopts five AI ethical principles and designates the Joint Artificial Intelligence Center (now CDAO) as DoD's lead for coordination and implementation of the principles.

Principle	Description
Responsible	DoD personnel will exercise **appropriate levels of judgment and care,** while remaining responsible for the development, deployment, and use of AI capabilities.
Equitable	The department will take deliberate steps to **minimize unintended bias** in AI capabilities.
Traceable	The department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with **transparent and auditable methodologies, data sources, and design procedure and documentation.**
Reliable	The department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to **testing and assurance** within those defined uses across their entire life cycles.
Governable	The department will design and engineer AI capabilities to fulfill its intended functions while possessing the ability to **detect and avoid unintended consequences** and the ability to **disengage or deactivate deployed systems** that demonstrate unintended behavior.

CDAO

RAI Strategy and Implementation Pathway



U.S. DEPARTMENT OF DEFENSE
RESPONSIBLE ARTIFICIAL
INTELLIGENCE STRATEGY AND
IMPLEMENTATION PATHWAY

Prepared by the DoD Responsible AI Working Council in accordance with the memorandum issued by Deputy Secretary of Defense Kathleen Hicks on May 26, 2021, Implementing Responsible Artificial Intelligence in the Department of Defense.
June 2022

June 2022

Outlines the Department’s Strategy for Operationalizing the Ethical Principles

Six Tenets:

1. **RAI Governance:** Modernize structures for continuous oversight.
2. **Warfighter Trust:** Achieve justified confidence through training and education and test and evaluation and verification and validation.
3. **AI Product and Acquisition Life Cycle:** Identify and mitigate risks throughout life cyle.
4. **Requirements Validation:** Ensure AI systems are aligned with operational needs.
5. **Responsible AI Ecosystem:** Promote shared understanding through domestic and international engagements.
6. **AI Workforce:** Build, train, equip, and retain an RAI-ready workforce.

CDAO

Examples of RAI Tools and Capabilities

RAI tools function in a number of ways to support the operationalization of DoD's AI ethical principles for capability developers, RAI practitioners, and senior leaders.

What	Function	Example Tools
Technical or Software Based	Helps developers and testers assess factors such as bias, reliability, and safety	Data Bias Detection Tools Explainability Tools T&E Harness
Documentation and Artifacts	Provides traceability of data sources, model limitations, risk identification, and mitigation efforts	Use Case/Harms Analysis Data Cards Model Cards
Frameworks and Checklists	Provides prompts to guide users in creating muscle memory around new processes for risk assessment and ethical considerations	Common Failure/Mishap List Algorithmic Impact Assessments Ethics Maturity Assessments User Research and Design Tools
Knowledge Sharing	Provides centralization for information sharing, learning, and common lexicon, practices, etc.	Use Case Repositories Information Management Systems
Executive Dashboards	Provides visibility into organizational compliance, status, and risk	Key Performance Metrics Progress Tracking

RAI Toolkit

- **The Responsible AI toolkit is our organizing framework to make the capabilities being built out under the RAI Strategy & Implementation Pathway:**
  - Findable
  - Usable
  - Interoperable
- **Living document and web application (currently in minimum viable product form) building upon and incorporating:**
  - Industry best practices and tools (currently 70+ listed in the toolkit) and academic innovations
  - DIU RAI Guidelines and Worksheets, NIST AI RMF + Playbook, IEEE 7000, etc.
  - Tools being built through the RAI Strategy and Implementation Pathway

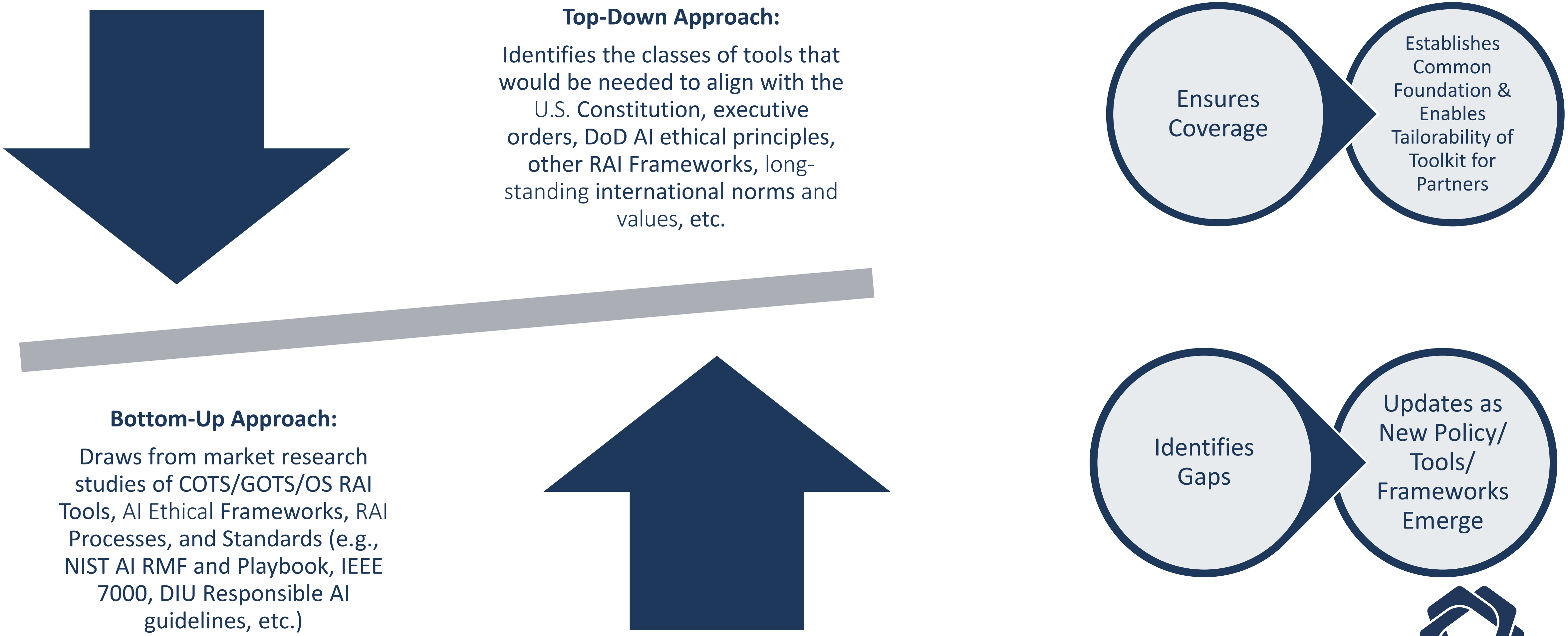
CDAO

RAI Toolkit Priorities

- **Provides a process for demonstrating consistency/alignment with the DoD AI ethical principles**
- **Enables traceability and promotes assurance**
- **Provides a mechanism for collecting lessons learned that can serve as inputs to policy**
  - Enables empirical tracking of how RAI influences mission success
- **Provides common framework for partners and allies to develop shared assurance cases**
  - Aids interoperability and trust
  - Is co-developed NATO version of Toolkit
  - Is developing collaborations over the toolkit with IC, interagency, other allies and partners

CDAO

Approach to Toolkit



Top-Down Approach:

Identifies the classes of tools that would be needed to align with the U.S. Constitution, executive orders, DoD AI ethical principles, other RAI Frameworks, long-standing international norms and values, etc.

Bottom-Up Approach:

Draws from market research studies of COTS/GOTS/OS RAI Tools, AI Ethical Frameworks, RAI Processes, and Standards (e.g., NIST AI RMF and Playbook, IEEE 7000, DIU Responsible AI guidelines, etc.)

Ensures Coverage
Establishes Common Foundation & Enables Tailorability of Toolkit for Partners
Identifies Gaps
Updates as New Policy/ Tools/ Frameworks Emerge

CDAO

Design Challenges and Principles

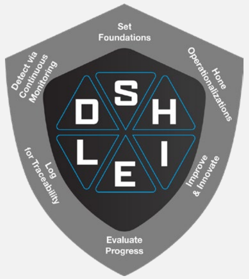
The RAI Toolkit aims to seamlessly assist users to plan and execute the necessary RAI activities and select appropriate supporting artifacts and tools.

Challenges		Principles
Wide diversity of use cases and priorities across the DoD.	▶	**Modular and Tailorable**
Demonstrate alignment with DoD AI ethical principles.	▶	**Traceable**
Existing assessment processes can overwhelm a small team.	▶	**Lightweight**
RAI processes require coordination among diverse team roles and stakeholder considerations.	▶	**Holistic**
RAI activities should take place during all phases of AI development.	▶	**Integrated**
Existing approaches assume expert RAI knowledge.	▶	**Upskilling**
RAI research and practice is still evolving.	▶	**Iterative (Living Document)**

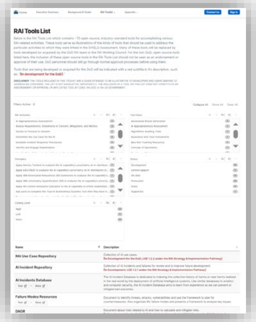
CDAO

RAI Toolkit Components

Currently Available in Toolkit MVP



Planning Tools
Identify and document potential risks and plan RAI activities for mitigation



RAI Tools List

Below is the RAI Tools List which contains ~70 open-source, industry-standard tools for accomplishing various RAI-related activities. These tools serve as illustrations of the kinds of tools that should be used to address the particular activities to which they were linked in the SHIELD Assessment. Many of these tools will be replaced by tools developed (or acquired) by the DoD RAI team or the RAI Working Council. For the non-DoD, open-source tools listed here, the inclusion of these open-source tools in the RAI Tools List should not be seen as an endorsement or approval of their use; DoD personnel should still go through normal approval processes before using them.

Tools that are being developed or acquired for the DoD will be indicated with a red subtitle in it's description, such as: "(In development for the DoD)".

DISCLAIMER: [illegible]

Filters Active : 0

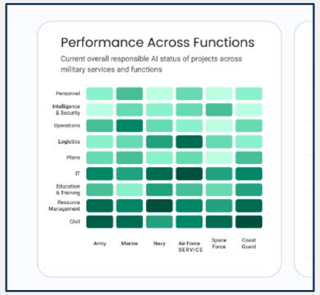
Name	Description
RAI Use Case Repository	Collection of AI use cases. [In Development for the DoD; LOE 1.2.2 under the RAI Strategy & Implementation Pathway]
AI Incident Repository	Collection of AI incidents and failures for review and to improve future development. [In Development; LOE 1.2.1 under the RAI Strategy & Implementation Pathway]
AI Incidents Database	The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes.
Failure Modes Resources	Document to identify threats, attacks, vulnerabilities and use the framework to plan for countermeasures. Also organizes ML failure modes and presents a framework to analyze key issues.
DAGR	Document about risks related to AI and how to calculate and mitigate risks. [illegible]

Tools and Resource Database
Provide resources for implementing RAI plan

Software Tools, Guidance and Best Practices, Checklists, Metrics

In Development

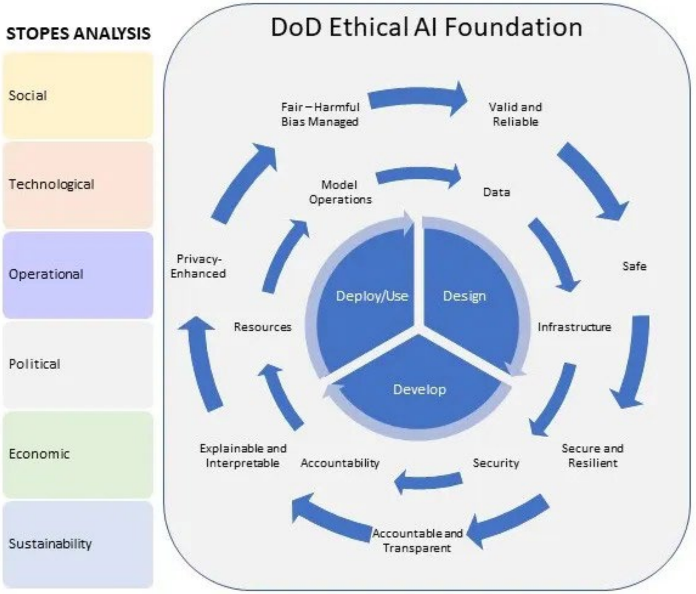
Evaluation Tools
Evaluate progress against RAI plan



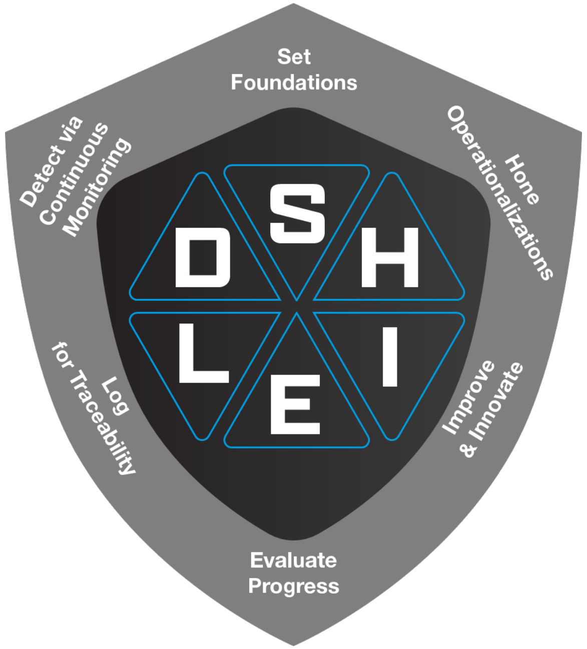
Oversight Dashboard
Monitor RAI progress and risk profile across programs and portfolios

The RAI toolkit assists users to plan and execute RAI activities, and select appropriate supporting artifacts and tools.

How: RAI Planning and Assessment



Initiates



Defense AI Guide on Risk Assessment

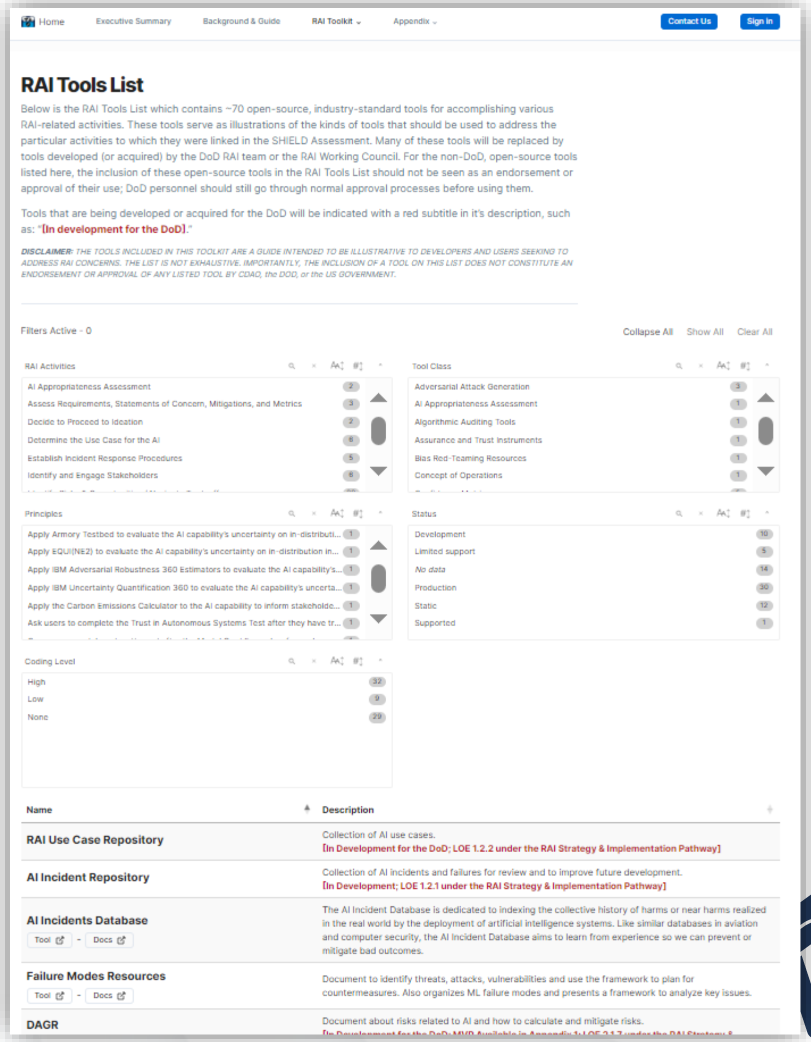
- Risk management guidance for DoD is aligned to NIST AI Risk Management Framework
- Risk management process initiates a SHIELD Assessment
- Supporting tools in development

SHIELD Planning Process

- A series of six sequential classes of activities identifies RAI-related issues for tracking and mitigation
- List of issues are tracked throughout the life cycle via statements of concern
- Elements in the SHIELD assessment route the user to relevant tools within the tools database

Tools and Resource Database MVP

- **Searchable Database (70+ items) of COTS/GOTS/open-source RAI Tools:**
  - Informed by CDAO market research and RAI FY22 tool survey
    - Industry best practices and tools (70+)
    - Academic methodologies
    - DIU Responsible AI guidelines & worksheets
    - NIST AI RMF + Playbook
    - IEEE 7000
- **Customizable User Interface:**
  - Tailorable labels for ethical principles, development lifecycle phases, category names, roles, and disciplines
  - Interactive search and exploration



RAI Tools List

Below is the RAI Tools List which contains ~70 open-source, industry-standard tools for accomplishing various RAI-related activities. These tools serve as illustrations of the kinds of tools that should be used to address the particular activities to which they were linked in the SHIELD Assessment. Many of these tools will be replaced by tools developed (or acquired) by the DoD RAI team or the RAI Working Council. For the non-DoD, open-source tools listed here, the inclusion of these open-source tools in the RAI Tools List should not be seen as an endorsement or approval of their use; DoD personnel should still go through normal approval processes before using them.

Tools that are being developed or acquired for the DoD will be indicated with a red subtitle in it's description, such as: "**[In development for the DoD]**."

DISCLAIMER: *THE TOOLS INCLUDED IN THIS TOOLKIT ARE A GUIDE INTENDED TO BE ILLUSTRATIVE TO DEVELOPERS AND USERS SEEKING TO ADDRESS RAI CONCERNS. THE LIST IS NOT EXHAUSTIVE. IMPORTANTLY, THE INCLUSION OF A TOOL ON THIS LIST DOES NOT CONSTITUTE AN ENDORSEMENT OR APPROVAL OF ANY LISTED TOOL BY CDAO, the DOD, or the US GOVERNMENT.*

Filters Active - 0

RAI Activities	
AI Appropriateness Assessment	2
Assess Requirements, Statements of Concern, Mitigations, and Metrics	3
Decide to Proceed to Ideation	2
Determine the Use Case for the AI	6
Establish Incident Response Procedures	5
Identify and Engage Stakeholders	6

Tool Class	
Adversarial Attack Generation	3
AI Appropriateness Assessment	1
Algorithmic Auditing Tools	1
Assurance and Trust Instruments	1
Bias Red-Teaming Resources	1
Concept of Operations	1

Principles	
Apply Armory Testbed to evaluate the AI capability's uncertainty on in-distributi...	1
Apply EQUI(NE2) to evaluate the AI capability's uncertainty on in-distribution in...	1
Apply IBM Adversarial Robustness 360 Estimators to evaluate the AI capability's...	1
Apply IBM Uncertainty Quantification 360 to evaluate the AI capability's uncerta...	1
Apply the Carbon Emissions Calculator to the AI capability to inform stakeholde...	1
Ask users to complete the Trust in Autonomous Systems Test after they have tr...	1

Status	
Development	10
Limited support	5
No data	14
Production	30
Static	12
Supported	1

Coding Level	
High	32
Low	9
None	29

Name	Description
RAI Use Case Repository	Collection of AI use cases. **[In Development for the DoD; LOE 1.2.2 under the RAI Strategy & Implementation Pathway]**
AI Incident Repository	Collection of AI incidents and failures for review and to improve future development. **[In Development; LOE 1.2.1 under the RAI Strategy & Implementation Pathway]**
AI Incidents Database Tool - Docs	The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes.
Failure Modes Resources Tool - Docs	Document to identify threats, attacks, vulnerabilities and use the framework to plan for countermeasures. Also organizes ML failure modes and presents a framework to analyze key issues.
DAGR	Document about risks related to AI and how to calculate and mitigate risks. **[In Development for the DoD; MVP Available in Appendix 1; LOE 2.1.7 under the RAI Strategy &**

Who: RASCI and Personas List

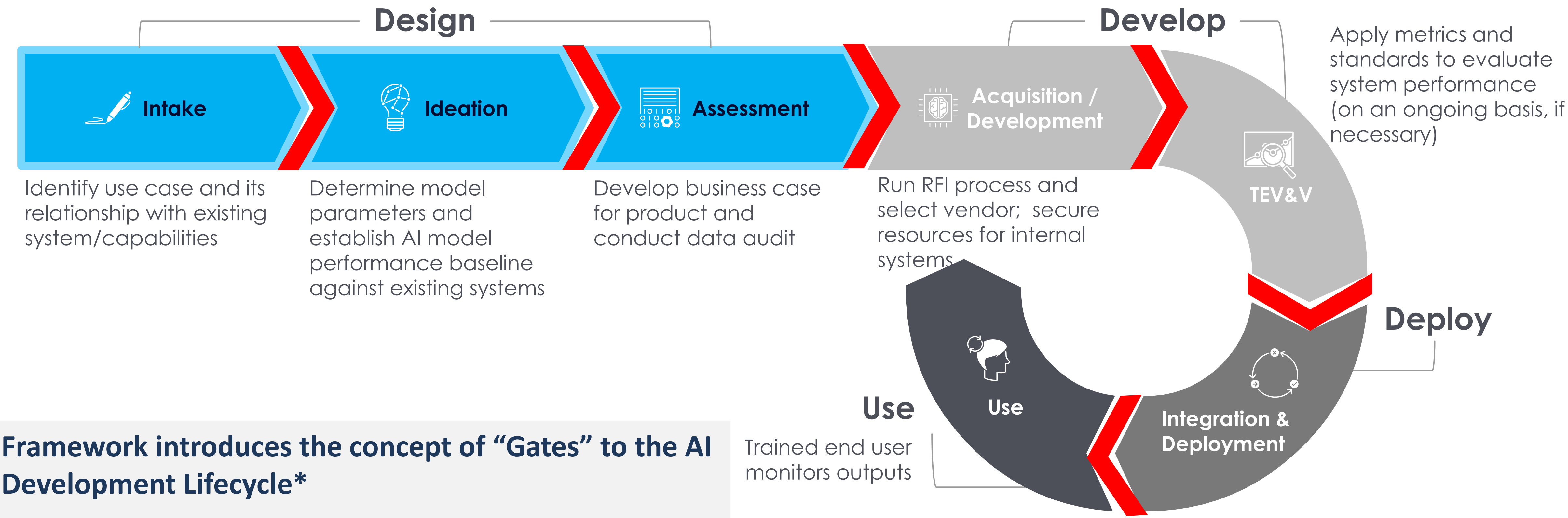
RAI Role**
Users/Stakeholders
Mission Commanders
Senior Leader/AI Innovation Leader
Functional Requirements Owner
Program Manager
AI Ethics and Risk Specialist
Relevant Legal, Ethical, or Policy Expert
UX/Design/HMT/AI Adoption Specialist
AI Development Team System Architect Data Architect Data Operations Specialist Data Analyst Data Scientist Data Officer AI Engineer/AI/ML Specialist Data Steward
AI Test and Evaluation Specialist
IT/Cyber Expert

RASCI matrix is built for each role to clarify RAI tasking

Role*	Definition
Responsible	The person who does the work to complete the task or create the deliverable.
Accountable	The person ultimately accountable for the work or decision being made; this person gives final approval.
Supporting	The person who provides support for those who are responsible or accountable; participates in doing the work of a task.
Consulted	Anyone who must be consulted with or add input prior to a decision being made and/or the task being completed.
Informed	The people who need to be updated on project status or informed when a decision is made or work completed.

**Individuals or teams may be dual hatted;
Roles map to DoD Cyber Workforce (DCWF) roles – **BLUE text indicates relevant DCWF Role**

When: RAI Development Life Cycle



Identify use case and its relationship with existing system/capabilities

Determine model parameters and establish AI model performance baseline against existing systems

Develop business case for product and conduct data audit

Run RFI process and select vendor; secure resources for internal systems

Apply metrics and standards to evaluate system performance (on an ongoing basis, if necessary)

Trained end user monitors outputs

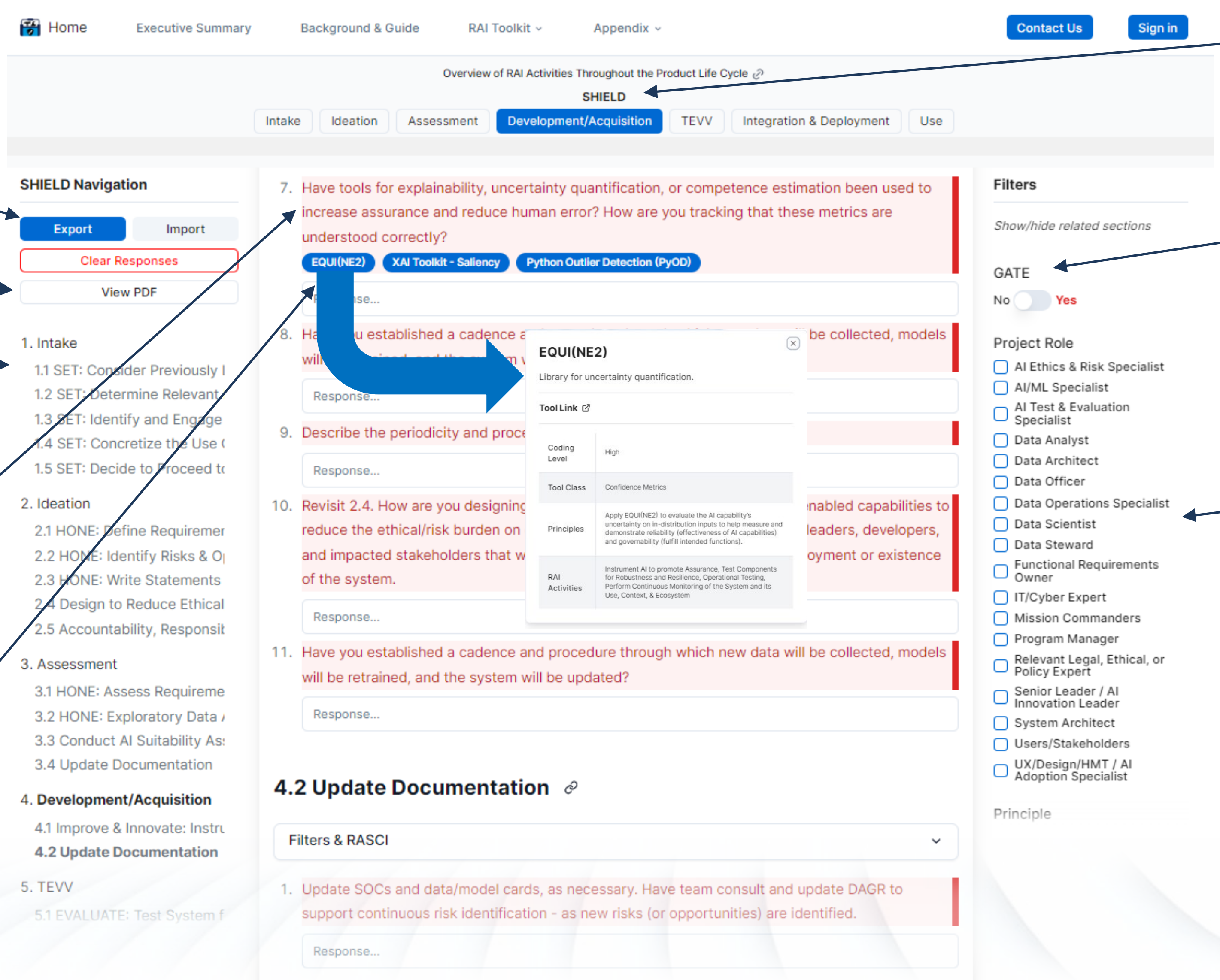
Framework introduces the concept of "Gates" to the AI Development Lifecycle*

- "Gates" indicate recommended considerations for progression to the next phase of development

**DoD RAI Strategy & Implementation Pathway (p13)*

CDAO

RAI Toolkit Current Features



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 tc
2. Ideation
   - 2.1 HONE: Define Requiremer
   - 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: Instru
   - **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?

...se...

8. Ha... u established a cadence a... be collected, models wil...

Response...

9. Describe the periodicity and proce...

Response...

10. Revisit 2.4. How are you designing ...nabled capabilities to reduce the ethical/risk burden on ... leaders, developers, and impacted stakeholders that w... oyment 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...

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

4.2 Update Documentation

Filters & RASCI

1. Update SOCs 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...

Filters

Show/hide related sections

GATE

No ○ Yes

Project Role

- □ AI Ethics & Risk Specialist
- □ AI/ML Specialist
- □ AI Test & Evaluation Specialist
- □ Data Analyst
- □ Data Architect
- □ Data Officer
- □ Data Operations Specialist
- □ Data Scientist
- □ Data Steward
- □ Functional Requirements Owner
- □ IT/Cyber Expert
- □ Mission Commanders
- □ Program Manager
- □ Relevant Legal, Ethical, or Policy Expert
- □ Senior Leader / AI Innovation Leader
- □ System Architect
- □ Users/Stakeholders
- □ UX/Design/HMT / AI Adoption Specialist

Principle

Navigation by AI Product Lifecycle Stage

Export/Import function to save and share progress

Export as PDF

Navigation by Type of RAI Activity

"GATE" filter (displays most essential assessment questions)

Filters assessment questions by persona/project role, AI ethical principle, discipline, etc.

SHIELD Assessment identifies risk and opportunities

Links to tools to address identified risks and opportunities

CDAO

RAI Toolkit Web Application

RAI Toolkit Web App:
https://rai.tradewindai.com/

Toolkit Way Ahead

- **Develop versions of toolkit to support approvals and reviews for various use cases**
  - Deconflict with other required processes and documentation to support creation of integrated template or documentation process.
  - Use Tabletop Exercises/'Mock Reviews' to refine documentation process.
- **Develop versions of toolkit focused on generative AI/LLMs**
- **Pilot on other use cases throughout DoD, interagency, international partners**
  - Codevelop shared versions of the toolkit with partners.
  - Collect, organize, and share lessons learned.
- **RAI Toolkit tabletop exercises & technical exchanges with key allies & partners to aid interoperability**
- **Develop acquisitions-focused version of toolkit**
- **Integrate into AI training courses**
- **Continue to add functionality**
  - Develop user interface.
  - Add further tailorability features (data & model type, use case, risk profile, etc.).
  - Continue dashboard development.
  - Integrate with other tools (T&E, cyber) and with platforms.
  - Integrate feedback.

CDAO

Closing Thoughts

> "…ultimately, AI systems only work when they are based in trust. We have a principled approach to AI that anchors everything that this Department does. We call this Responsible AI, and it's the only kind of AI that we do. Responsible AI is the place where cutting-edge tech meets timeless values." - General Lloyd J. Austin III, Secretary of Defense

Thank You and Questions

Drew Brooks

Lead Scientist for Responsible AI Tools

U.S. Department of Defense

andrew.l.brooks.civ@mail.mil

RAI Toolkit Web App:
https://rai.tradewindai.com/

CDAO