



U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND ARMAMENTS CENTER

**AI4SE & SE4AI:
Assessing Consistency with Ethical Principles**

SEP 2026

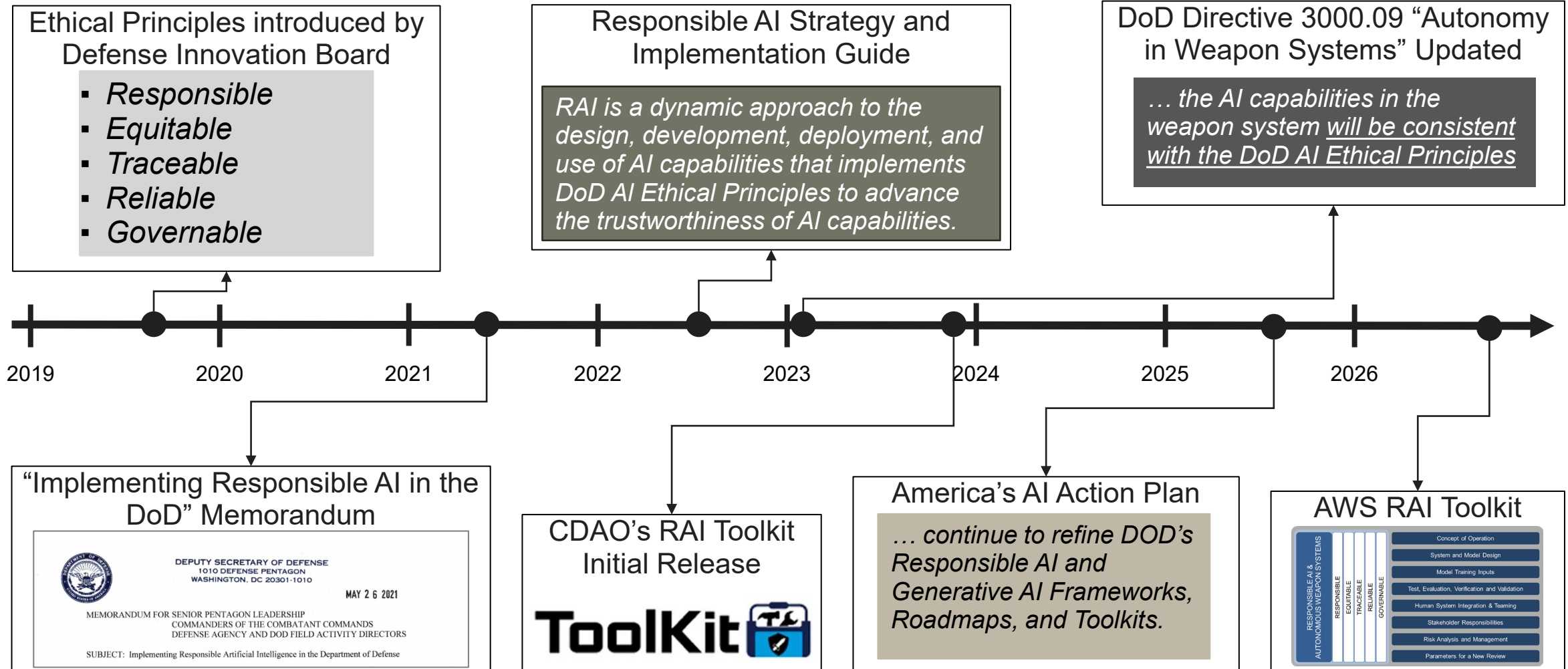
MR. BENJAMIN WERNER, DEVCOM AC



AGENDA

- Responsible AI Timeline
- Challenge
- Ethical Principles
- Areas of Consideration
- Toolkit Sample
- Toolkit Results
- Path Forward

RESPONSIBLE AI



CHALLENGE/PROBLEM STATEMENT



- Continued reinforcement from policy and guidance for adherence to Responsible AI and the Ethical Principles
- No explicit means to demonstrate consistency as currently required in DoDD 3000.09
- General confusion leading to delay in pursuit of capabilities given uncertainty of how to assess this requirement
- Objective: develop guidance and means to assess programs

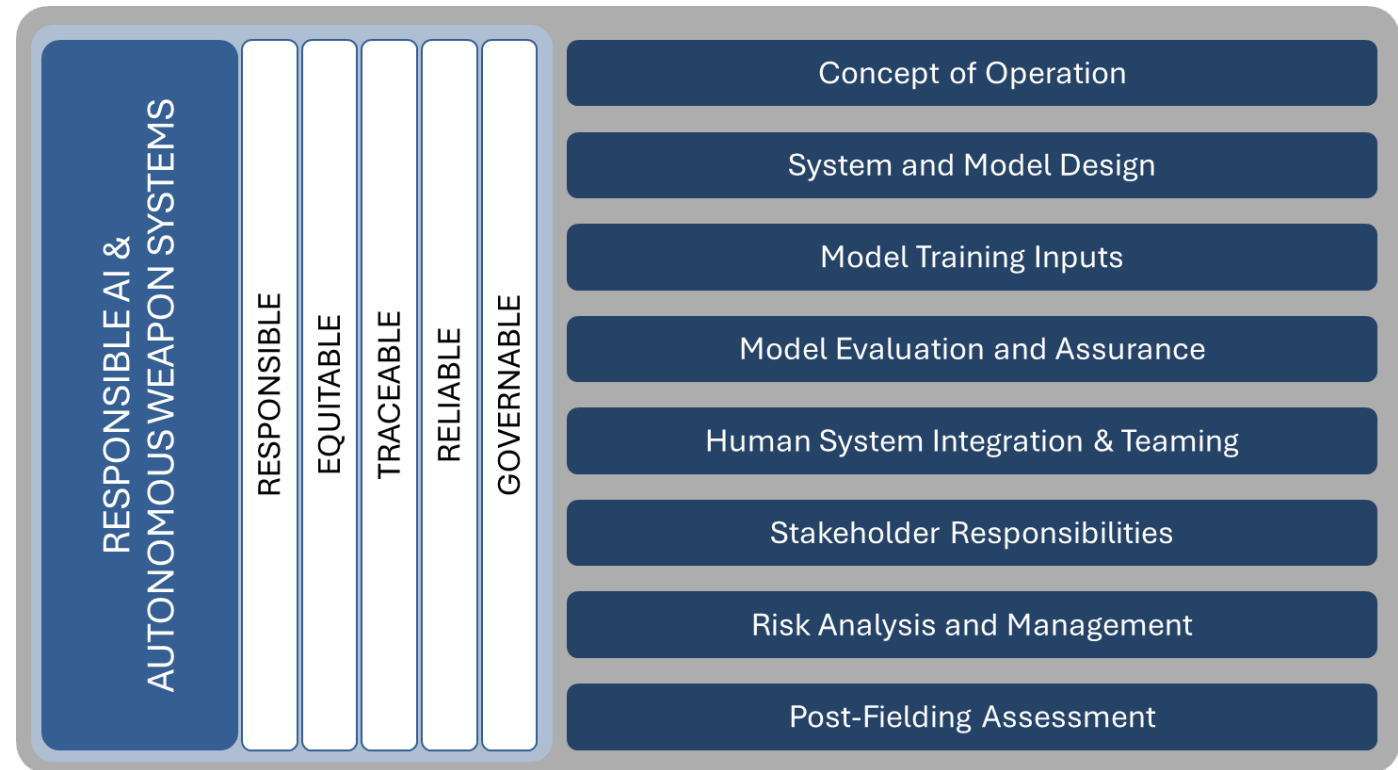
DoDD 3000.09: Guidelines for Reviews of Certain Autonomous Weapon Systems

For systems incorporating AI capabilities, the deployment and use of the AI capabilities in the weapon system will be consistent with the DoD AI Ethical Principles and the DoD Responsible AI Strategy and Implementation Pathway.

INTRODUCTION TO THE TOOLKIT



- The Autonomous Weapon System (AWS) Responsible AI (RAI) Tool delivers an adherence framework for developers and programs.
 - Practical guide for preparing for mandatory senior-level reviews before formal development and fielding.
 - Comprehensive look across the lifecycle of development and deployment
 - Leverages activities, artifacts, and evidence collected during capability development to demonstrate adherence to AI Ethical Principles.



Assess AWS AI Capability's Consistency with AI Ethical Principles

ETHICAL PRINCIPLES



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 DoD will take deliberate steps to minimize unintended bias in AI capabilities.

Traceable.

The DoD'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 procedures and documentation.

Reliable.

The DoD'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 DoD will design and engineer AI capabilities to fulfill their 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.

KEY AREAS OF CONSIDERATION



Concept of Operation: Defines use cases, operational environments, and human decision points. Explores sensors and data used by the AI model, and accounts for risks to non-targets.

System and Model Design: Details where AI components operate and inputs/outputs. Includes configuration management plans and design aspects to mitigate unintended engagements.

Model Training Inputs: Ensures data suitability for intended domains. Covers provenance documentation, bias mitigation, and data maintenance procedures.

Model Evaluation & Assurance: Focuses on TEVV (Test, Evaluation, Verification, Validation). Assesses models against perturbations, red-teaming, and cyber security threats.

Human System Integration: Ensures AI outputs are effectively understood by operators. Develops training plans focusing on AI capabilities, limitations, and potential human/model errors. Evaluates operator acceptance and confidence.

Stakeholder Responsibilities: Defines roles for operators, commanders, and maintainers. Ensures appropriate human judgment is maintained and authorizations are clearly delegated.

Risk Analysis and Mitigation: Provides risk matrices with planned mitigation actions (accept, transfer, avoid). Communicates risks to stakeholders and tracking disposition actions.

Post-Fielding Assessment: Plans to monitor AI use for misuse, emergent behavior, or degradation. Defines capability and operational bounds that, if breached, necessitate action to be taken to include new review.

AWS RAI TOOLKIT SAMPLE SECTION



Tool developed leveraging prompts from AWS RAI Paper

Intent is for assessments based on programmatic documentation and artifacts for Ethical Principle consistency

Scores calculated by Principle and on aggregate

Responses are close-form dropdowns to enable assessment

Concept of Operation (Pre-Fielding)

Are use cases and CONEMP clearly defined for the AI?

The operation of the AWS, and any embedded AI, should be aligned to the intended use cases under which the DoDD 3000.09 review will take place. The intended use cases and CONEMP should define the expected operation of the AWS and AI. Given the autonomous nature of the capability, and the challenges with executing a full coverage test suite, designing and deploying the system bounded in its operation to the expected use cases and operational conditions may help mitigate associated risks. It informs operators about scenarios that may impact operational performance and reliability.

Concept of Operations (Pre-Fielding)	Response	Rationale/Evidence & Corresponding Artifacts	Related Ethical Principle(s)
1. Provide the final concept of employment (CONEMP) to include appropriate use cases and operational environments	Can answer with evidence/artifacts	CONOP/OV1	Reliable, Traceable
a. Define where human operators make decisions and (where applicable) any opportunities for intervention	Can answer anecdotally	CONEMP, HSI Assessment	Responsible, Governable
b. Define what functions the AI may perform in the system/mission CONOPS to include where in the kill chain it may function *	The program is engaged in addressing this	Functional Architecture, State Diagram	Reliable, Traceable
c. Describe how potential users were involved in the development of the CONEMP or CONOPS	Unsure if this information is available	CDD, Soldier Touchpoints	Traceable
2. What sensors and data feed into the AI, how were these represented or replicated in development and testing	There is no intent to address this	System Architecture, Data Assurance	Equitable, Reliable
a. Is the system dependent on any other capabilities to provide data once deployed or are all sensors/data ingestion capabilities organic to the system?	Not Applicable	CONEMP, System Design	Reliable
3. Where applicable, describe how the CONOP impacts how the AI capabilities may (or may not) account for risks to non-targets, consistent with commander and operator intent	Can answer with evidence/artifacts	Risk Assessment	Responsible, Traceable, Reliable

TOOLKIT IN OPERATION



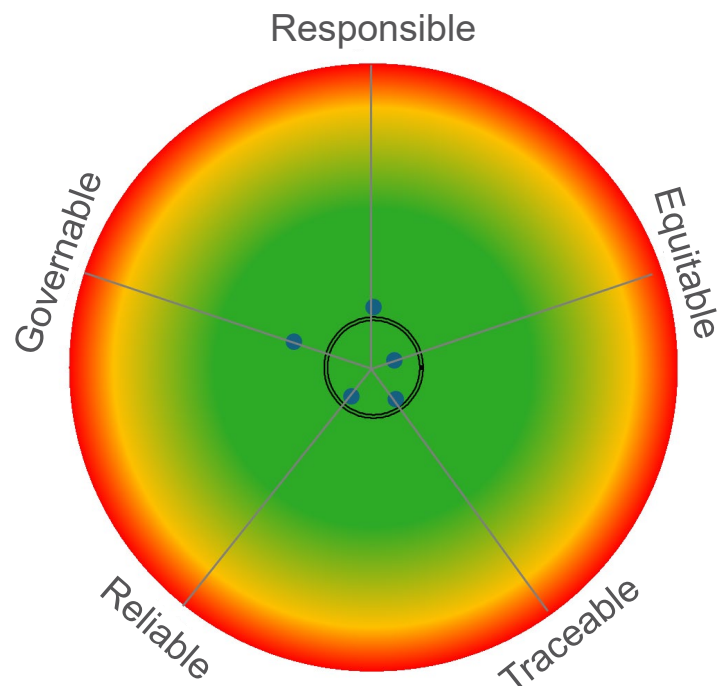
- Users review prompts for Requested Information
- Consider available resources, artifacts, and program documentation
 - Design Specifications
 - System Architectures
 - Design Reviews
 - Test Plans
- Select response from dropdown menu and submit justification or corresponding artifacts to support selection
- As selections are made the results chart update to reflect the inputs

Response	
Select Response ...	▼
Select Response ...	
Can answer with evidence/artifacts	
Can answer anecdotally	
The program is engaged in addressing this	
Unsure if this information is available	
There is no intent to address this	
Not Applicable	

AWS RAI TOOLKIT SAMPLE RESULTS



Assessment of Consistency



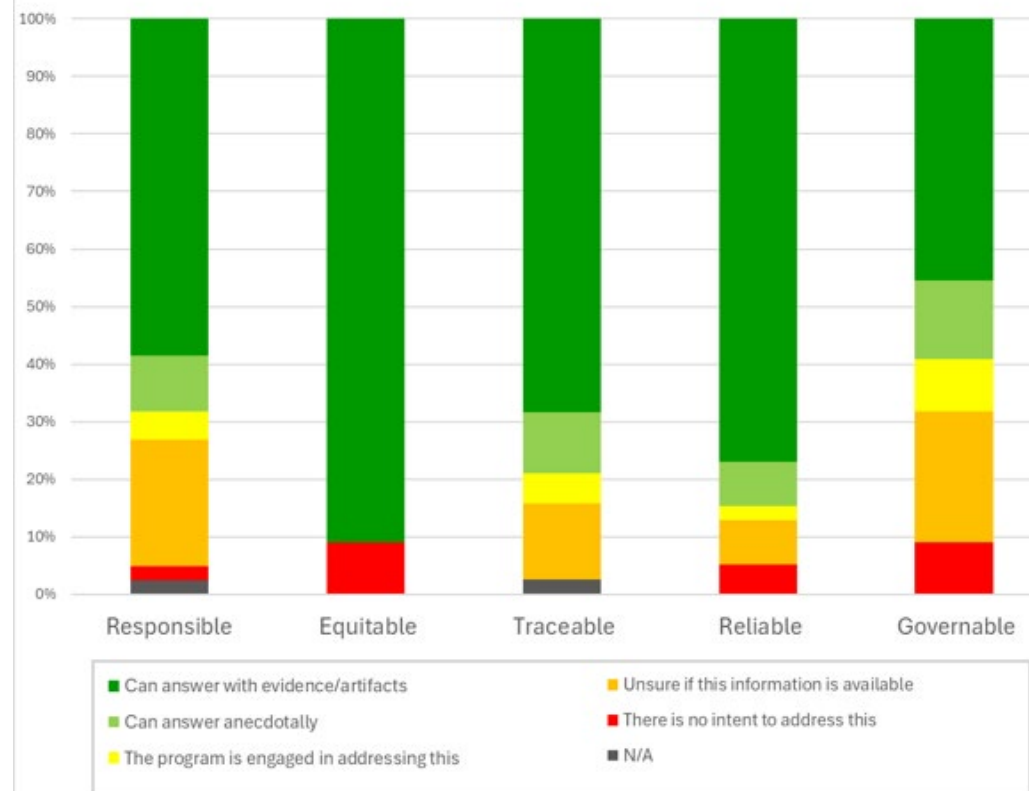
More Consistent - - - - - Less Consistent



● Consistency by Ethical Principle

○ Average Across Ethical Principles

Results by Principle



Results generated assess consistency by Ethical Principle
Identify where there are inconsistencies to correct

PATH FORWARD



- AWS RAI Toolkit (ART) delivers a capability to meet DoDD 3000.09 requirements
 - Grounded in engineering and assurance best practices
 - Enables the development and delivery of emerging technology to the warfighter
- Demonstrate to programs and engage stakeholders
- Ensure consistency to Ethical Principles through engineering best practices
- Investigate scalability to any AI enabled capability, beyond narrow AWS scope
- Pursue expanding guidance to align with NSPM-11 Assurance objectives:
 - Reliability
 - Robustness
 - Controllability
 - Steerability