



U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND ARMAMENTS CENTER

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

SEP 2026

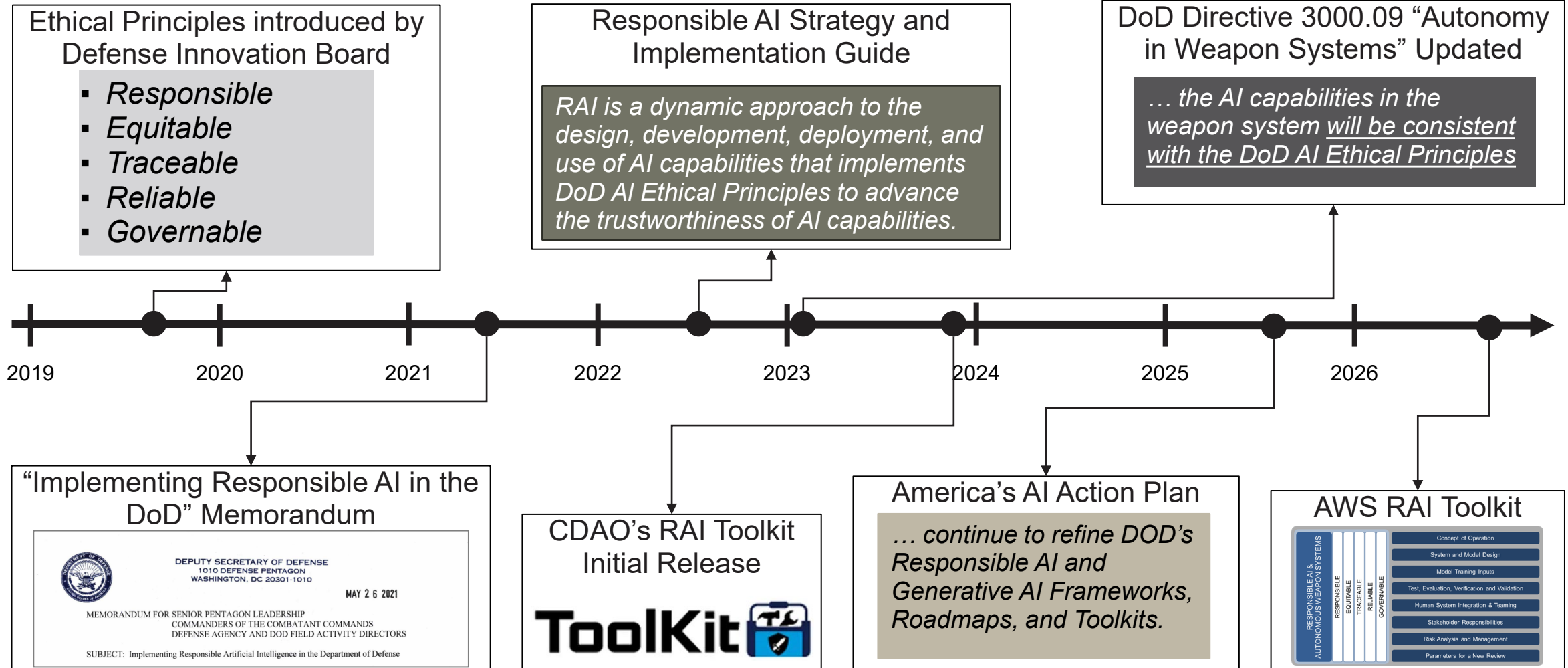
MR. BENJAMIN WERNER, DEVCOM AC



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

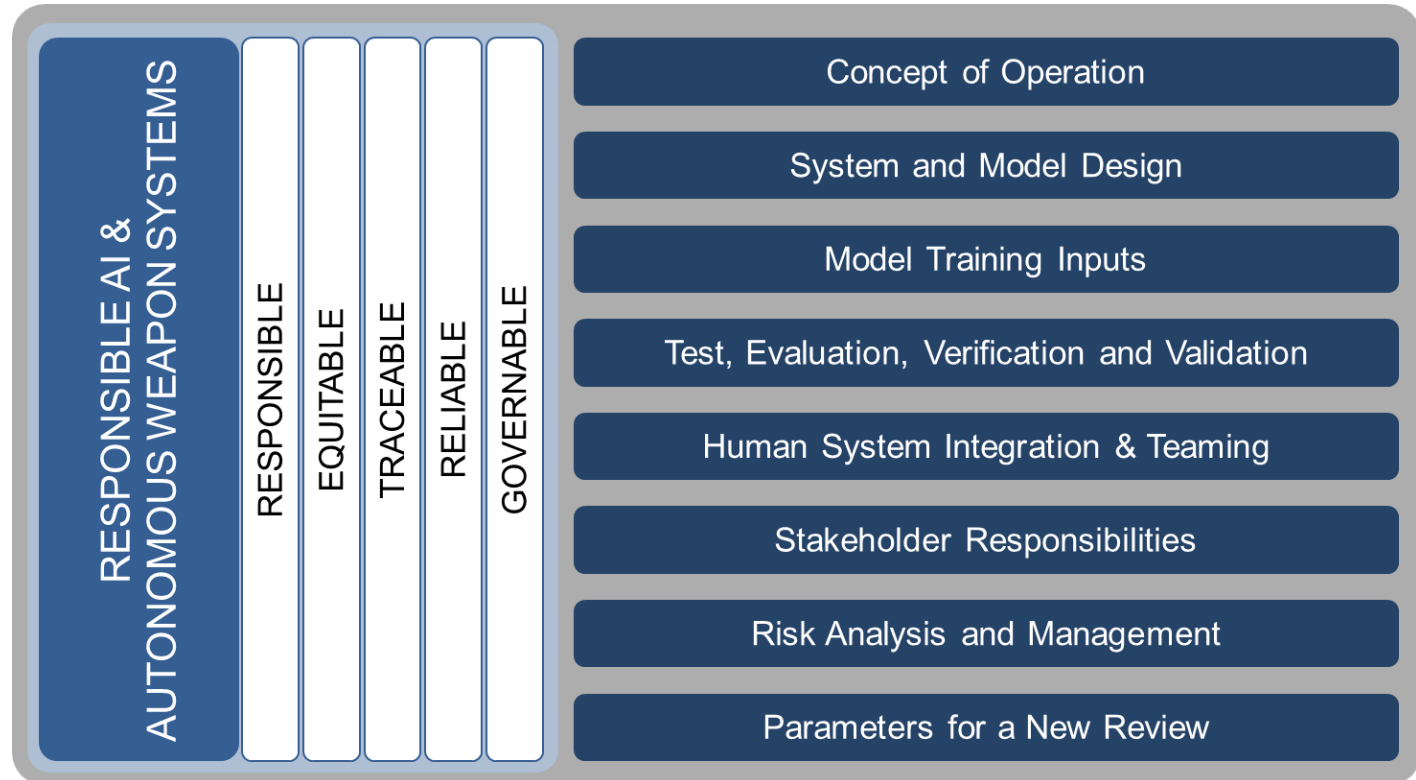
DoDD 3000.09: Guidelines for Reviews of Certain Autonomous Weapon Systems

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



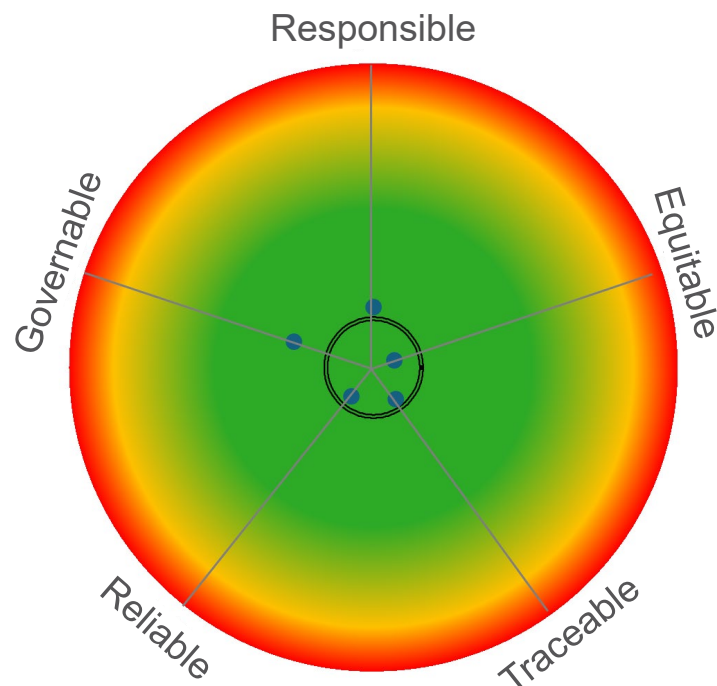
- Users review prompts for Requested Information
- Consider available resources, artifacts, and program documentation
 - Design Specifications
 - System Architectures
 - Design Reviews
 - Test Plans
- Select response from dropdown menu and submit justification or corresponding artifacts to support selection
- 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



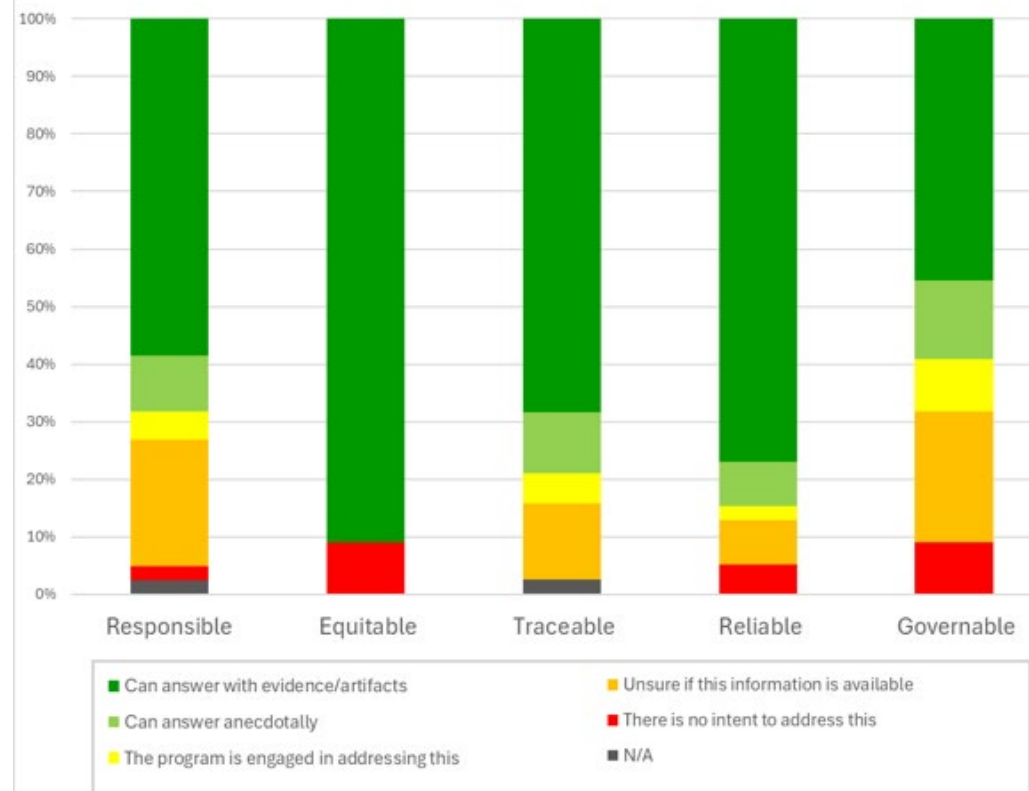
More Consistent - - - - - Less Consistent



● 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

BACKUP INFORMATION



CONCEPT OF OPERATION



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.

Recommended Information

1. Provide the concept of employment (CONEMP) to include appropriate use cases and operational environments.
 - a. Define where human operators make decisions and (where applicable) any opportunities for intervention.
 - b. Define what functions the AI may perform in the system/mission CONOPS to include where in the kill chain it may function.
 - c. Describe how potential users may have been involved in the development of the CONEMP or CONOPS.
 - d. Will the capability be transitioning to or operated by another organization? Has this organization accounted for the possible deltas in CONOP and mission use?
2. What sensors and data is used by the AI model, how were these represented or replicated in development and testing?
 - 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?
3. Where applicable, describe how the CONOP enables monitoring or assessment of nominal system performance/actions, and continued alignment to defined operational domain.
4. Where applicable, describe how the CONOP accounts for risks to non-targets consistent with commander and operator intent.

SYSTEM AND MODEL DESIGN



What is the system and AI model design?

An overall system design implements the intended functions and capabilities defined by the user needs and requirements. The AI model may be considered an integral component of the system design for the purposes of an AWS review. Therefore, the design and functions of an AI model should include the integration platform and system, which may introduce additional limitations and constraints. The model design should also account for requirements to maintain alignment with external system policies and directives. A state or flow diagram that explains the autonomy and AI interactions can highlight critical decision points, actions, or effects of importance for senior reviewers. While the overall AWS system design is identified as a critical piece in the directive policy, AI model design will still inform those same requirements. That can include the responsibility to ensure the model is governable in accordance with ethical principles.

Recommended Information

1. Provide the system design and identify where in the system the AI components operate, to include the inputs/outputs to the AI model from the system enablers (e.g. sensors).
 - a. Provide a sequence graph, flow chart, or similar process artifact that outlines where the AI is performing functions.
 - b. Identify AI-enabled functions that impact system safety and associated SME assessments.
 - c. How are the inputs filtered or limited or the model outputs monitored or regulated?
2. Provide the model design as a separate yet integrated component within the overall system design.
 - a. At each step where the AI is performing a function identify the model type and how it's appropriate for the function.
3. Provide the plan for configuration management of AI model updates.
 - a. What is the process to update AI models when needed or required?
 - b. Are there particular time, event, or performance triggers for updating the model?
4. Where applicable, describe how the integrated system design enables monitoring or assessment of nominal system performance/actions, and continued alignment to defined operational domain.
5. Where applicable, describe how the integrated system design accounts for risks to non-targets, consistent with commander and operator intent.

MODEL TRAINING INPUTS



How were the model training inputs sourced, collected, documented, and updated?

Model training inputs define the development and operation of an AI-enabled system. These can range from images to tabular data to rules and rewards for learning agents. As such, it is imperative to properly define and utilize those inputs to ensure they are aligned with, and traceable to, commander and operator intent, the intended CONEMP, defined functions, the system performance, and general reliability and resilience. There is a direct impact on the executed functions of the autonomous weapon system from the AI models that are rooted in the training inputs. Assessment and assurance of the suitability and quality of the model training inputs is critical and aligns with adherence of the ethical principles and DoDD 3000.09.

Recommended Information

1. Describe the measures taken to ensure that the model training inputs used for AI development were suitable and applicable for each function as employed within the intended domain, use cases and tasks of the AWS.
2. Show how the breadth of the inputs represents the suite of operational conditions it is expected to be employed in.
3. Describe how the inputs were developed, validated against ground truth, and split between train/test/validate where applicable.
4. How have the source, provenance, and key parameters of the model inputs (to include who, what, where, when, and why) been documented?
5. What is the method and frequency by which the model training inputs are maintained, updated, and appropriately documented and configuration controlled to ensure timeliness, relevance, and suitability?
6. What is the method for how these model training inputs were collected and stored in a secure manner?
7. Who owns the data and on what system are models trained (USG v CTR)?
8. What steps were taken to detect and mitigate unintended biases and control for sources of human bias (e.g., in data selection and preparation)?
9. Provide the results of any assessments on the model training inputs with respect to functions and CONEMP for quality, suitability, and completeness.

MODEL EVALUATION & ASSURANCE



What plans are in place for evaluation and assurance of the AI models?

Test and evaluation along with verification and validation (TEVV) provides evidence of model performance to deliver assurance to operators and evaluators that the system is capable and suitable for the intended operation. TEVV plans, often formally documented in a T&E strategy, establish plans to assess reliability, effectiveness, and suitability under realistic operational conditions. With respect to AWS, this would include conditions with adversarial or red team kinetic, cyber, and environmental actions that could impact model inputs yielding unexpected or corrupted outputs that would impact system functions in the intended CONEMP. Including TEVV throughout the lifecycle enables continuous validation of the models so that risks can be addressed and mitigated. The ability for post-fielding evaluation through software hooks enables stakeholders to monitor if models continue to perform as expected once deployed into operational environments. Successful TEVV can assess and assure adherence to elements of the ethical principles.

Recommended Information

1. Provide the plan(s)/strategy(ies) that outline the test and evaluation process for the AI model.
 - a. Describe how the AI components were tested, to include in isolation as well as with enabling components (e.g. sensors) across representative conditions where applicable, for performance, safety, maintainability, and suitability, to include robustness, resilience and reliability
 - b. Describe what steps were taken to detect and mitigate bias in the model that would impact function performance within the intended CONEMP.
 - c. Detail how the model was assessed against perturbations, adversarial attack, drift, poisoning, hallucinations or other operating challenges as expected in the CONEMP.
 - d. What metrics were assessed and are they relevant for the intended mission?
2. Provide the cyber security assessment results of the AI to include relevant testing and analysis.
3. Provide results that verify AI performance and resilience to counter AI red-teaming to include relevant testing and analysis.
4. Provide cyber security review and approval authorizations as appropriate.
5. For systems where the AI model can be updated, provide the results of the verification and validation for the update process.

HUMAN SYSTEM INTEGRATION AND TEAMING



What interactions do humans have with the AI and how are the humans trained and evaluated?

Human system integration (HSI) and teaming are critical in the design, development, deployment, and operation of responsible autonomous weapon systems. AWS must be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force. For that reason, it is imperative that part of the design and development of the system includes the development of training and evaluation of users' trust and understanding of the system's capabilities and limitations to enable appropriate judgement. Some AWS variations may have human user interfaces, experiences, or overwatch requirements and capabilities. For these it is necessary to ensure such integrations consider the user's needs in their design such that they can effectively team with the developed and deployed capabilities when applicable. The following information can help to ascertain how the integration of operators throughout the lifecycle processes enables the ethical principles and provides appropriate judgement on use.

Recommended Information

1. Describe any Human Readiness Level analysis and assessment of the AI system.
2. Describe, when applicable, any human interaction with the system once employed or activated, for example decision making, setting selection, observation, intervention.
3. Outline, where applicable, a procedure for when automated decisions or activities of the AI will require human approval or to be marked out for human oversight.
4. If the system provides AI outputs to an operator (as part of deciding to deploy/employ or as part of monitoring), how is it ensured and evaluated that the model outputs are appropriately understood by a human operator?
5. Identify, where applicable, features, to include training and TTPs, that may provide information to operators about the AI's capabilities, limitations, and included targets to help the operator make the appropriate decision concerning employment.
6. Describe how stakeholders and potential operators were included in the development process to design and incorporate TTPs and CONEMP specific constraints into the user workflow to ensure appropriate trust, understanding, reliance, usability, and confidence of the capabilities and limitations of an AI-enabled system.
7. Define the procedure and scope for user testing throughout the lifecycle and how operator acceptance, confidence and trust in the AI-enabled system was evaluated and maintained.
8. Provide the training plan for operators on the understanding of AI-enabled components.
 - a. Does the associated operator training or UI/UX design support understanding of potential sources of human or model error, and constraints & limitations of the AI and AWS?
 - b. How are the operational and environmental limitations of the system, specifically the AI capabilities, communicated to the operators or commanders?

STAKEHOLDER RESPONSIBILITIES



What are the roles and responsibilities of the stakeholders in development, deployment, and use?

The roles and responsibilities of stakeholders, across administrative, developmental, and operational domains, will directly impact how the AI and AWS can be expected to function and how it may be assured. Most critically for an AWS, DoDD 3000.09 cites the requirement for “appropriate levels of human judgement” in the use of force. Information from this section will inform senior leaders of how that is established through the roles and responsibilities of the commander and operators. However, the commander and operator are not the only stakeholders that share responsibility in the utilization of weapon systems as part of Responsible AI. There are stakeholders during development, such as model and data engineers, as well as during review processes, such as test engineers and review authorities, that have roles and responsibilities. The individual or organization responsible for the development and review of the capabilities should have roles identified and defined. This enables accountability to ensure solutions are expected to be governable.

Recommended Information

1. Provide concept of employment (CONEMP) to include clear indication of role and responsibility of operators, commanders, and maintainers of the system.
 - a. Clearly articulate how “appropriate levels of human judgement” will be actioned through responsible stakeholders in operations to the operator or commander in the use of force (employment) of the system.
2. Define, if applicable, what stakeholders and who is responsible for the monitoring of model performance, requesting and/or performing updates to the model, evaluating the performance of the updated model and expected impact if re-deployed under the intended operating conditions, and re-deploying the model to the fielded capability.
3. How will information about the system, its risks and benefits, and significant changes in its development, configuration, or use be communicated to relevant stakeholders, at the appropriate level of detail?
 - a. Identify the responsible role or organization for the development of the AWS and for each AI component (if different).
 - b. Identify the responsible role (at the lowest level) for the deployment (fielding) of the AWS.
 - c. Identify the responsible role for the authorization of use of the AWS and if it may be delegated under particular conditions.
 - d. Identify the responsible role (at the lowest level) for the activation (use) of the AWS.

RISK ANALYSIS AND MANAGEMENT



What are the risks and how are they tracked and mitigated?

Risk analysis and management provide a method to identify potential issues, analyze possible outcomes, and address them. Risks can impact any element during the program lifecycle, to include all areas discussed in this paper. Through this analysis, various issues or potential negative effects can be identified and addressed early through program or system adjustments and validated with stakeholders and risk managers for proper disposition. The tracking of risks helps ensure proper traceability and implementation during the program lifecycle. This tracking can be used to develop or inform other programmatic documentation such as design plans, test strategies, and operational usage. The AI risk analysis should be supplemental to an existing risk management process for a program or project, and many recommended activities and information may already exist. Risks related to AI or DoDD 3000.09 requirements can be flagged as part of the larger risk tracking.

Recommended Information

1. Provide the final risk matrix for existing and potential AI risks – this may be a matrix tracing identified risks with corresponding disposition actions, artifacts/documents, and status.
 - a. Identify risk disposition (e.g. accept, mitigate, transfer, avoid), who made the disposition, and any associated technical or functional authority involved.
 - b. Identify actions (planned or taken) to mitigate risks (e.g. design changes, training/doctrine updates, assurance activities, or TEVV plans/results).
2. Provide means for communicating risk to stakeholders.

POST FIELDING ASSESSMENT



What needs to be monitored post-fielding to ensure consistent performance with expectations, avoid unintended consequences, and anticipate need for additional review?

Once the system has been fielded, monitoring will be critical to ensure the AI maintains performance, is not exhibiting troubling emergent behavior, is robust to new environments, or can execute different missions. The program should define a set of bounds and criteria that if breached, action may be taken. That action may be a change in operation or training, or a update or patch to a model. DoDD 3000.09 outlines a requirement to execute another review when a modification substantially differs from the previous approval. It is critical to establish at the onset of the program these criteria for what would be substantial. This enables monitoring to ensure that such changes do not creep in to the system that may lead to undetected degradation in performance or invalidate an original review and approval. Defining criteria prior to fielding to be monitored in operation for action enables governable AI that can fulfill the intended functions.

Recommended Information

1. Identify the bounds on the mission capabilities that, if breached, may result in substantial changes to the system and/or mission under the original review.
 - a. Identify what may constitute substantial changes in mission set, e.g. target types, target sets, environments, host platform, adversarial countermeasures, or human interfaces.
 - b. Determine if the actions require a CONOPS or usage change or would require a technical change that would require a new 3000.09 Review.
2. Is there a plan in place to continuously monitor the use and performance of the AI models (e.g. for misuse, emergent behavior, or degradation) and take action once the system is deployed?
 - a. Define the evidence or behaviors that would indicate the AI-enabled functions have degraded to an unacceptable level due to environment or adversary changes.
 - b. Identify actions to take or how the system's training, use, or deployment may be altered to account for degraded or limited performance during operations.