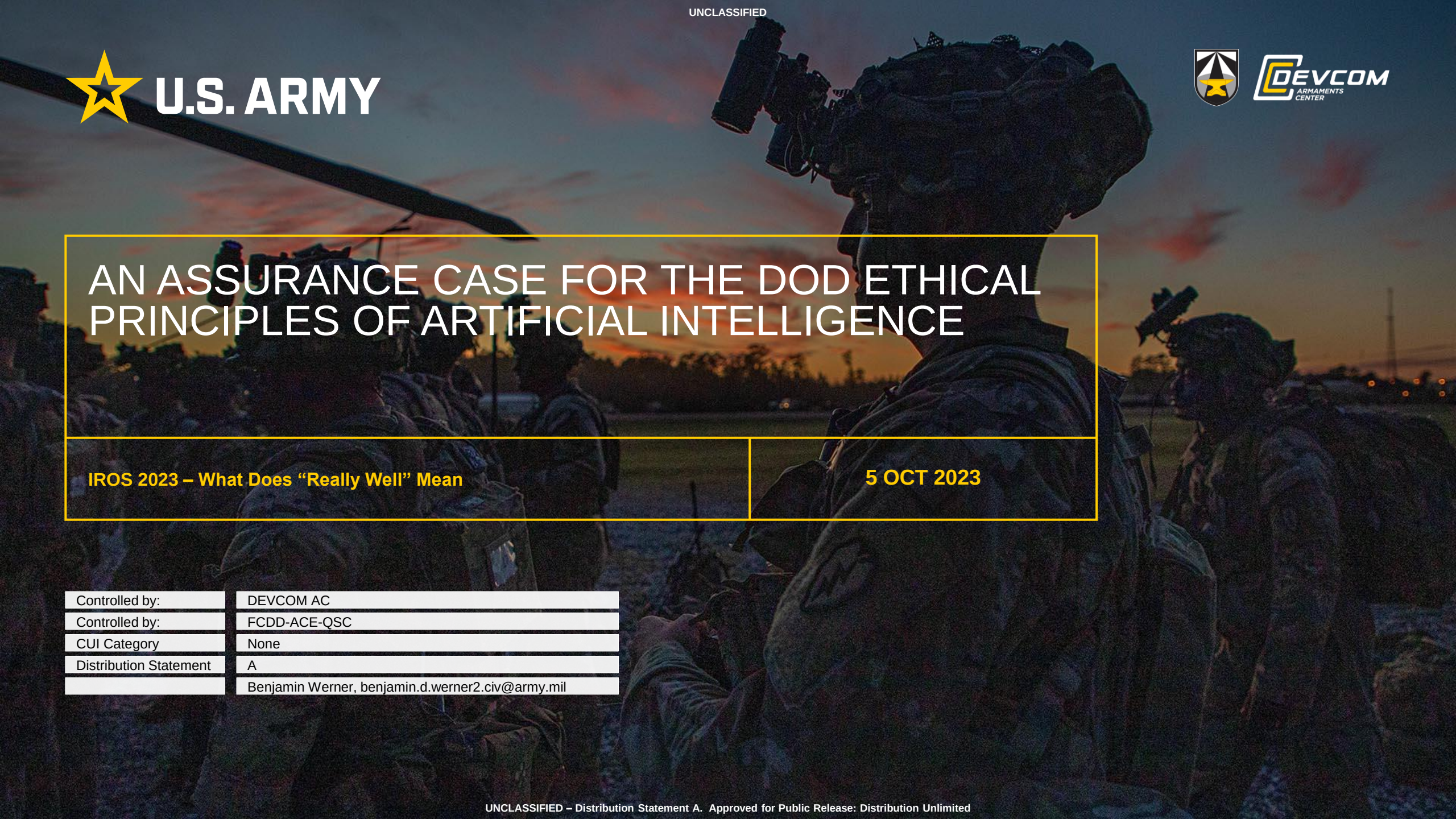


U.S. ARMY

DEVCOM ARMAMENTS CENTER

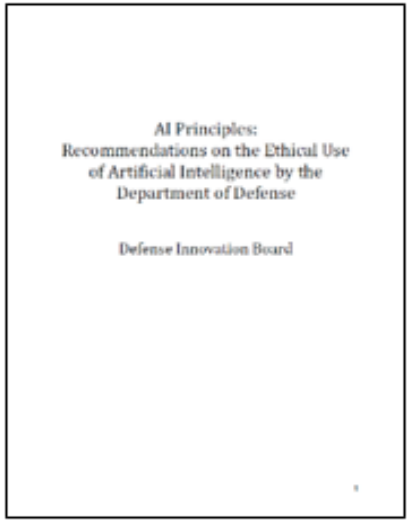
AN ASSURANCE CASE FOR THE DOD ETHICAL PRINCIPLES OF ARTIFICIAL INTELLIGENCE

IROS 2023 – What Does "Really Well" Mean

5 OCT 2023

Controlled by:	DEVCOM AC
Controlled by:	FCDD-ACE-QSC
CUI Category	None
Distribution Statement	A
	Benjamin Werner, benjamin.d.werner2.civ@army.mil

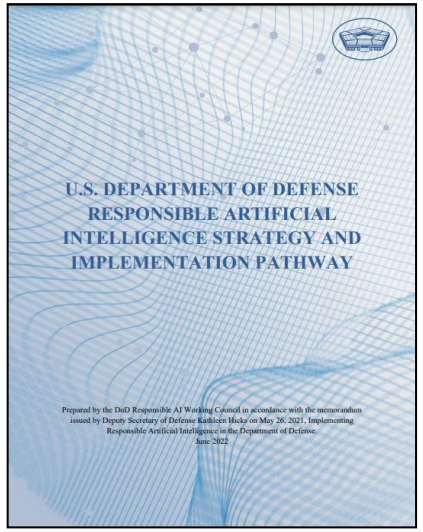
EVOLUTION OF ETHICAL PRINCIPLES



Defense Innovation Board Report (2019)
- Ethical Principles 1st Published

Responsible AI Memorandum (2021)
- Ethical Principles Championed
- Responsible AI Working Council

Responsible AI Strategy (2022)
- Tenets and Lines of Effort identified to embody the Ethical Principles



DOD ETHICAL PRINCIPLES OF AI

- ***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 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.*

MATERIEL RELEASE QUESTIONS & ARTIFACTS

PROCESS THAT CERTIFIES THAT ARMY MATERIEL
IS **SAFE, SUITABLE AND SUPPORTABLE** BEFORE ISSUED TO THE FIELD

SAFETY

Questions:

- Is the system safe?
- Have hazards to Soldiers, civilians, and equipment been identified and mitigated or accepted?
- Has AEC confirmed the system is safe?
- Have hazards related to health, EOD, energetics, or environment been identified and mitigated or accepted?

Artifacts:

- Safety Certification & Safety Data Package or Safety & Health Data Sheet
- AEC Safety Confirmation
- Mishap Risk Acceptance or System Safety Risk Assessment (SSRA)
- Health Hazard Assessment
- Surface Danger Zone
- ATEC Assessment or Evaluation
- Final Hazard Classification
- Army Fuze Safety Review Board Certification
- Energetic Materials Qualification Board Statement
- EOD Support Statement
- Environmental Support Statement
- Nuclear Regulatory Commission Licensing
- Air Worthiness Release
- Ignition System Safety Review Board Certification
- Hazards of Electromagnetic Radiation to Ordnance (HERO) Certification

SUITABILITY

Questions:

- Is the system suitable?
- Does the system meet requirements?
- Has the system been evaluated by ATEC? Do they concur?
- How will it function in operational setting?
- Does the system have sufficient reliability for intended missions?
- Have cyber security vulnerabilities been identified and mitigated?
- Has the software been assessed?
- Can the system be used on the network and interface?
- Are TIR/PCRs documented and resolutions effective?
- Have physical and functional configuration audits been conducted?

Artifacts:

- ATEC Assessment or Evaluation
- Quality and Reliability Statement
- Army Interoperability Certification
- Risk Management Framework
- Software Quality Statement
- Human Systems Integration (HSI) Assessment

SUPPORTABILITY

Questions:

- Is the system supportable?
- Has the sustaining command approved of the plan?
- How will software be supported?
- Has test and diagnostic equipment been identified?
- Has training been developed and approved?
- What is the fielding plan?
- Have the Gaining Commands been notified of the system that will be fielded?

Artifacts:

- Proof of TC-STD
- Logistics Certification from Sustainment Organization
- Software Supportability Statement
- Test, Measurement and Diagnostic Equipment (TMDE) Support Statement
- Signed Materiel Fielding Agreement (MFA)/Materiel Fielding Plan (MFP)/ Memorandum of Notification (MON)
- Training Assessment from Capability Developer

RESPONSIBLE

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

- Materiel Release process ensures responsible risk mitigation across domains
- Materiel Release is a requirement for Army systems prior to fielding or deployment
- Upon deployment systems may only be used for the intended form, fit, and function, use outside of these bounds would be cause for a new evaluation

Materiel Release Artifacts

Safety

- Safety Certification & Safety Data Package or Safety & Health Data Sheet
- AEC Safety Confirmation
- Mishap Risk Acceptance or System Safety Risk Assessment (SSRA)
- Health Hazard Assessment
- Surface Danger Zone
- Final Hazard Classification
- Army Fuze Safety Review Board Certification
- Energetic Materials Qualification Board Statement
- EOD Support Statement
- Environmental Support Statement
- Nuclear Regulatory Commission Licensing
- Air Worthiness Release
- Ignition System Safety Review Board Certification

Supportability

- Proof of TC-STD
- Army Logistician Assessment
- Logistics Certification from Sustainment Organization
- Software Supportability Statement
- Test, Measurement and Diagnostic Equipment (TMDE) Support Statement
- Materiel Fielding Agreement /Materiel Fielding Plan /Memorandum of Notification Training Assessment from Capability Developer

Suitability

- ATEC OMAR/OER Support Statement
- Quality and Reliability Statement
- Army Interoperability Certification
- Risk Management Framework
- Networthiness Certification
- Software Quality Statement

EQUITABLE

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

- There are multiple artifacts and deliverables to measure system bias and performance
  - Operational Test Agency (OTA) Milestone Assessment Report, OTA Evaluation Report
    - Test and Evaluation community report on operational performance of the system, would capture observed biases
  - Quality Assurance / Reliability Availability Maintainability (QA/RAM) Statement
    - Documents system reliability across the spectrum of expected operating conditions
    - Would identify and measurable bias in performance
    - Bias can be interpreted as a failure, intent of reliability engineering is to mitigate failures
- Design for Assurance tools and practices promote building in equitability
  - Design of AI/ML is predicated on the training data sets
  - Risk and readiness analysis of data sets to include diversity, distribution and completeness

TRACEABLE

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.*

- Materiel Release elements include Manpower/Personnel evaluations
  - Correct Military Occupational Series (MOS)
  - Identify proper skillsets, training, knowledge
- Program of instruction for all stakeholders
- Human System Integration (HSI) assessment completed
  - Understand how and where system is capable or limited
  - Develop mental models to make clear to users the data and methodology used
  - Measures/metrics of human interaction, capabilities
- Training evaluated and exercised in a Logistics Demonstration

RELIABLE

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.*

- Reliability: probability that the system will perform without failure over a specific interval, under specified conditions
  - Documented in OMAR/OER and QA/RAM Statement
  - Measured against specific use cases for testing and assurance
- Safety: ensuring that hazards to human, equipment and environment have been mitigated to acceptable levels
  - Documented through Safety Certification, System Safety Risk Assessment, Safety Confirmation
  - Critical requirement for armaments systems
- Security: cybersecurity is a critical element of Materiel Release
- Effectiveness, testing and assurance
  - Testing for effectiveness is documented through the OMAR/OER
  - Assurance in reliability/safety/security throughout the lifecycle

GOVERNABLE

Governable. *The Department 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.*

- Design - Design for Assurance / Reliability / Safety
- Intended functions – high reliability
- Detect and Avoid – leverage reliable, safe, robust design practices
- Unintended consequences – reliability failures, safety hazards
- Disengage or deactivate – leverage redundancies, guardrails, governors, mitigations
- Unintended behavior – reliability failures, safety hazards

SUMMARY

- DoD Ethical Principles provides design guidance

- This design guidance can be embodied through the activities and deliverables required for Materiel Release

- The Materiel Release process presents an assurance case for the Ethical Principles

  – Processes and artifacts already in place to ensure Army develops responsibly
  – Identification of gaps can garner new initiatives and capabilities (ie data quality)

- Align assurance case concept to the new DoD Responsible AI Strategy

– Current activities and deliverables may present solutions to the documented tenets and lines of effort


U.S. ARMY