DoD Responsible AI (RAI) Strategy and Implementation Pathway

Carol J. Smith
Senior Research Scientist, Human-Machine Interaction
AI Division

cjsmith@sei.cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

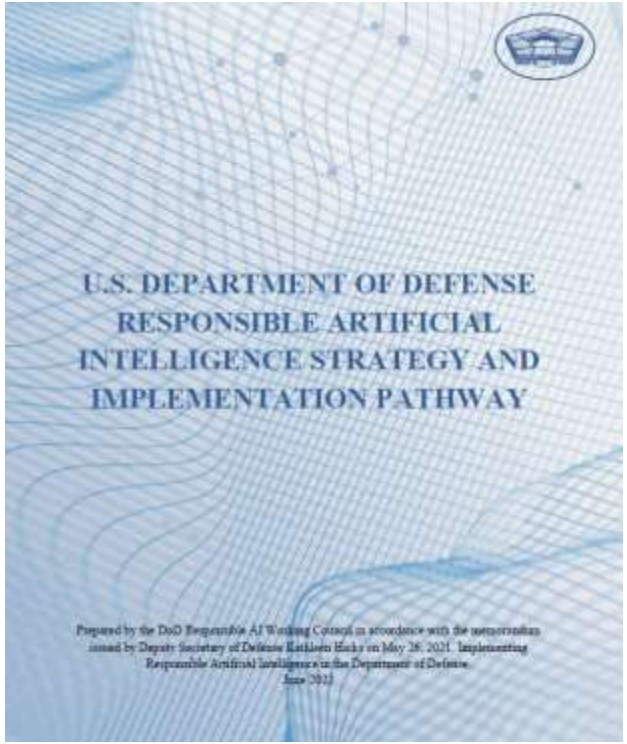
Copyright Statement

DoD: Responsible AI (RAI)

Responsible AI is the approach for how the Department must conduct AI design, development, deployment, and use.

RAI is a journey to trust.

This approach ensures the safety of DoD systems and their ethical employment.



Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).
PDF: https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

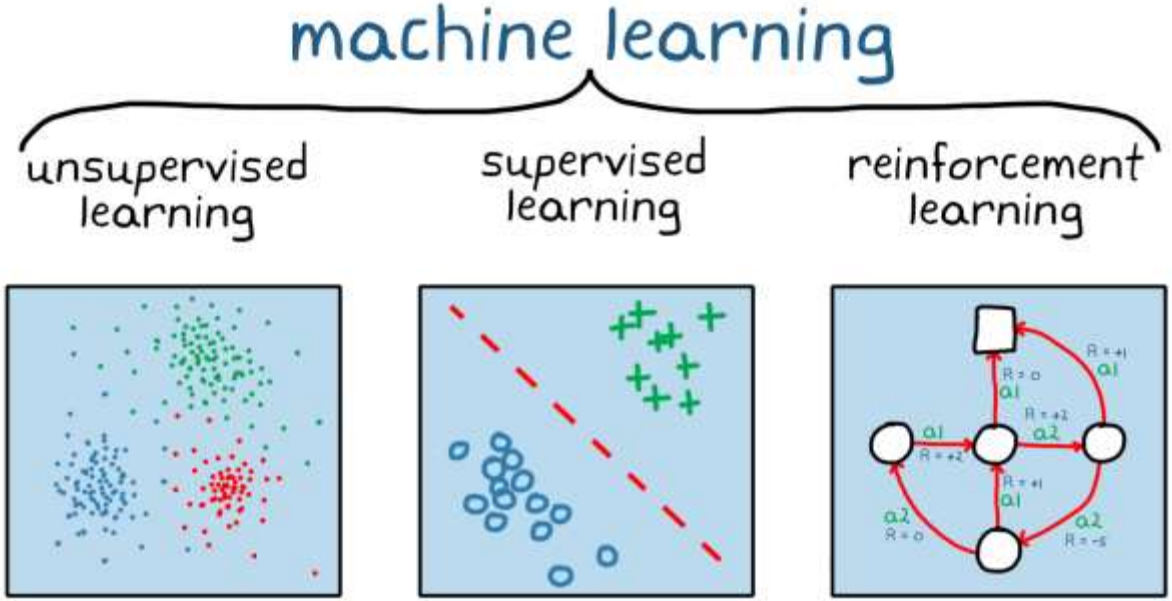
AI Primer

AI systems can

- recognize patterns
- create predictions
- make decisions, and/or
- generate new content.

Without being explicitly programmed to do so.

AI is typically created with machine learning (ML) methods



+ deep learning, neural networks, etc.

Image What Is Reinforcement Learning? 3 things you need to know. © 1994-2022 The MathWorks, Inc.
https://www.mathworks.com/discovery/reinforcement-learning.html

Machine Learning

Requirements for ML

1. **Data**: pre-existing, machine readable, relevant (amount vary)
2. **Math**: appropriate for data and context (statistics, probability, calculus…)
3. **Programming**: Python, C/C++, R, Java, JavaScript…

Math + programming = algorithm

Data + algorithm = ML model*

*The term model is used inconsistently. Model sometimes refers to an algorithm without data.

Responsible AI (RAI)

Data are the core of AI systems

Data are collected and curated for a reason (bias).

The AI system is being created for a reason (bias).

All systems have some form of bias. Complete objectivity is misleading.

Bias can be unintentional or have purpose and be helpful.

Goal of RAI: Reduce unintended, unwanted, and/or harmful bias.

DoD Ethical Principles for Artificial Intelligence

- Responsible
- Equitable
- Traceable
- Reliable
- Governable

IMMEDIATE RELEASE

DOD Adopts Ethical Principles for Artificial Intelligence

FEB. 24, 2020

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.

The recommendations came after 15 months of consultation with leading AI experts in commercial industry, government, academia and the American public that resulted in a rigorous process of feedback and analysis among the nation's leading AI experts with multiple venues for public input and comment. The adoption of AI ethical principles aligns with the DOD AI strategy objective directing the U.S. military lead in AI ethics and the lawful use of AI systems.

"The United States, together with our allies and partners, must accelerate the adoption of AI and lead in its national security applications to maintain our strategic position, prevail on future battlefields, and safeguard the rules-based international order," said Secretary Esper. "AI technology will change much about the battlefield of the future, but nothing will change America's steadfast commitment to responsible and lawful behavior. The adoption of AI ethical principles will enhance the department's commitment to upholding the highest ethical standards as outlined in the DOD AI Strategy, while embracing the U.S. military's strong history of applying rigorous testing and fielding standards for technology innovations."

DoD Memorandum, "Artificial Intelligence Ethical Principles for the Department of Defense." (Feb 2020)

"*RAI is a journey to trust*." - RAI S&I Pathway

Ensure AI programs are built with principles
of *fairness, accountability, and transparency*
at each step in the development cycle.

Deputy Secretary of Defense directed DoD officials
to "develop tools, policies, processes, systems, and guidance"
that ensure that AI technology systems
comply with ethical development principles
as part of acquisition policies.

- Memorandum, May 27, 2021

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).
Responsible AI Guidelines in Practice: Lessons Learned from the DIU AI Portfolio. Defense Innovation Unit. https://www.diu.mil/responsible-ai-guidelines

DoD's RAI approach is broad.

RAI manifests in…

- ethical guidelines,
- testing standards,
- accountability checks,
- employment guidance,
- human systems integration, and
- safety considerations.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

How does the DoD enable RAI?

DoD Ethical Principles

- Responsible
- Equitable
- Traceable
- Reliable
- Governable

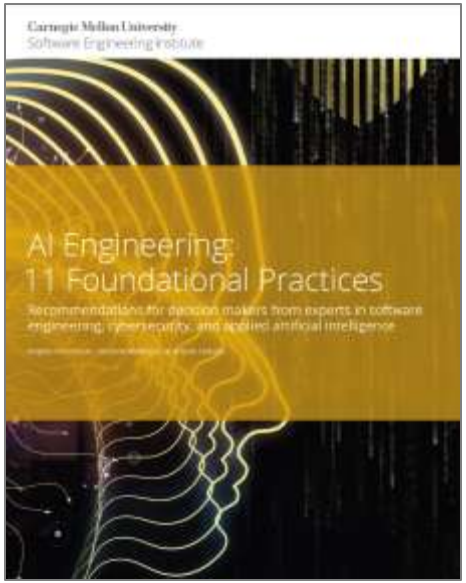
→ ? → Responsible AI

DoD Memorandum, "Artificial Intelligence Ethical Principles for the Department of Defense." (Feb 2020)

SEI’s RAI Efforts

2019: AI Engineering

AI Engineering: 11 Foundational Practices



Angela Horneman, Andrew Mellinger, and Ipek Ozkaya. 2019.
AI Engineering: 11 Foundational Practices. (September 2019)
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=633647

Framework and Checklist for Designing Ethical/Trustworthy AI

Carol J. Smith. 2019. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development.
arXiv:1910.03515 [cs] (October 2019). http://arxiv.org/abs/1910.03515

2021: SEI AI Engineering Qualities - White Papers

Scalable
Accommodate the size, speed, and complexity of mission needs

- Scalable management of data and models
- Enterprise scalability of AI development and deployment
- Scalable algorithms and infrastructure

Robust and Secure
Operate reliably when faced with uncertainty or threat

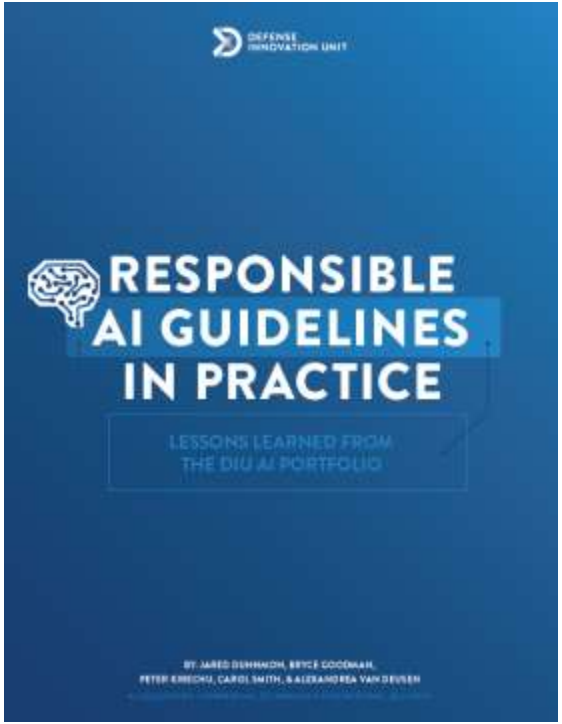
- Robustness of AI components and systems
- Designing for security challenges in modern AI systems
- Testing, evaluating, and analyzing AI systems

Human-Centered
Design with the goal of working with, and for, people

- Understand context of use, sense changes over time
- Scope and facilitate human-machine teaming
- Methods, mechanisms, and mindsets for critical oversight

SEI White Papers. June 2021. CMU Software Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=735452

2021-22: Defense Innovation Unit
RAI Report, Guidelines, Worksheets, and Workshops



Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

2021-22: ODNI - AI Engineering Research

Usable Hazard Analysis Processes for AI Engineering

Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro,[1] Carol J. Smith,[2] Tamara Zilovic[1]
HCI Institute - Carnegie Mellon University,[1] Software Engineering Institute, Carnegie Mellon University[2]
nikmart@cmu.edu, cjsmith@sei.cmu.edu, tzilovic@andrew.cmu.edu

Abstract

Embedding artificial intelligence into systems introduces significant challenges to modern engineering practices. Hazard analysis tools and processes have not yet been adequately adapted to the new paradigm. This paper describes initial research and findings regarding current practices in AI-related hazard analysis and on the tools used to conduct this work. Our goal with this initial research is to better understand the needs of practitioners and the emerging challenges of considering hazards and risks for AI-enabled products and services. *Our primary research question is: Can we develop new structured thinking methods and systems engineering tools to support effective and engaging ways for proactively considering failure modes in AI systems?* The preliminary findings from our review of the literature and interviews with practitioners highlight various challenges around integrating hazard analysis into modern AI development processes and suggest opportunities for exploration of usable, human-centered hazard analysis tools.

implications for the organizations that develop these products. While the use of new technologies always comes with the possibility of unintended consequences, we believe that many of these examples could have been prevented through strategic and thoughtful consideration when these systems are being designed and engineered.

Within systems engineering, strategies for hazard analysis can be used by teams to identify risks and potential failures with the goal of developing more robust and safe engineered systems. While many formal hazard analysis techniques exist, these activities largely center around helping teams determine potential risks and/or sources of failure *before* products have begun the development process. However, Rae et al. (2020) have called into question the utility of such methods as it is unclear if and how such methods are used by practitioners (Privan et al., 2017).

tinyurl.com/hazards-ai-eng

Nikolas Martelaro, Carol J. Smith, and Tamara Zilovic. 2022. Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering. Presented at 2022 AAAI Spring Symposium Series Workshop on AI Engineering: Creating Scalable, Human-Centered and Robust AI Systems. arXiv:2203.15628 [cs] (March 2022).

2022: SEI Representation, CMU Responsible AI Initiative

Carnegie Mellon University
Responsible AI

University-wide initiative at the Block Center.

Brings together researchers and educators across CMU to collaborate.

Carol Smith is on the CMU RAI Initiative Advisory Council and Interim Leadership Team: https://www.cmu.edu/block-center/responsible-ai/index.html

DoD RAI Strategy and Implementation Pathway: **Foundational Tenets**

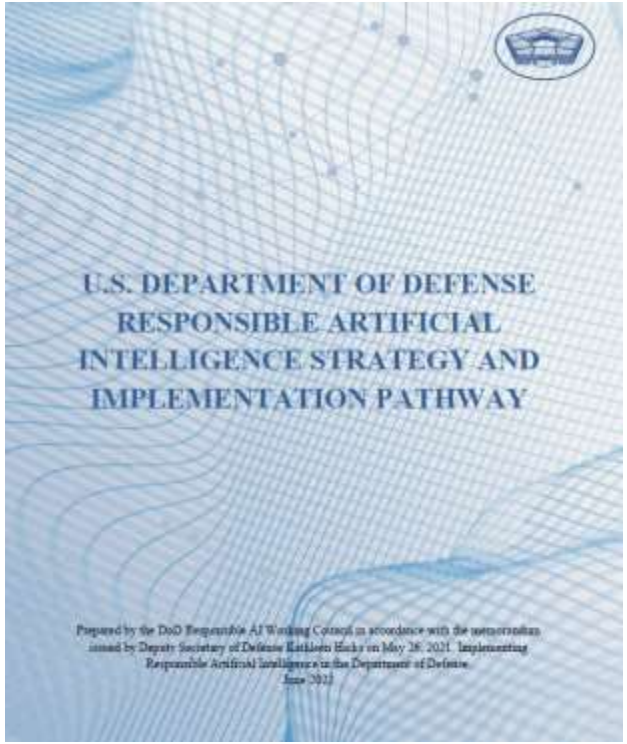
DoD: Responsible AI (RAI)

AI design, development, deployment, and use.

Journey to trust.

Ensures safety and ethical employment.

- Led by Office of the Chief Digital and Artificial Intelligence Officer (CDAO).
- SEI AI Division is supporting CDAO with multiple workplans.



Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

Foundational Tenets

DoD will implement RAI in accordance with 6 foundational tenets:

1. RAI Governance
2. Warfighter Trust
3. AI Product and Acquisition Lifecycle
4. Requirements Validation
5. Responsible AI Ecosystem
6. AI Workforce

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

1. RAI Governance

Goal: Modernize governance structures and processes that allow for continuous oversight of DoD use of AI, within context of use.

Topics of interest

- Measures of RAI adoption and progress
- Methods for users and developers to report concerns about implementation of the DoD AI Ethical Principles.
- Repository for exemplary AI use cases and regular knowledge-sharing of best practices and risk mitigation.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

2. Warfighter Trust

Goal: System operators gain tech familiarity and proficiency to achieve justified confidence in AI capabilities.

Topics of interest

- Education and training for warfighters
- Test, evaluation, verification, and validation (TEVV) framework to articulate how test and evaluation is intertwined across lifecycle
- Security and defense research and guidance (adversarial attacks)
- Integration of end-to-end real-time AI monitoring, confidence metrics, and feedback (explainable AI)

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

3. AI Product and Acquisition Lifecycle

Goal: Exercise appropriate care in the acquisition lifecycle

Topics of interest

- RAI evaluation, implementation, and continuous engagement
- Assess and understand bias and potential AI risks
- Mitigation planning to test and act on informed risk assessments

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

4. Requirements Validation

Goal: Ensure capabilities that leverage AI are aligned with operational needs while addressing relevant AI risks, for better traceability, accountability, and both internal and external oversight.

Topics of interest

- Process to develop AI requirements, incorporating RAI, that are testable, operationally relevant, and reusable.
- Repository of AI-related requirements with common use cases, mission domains, and system architectures.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

5. Responsible AI Ecosystem

Goal: Promote a shared understanding of RAI design, development, deployment, and use.

Topics of interest

- RAI tool development
- Share best practices on AI ethics, safety and trust in defense

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

6. AI Workforce

Goal: Ensure that all DoD AI workforce members possess an appropriate understanding of the technology, its development process, and the operational methods applicable to implementing RAI.

Topics of interest

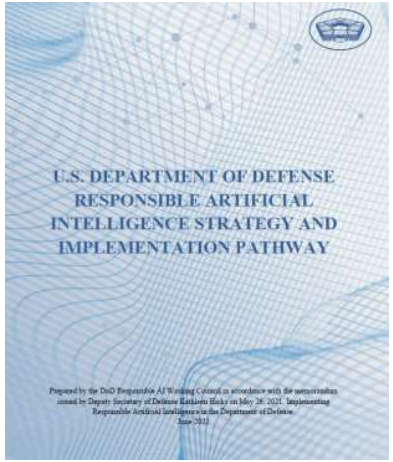
- Mechanism to track AI expertise and determine gaps
  - Career fields and pathways
  - Update standardized curricula and reshape
- AI Ethics Awareness training (similar to Cyber Awareness)
- RAI Champions Program Department-wide

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

SEI is a partner with the DoD on the RAI journey to trust

DoD Ethical Principles

- Responsible
- Equitable
- Traceable
- Reliable
- Governable



Responsible AI

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).
DoD Memorandum, "Artificial Intelligence Ethical Principles for the Department of Defense." (Feb 2020)

AI Workforce Development

SEI, AI Workforce Development https://www.sei.cmu.edu/our-work/projects/display.cfm?customel_datapageid_4050=343975

Responsible AI

SEI, Responsible AI: https://sei.cmu.edu/research-capabilities/all-work/display.cfm?customel_datapageid_4050=197910

Appendix

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

DoD Ethical Principles for Artificial Intelligence

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.

DoD RAI Strategy and Implementation Pathway: **Foundational Tenet Descriptions and Goals**

RAI Governance

Description: Ensure disciplined governance structure and processes at the Component and DoD-wide levels for oversight and accountability and clearly articulate DoD guidelines and policies on RAI and associated incentives to accelerate adoption of RAI within the DOD.

Goal: Modernize governance structures and processes that allow for continuous oversight of DoD use of AI, taking into account the context in which the technology will be used.

Governance structures and processes will enable the appropriate assessment of risks and the mitigation of unintended consequences or bias in AI capabilities. Users or developers will also have clear mechanisms to implement the DoD AI Ethical Principles and to report potential concerns.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

Warfighter Trust

Description: Ensure warfighter trust by providing education and training, establishing a test and evaluation and verification and validation (TEVV) framework that integrates real-time monitoring, algorithm confidence metrics, and user feedback to ensure trusted and trustworthy AI capabilities.

Goal: Achieve a standard level of technological familiarity and proficiency for system operators to achieve justified confidence in AI capabilities and AI-enabled systems. Trustworthiness is bolstered by the application of TEVV frameworks that allow for the monitoring of system performance, reliability, unintended behavior, and failure modes before fielding the system and during operation. The combination of these factors contributes to a greater understanding of an AI's capabilities and limitations, which will be critical for the development of an AI-ready force.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

AI Product and Acquisition Lifecycle

Description: Develop tools, policies, processes, systems, and guidance to synchronize enterprise RAI implementation for the AI product throughout the acquisition lifecycle through a systems engineering and risk management approach.

Goal: Exercise appropriate care in the AI product and acquisition lifecycle to ensure potential AI risks are considered from the outset of an AI project, and efforts are taken to mitigate or ameliorate such risks and reduce the likelihood of unintended consequences while enabling AI development at the pace the Department needs to meet the National Defense Strategy. This includes robust documentation to understand, test, and act on informed risk assessments, recognizing that needs will vary based on the level of technical maturity, sensitivity, and context in which the AI capability will be used.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

Requirements Validation

Description: Incorporate RAI into all applicable AI requirements, including joint performance requirements established and approved by the Joint Requirements Oversight Council, to ensure RAI inclusion in appropriate DoD AI capabilities.

Goal: Use the requirements validation process to ensure that capabilities that leverage AI are aligned with operational needs while addressing relevant AI risks. System performance requirements validation increases the reliability and safety of systems prior to and during deployment. A formalized requirements validation process also provides for better traceability, accountability, and both internal and external oversight.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

Responsible AI Ecosystem

Description: Build a robust national and global RAI ecosystem to improve intergovernmental, academic, industry, and stakeholder collaboration, including cooperation with allies and coalition partners, and to advance global norms grounded in shared values.

Goal: Promote a shared understanding of responsible AI design, development, deployment, and use through domestic and international engagements. Such engagements will facilitate knowledge-sharing exchanges with intergovernmental stakeholders as well as partners in industry, academic institutions, and civil society. Through this, the DoD will collaborate on common challenges, advance shared interests, promote democratic norms and values, and increase interoperability with partners.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

AI Workforce

Description: Build, train, equip, and retain an RAI-ready workforce to ensure robust talent planning, recruitment, and capacity-building measures, including workforce education and training on RAI.

Goal: Ensure that all DoD AI workforce members possess an appropriate understanding of the technology, its development process, and the operational methods applicable to implementing RAI commensurate with their duties within the archetype roles outlined in the 2020 DoD AI Education Strategy. DoD AI workforce education and training should promote consistent understanding across all DoD stakeholders and build a culture within the DoD that enables RAI. Proper training and education must be accompanied with strategies to recruit and retain the personnel whom the DoD trains and educates.

Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).

Supporting Materials

Training - incremental improvement to the model

A model is trained over many iterations to achieve performance goals

- Data adjusted based on performance
  - Variations: amount, breadth, depth, augmentation
  - Annotation: labeling, rules, etc.
- Monitored for performance goals
  - When performance goals are reached model is ready to be integrated into software-based system (new, or existing)

Know *only* what taught. Control *only* what given control of.

Design to work with, and for, people

Effective implementations

Minimize unintended consequences

1. Understand complexity of context
2. Design for human-machine teaming
3. Engage in critical oversight

Human-Centered AI, White Paper. June 2021. CMU's Software Engineering Institute.
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

Ongoing: Speculative Workshops

- Inspire creativity and curiosity regarding misuse and abuse
- Based on HCI methods
- Expose concerns/fears of those using/affected by AI system.