



C D A O

Chief Digital & Artificial
Intelligence Office

DAGR, SHIELD, and the DoD Responsible AI Toolkit

Dr Matthew Johnson
Acting Chief of Responsible AI
Chief Architect, DoD RAI Toolkit
US Department of Defense

CDR Michael Hanna
Director of Global Fleet Operations
Office of Naval Intelligence,
Hopper Global Communications Center

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


What does it mean to ‘win because of our values?’

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
08/11/2023 06:22 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.

Value Proposition of RAI: Assurance

RAI increases assurance, thereby sustaining our tactical edge:

- **Assurance for the Warfighter and Operational Commanders 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 DoD will responsibly steward their 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:**
 - Systems, tools, & processes grounded in shared values.
 - Crucial, given the increasing need for interoperability (e.g., CJADC2, Integrated Deterrence).

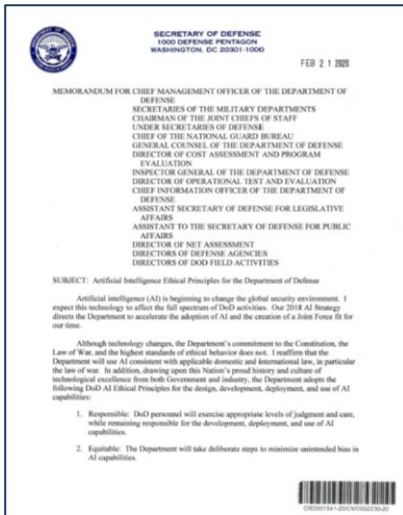


What is Responsible AI (RAI)?

- RAI translates high-level values and the DoD 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
 - There are risks to not innovating fast enough or failing to keep pace with near-peer adversaries
- RAI contributes to mission and military success through justified confidence and decision advantage



DoD AI Ethical Principles



February 2020: AI Ethical Principles Memo

The DoD formally adopts five AI ethical principles and designates the JAIC (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

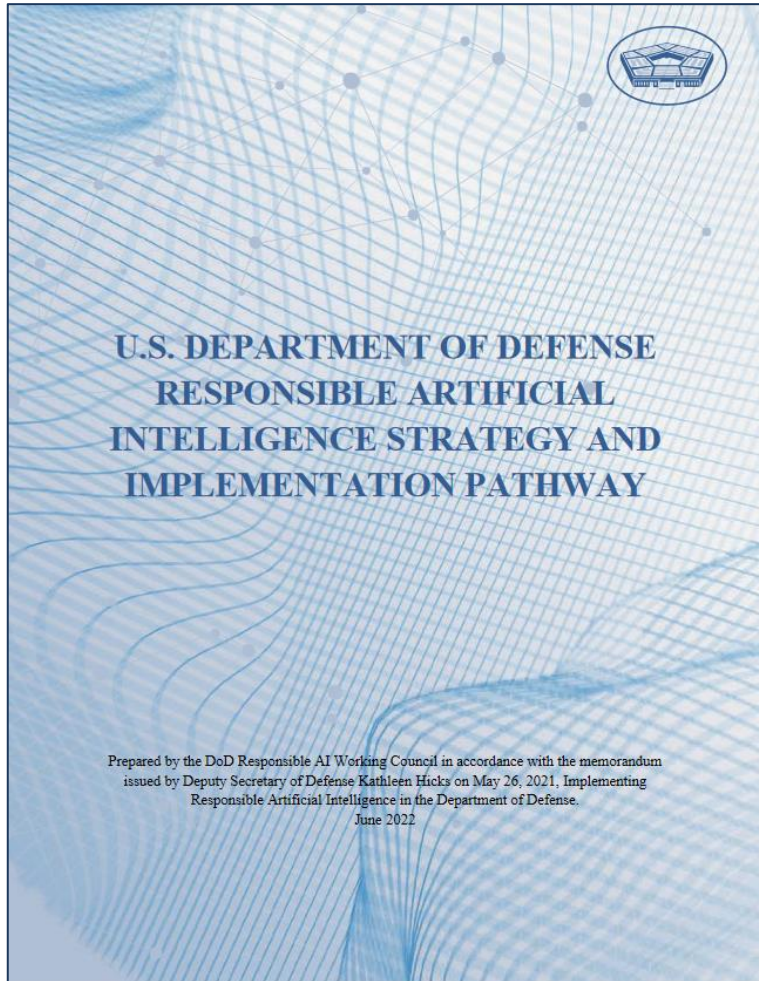
The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to **testing and assurance** within those defined uses across their entire life-cycles.

Governable

The Department will design and engineer AI capabilities to fulfill 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.



RAI Strategy and Implementation Pathway



Outlines the Department's Strategy for Operationalizing the Ethical Principles

Within this document, the Deputy Secretary of Defense:

- Explains the Department's approach to Responsible AI
- Establishes over 60 lines of effort aligned with the RAI implementation tenets
- Defines governance, roles, and responsibilities within the Department
- Directs Department to build RAI Tools and Capabilities

June 2022

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 to 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



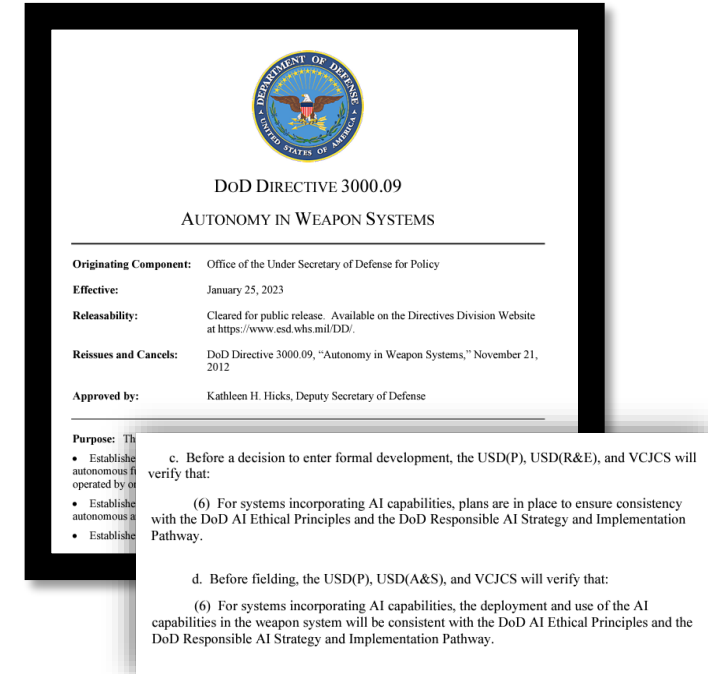
RAI Toolkit

- **The Responsible AI Toolkit is our organizing framework to make the capabilities being built out under the RAI Strategy & Implementation Pathway:**
 - Findable
 - Usable
 - Interoperable
- **Centralized process through which AI projects can identify, track, and improve alignment with RAI practices and the DoD AI Ethical Principles, and manage risk while capitalizing on opportunities for innovation.**
- **Living document and Web App (currently in MVP form) building upon and incorporating:**
 - Industry best practices and tools (currently 70+ listed in the Toolkit) and Academic innovations
 - DIU RAI Guidelines & Worksheets, NIST AI RMF + Playbook, IEEE 7000, etc.
 - Tools being built through the RAI Strategy & Implementation Pathway



RAI Toolkit Priorities

- **Provide a process for demonstrating consistency/alignment with the DoD AI Ethical Principles**
 - *viz.* 3000.09's pre-fielding and pre-development SRB requirements
 - Use 3000.09 as a pathfinder/validation for having this 'consistency'/alignment requirement in additional policy
- **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 (e.g. CJADC2)*
 - Developed version of Toolkit for NATO
 - Developing collaborations over the Toolkit with IC, Interagency, ROK

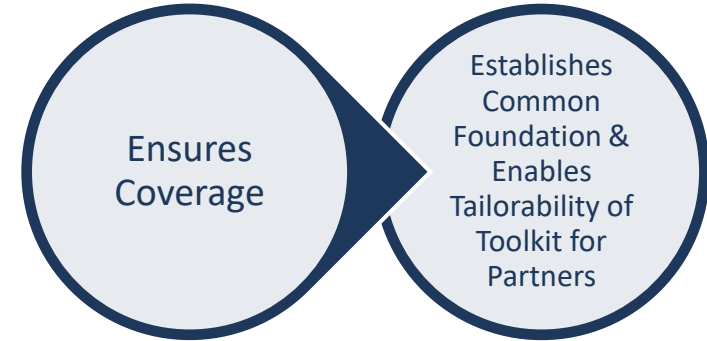


Approach to Toolkit



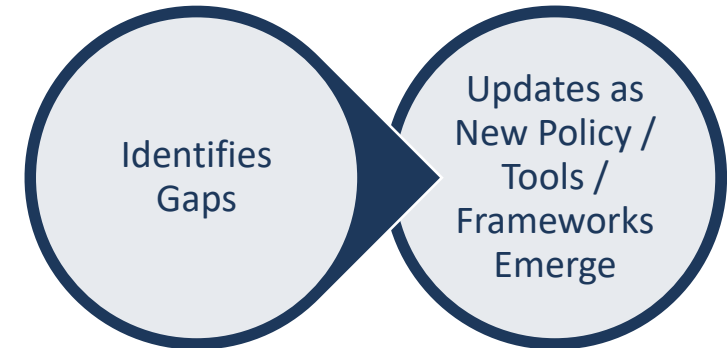
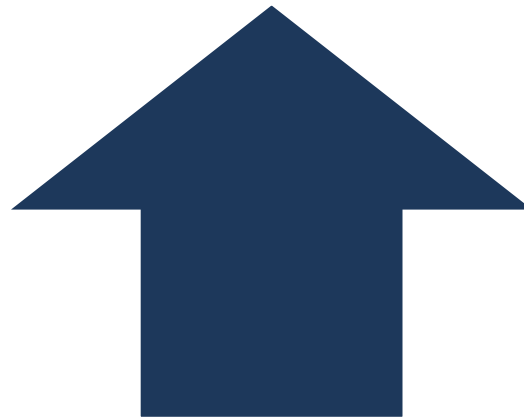
Top-Down Approach:

Identified 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:

Drew 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.)



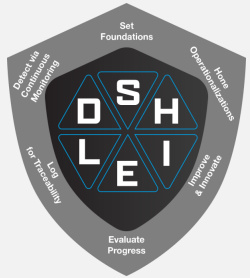
Design Challenges and Principles

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 process 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)

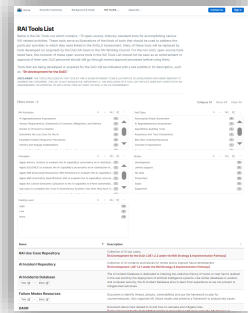
RAI Toolkit Components

Currently Available in Toolkit MVP



Planning Tools

Identify and document potential risks and plan RAI activities for mitigation



Tools and Resource Database

Provides resources for implementing RAI plan



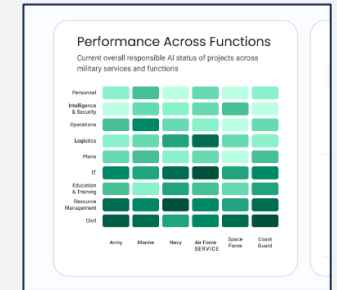
Software Tools, Guidance and Best Practices, Checklists, Metrics

In Development



Evaluation Tools

Evaluate progress against RAI plan

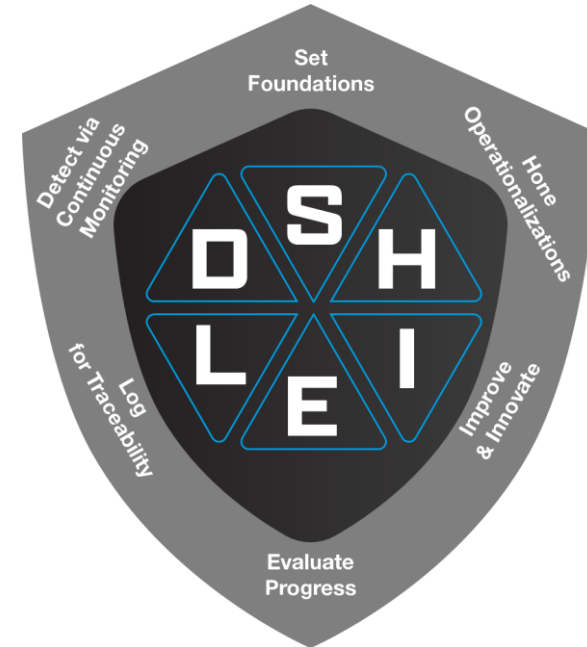
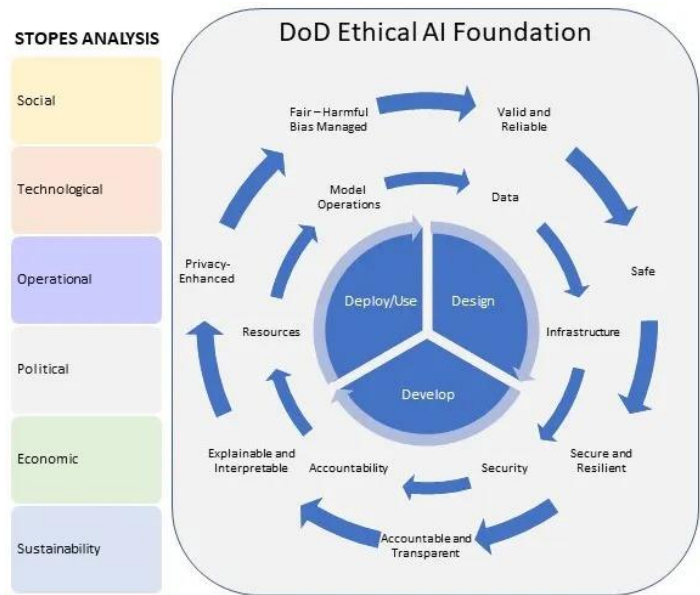


Oversight Dashboard

Monitor RAI progress and risk profile across programs and portfolios

RAI Toolkit assists users to plan and execute RAI activities, and select appropriate supporting artifacts and tools

How: RAI Planning and Assessment



DAGR Risk Assessment

- Risk management guidance for DoD, 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 that identify RAI-related issues for tracking and mitigation
- List of issues are tracked throughout the lifecycle via Statements of Concern (SOCs)
- Elements in the SHIELD Assessment route the user to relevant tools within the Tools Database

Defense AI Guide on Risk (DAGR)

DAGR: Holistic risk guide to mitigate risks and realize opportunities of AI capabilities.

Guiding Principles

Constitutional Rights, Democratic Way of Life, and Shared Values/Interests

Strategic Alignment

Guidance and Frameworks

1. Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI
2. DoD AI Ethical Principles
3. NIST AI RMF
4. International Considerations (Partners and Allies)
5. Best Practices (Industry and Academia)

Consolidation, Abstraction, and Systems Thinking

DAGR Components

1. AI Risk Relationship Dynamics – Shifting, bidirectional, and interconnected.
2. STOPES – Social, Technological, Operational, Political, Economic, and Sustainability.
3. DISARM Hierarchy of AI Risk – Data, Infrastructure, Security, Accountability, Resources, and Model Operations.
4. Risk Considerations – Facilitate effective and collaborative risk discussion.
5. Risk Evaluation Process – Qualitative/Quantitative evaluation of risk.

STOPES Analysis

Social

Technological

Operational

Political

Economic

Sustainability

AI Risk Evaluation Process

DAGR: Holistic risk guide to mitigate risks and realize opportunities of AI capabilities.

- Risk as a function of Likelihood and Consequence
- Risk in interrelated AI systems is a function of their relative dependency

Probability of Event	Consequence of Event			
	Minor	Modest	Major	Extreme
Almost Certain / Nearly Certain (95-99.9%)	7	14	21	28
Very Likely / Highly Probable (80-95%)	6	12	18	24
Likely / Probable (55-80%)	5	10	15	20
Roughly Even Chance / Roughly Even Odds (45-55%)	4	8	12	16
Unlikely / Improbable (20-45%)	3	6	9	12
Very Unlikely / Highly Improbable (5-20%)	2	4	6	8
Almost No Chance / Remote (0.1-5%)	1	2	3	4



CAPABILITY 1		
Social Factors	Risk Score	Residual Risk Score
Social Risk 1	6	3
Social Risk 2	12	4
Social Factor Sum	18	7
Technological Factors	Risk Score	Residual Risk Score
Technological Risk 1	1	1
Technological Risk 2	3	2
Technological Risk 3	3	2
Technological Factor Sum	7	5
Operational Factors	Risk Score	Residual Risk Score
Operational Risk 1	6	6
Operational Factors Sum	6	6
Political Factors	Risk Score	Residual Risk Score
No Political Risks	0	0
Political Factors Sum	0	0
Economic Factors	Risk Score	Residual Risk Score
Economic Risk 1	2	1
Economic Factors Sum	2	1
Sustainability Factors	Risk Score	Residual Risk Score
Sustainability Risk 1	9	6
Sustainability Risk 2	3	3
Sustainability Factor Sum	12	9

DAGR Framework

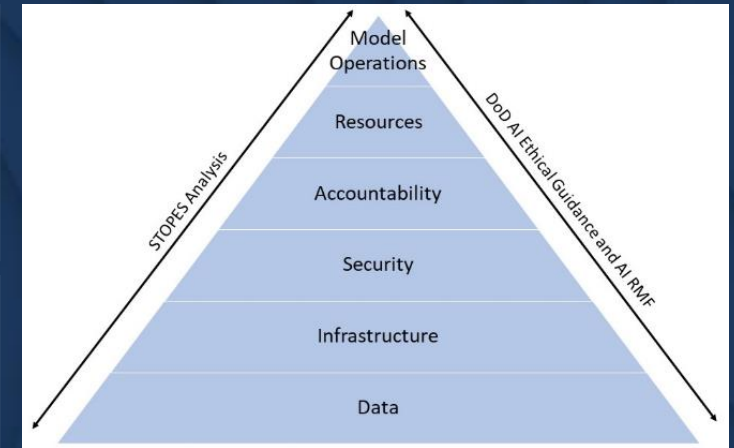
DAGR: Holistic risk guide to mitigate risks and realize opportunities of AI capabilities.

DAGR Components

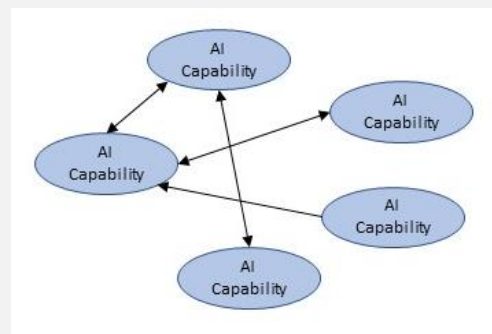
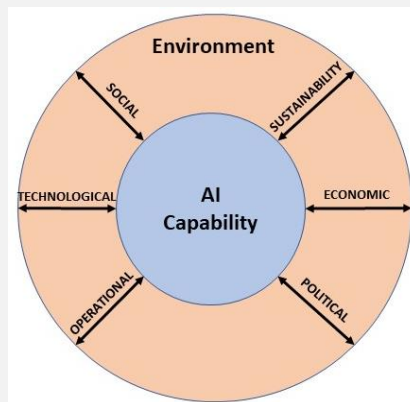
1. **AI Risk Relationship Dynamics** – Shifting, bidirectional, and interconnected.
2. **STOPES** – Social, Technological, Operational, Political, Economic, and Sustainability.
3. **DISARM Hierarchy of AI Risk** – Data, Infrastructure, Security, Accountability, Resources, and Model Operations.
4. **Risk Considerations** – Facilitate effective and collaborative risk discussion.
5. **Risk Evaluation Process** – Qualitative/Quantitative evaluation of risk.

Benefits

1. **Diverse Benefits** – Promote trustworthiness, responsibility, risk mitigation, operations, and realize opportunities.
2. **Alignment**
3. **Consolidation and Abstraction**

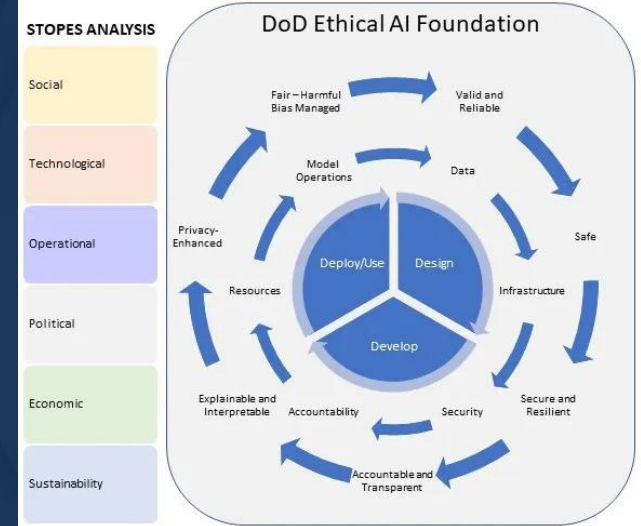


AI Risk Relationship Dynamics (Shifting, Bidirectional, and Interconnected)



STOPES Analysis

- Social
- Technological
- Operational
- Political
- Economic
- Sustainability



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

The screenshot displays the 'RAI Tools List' web application. The page has a navigation bar with links for Home, Executive Summary, Background & Guide, RAI Toolkit, and Appendix, along with 'Contact Us' and 'Sign In' buttons. The main content area is titled 'RAI Tools List' and includes a disclaimer: '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 CDAG, the DOD, or the US GOVERNMENT.' Below the disclaimer, there are filter sections for 'RAI Activities', 'Principles', 'Coding Level', 'Tool Class', and 'Status'. Each filter section contains a list of items with a search icon, a filter icon, and a count. For example, 'RAI Activities' includes 'AI Appropriateness Assessment' (2), 'Assess Requirements, Statements of Concern, Mitigations, and Metrics' (2), 'Decide to Proceed to Ideation' (2), 'Determine the Use Case for the AI' (5), 'Establish Incident Response Procedures' (5), and 'Identify and Engage Stakeholders' (5). The 'Tool Class' filter includes 'Adversarial Attack Generation' (2), 'AI Appropriateness Assessment' (5), 'Algorithmic Auditing Tools' (5), 'Assurance and Trust Instruments' (5), 'Bias Red-Teaming Resources' (1), and 'Concept of Operations' (1). The 'Status' filter includes 'Development' (10), 'Limited support' (5), 'No data' (14), 'Production' (26), 'Static' (12), and 'Supported' (1). At the bottom, there is a table with columns for 'Name' and 'Description'. The table lists several resources: '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.), and 'DAGR' (Document about risks related to AI and how to calculate and mitigate risks. [In Development for the DoD; ML/D Available in Appendix 1, LOE 3.1.3 under the RAI Strategy & Implementation Pathway]).



Who: RASCI and Personas List

RAI Role**
Users/Stakeholders
Mission Commanders
Senior Leader / AI Innovation Leader
Functional Requirements Owner
Program Manager
AI Ethics & 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 & Evaluation Specialist
IT / Cyber Expert

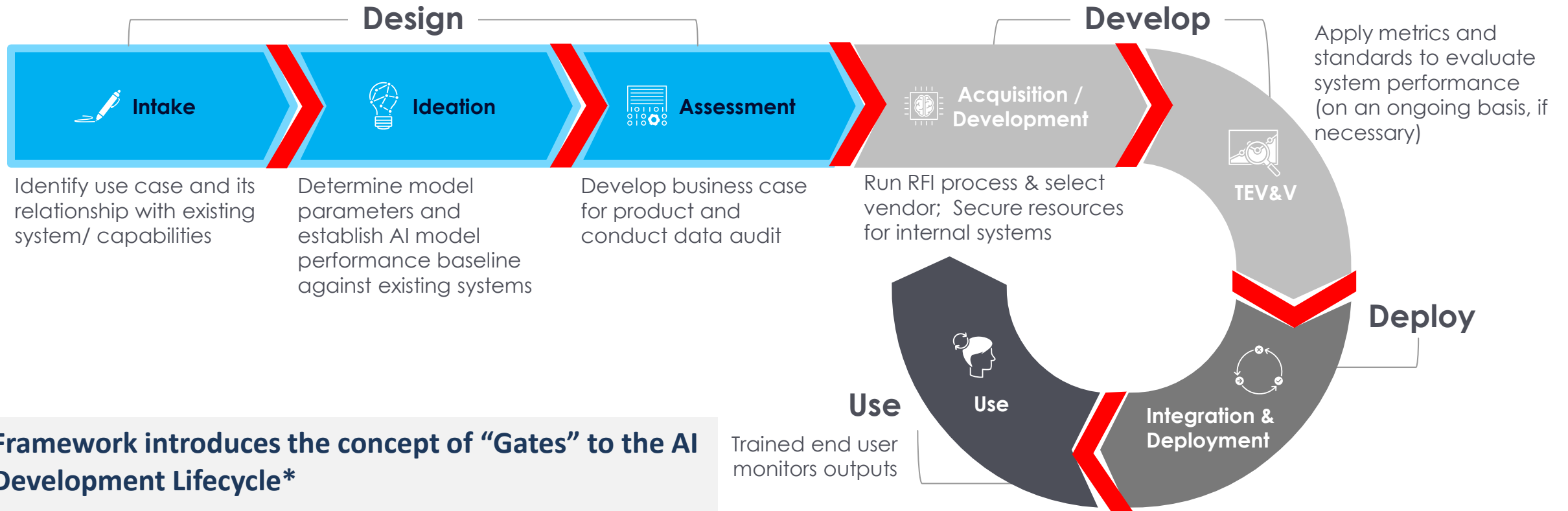
RASCI matrix is built for each role to clarify RAI taking

Role*	Definition
<u>R</u>esponsible	The person who does the work to complete the task or create the deliverable
<u>A</u>ccountable	The person ultimately accountable for the work or decision being made; this person gives final approval.
<u>S</u>upporting	Support for those who are responsible or accountable; participates in doing the work of a task
<u>C</u>onsulted	Anyone who must be consulted with or add input prior to a decision being made and/or the task being completed
<u>I</u>nformed	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 Lifecycle



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)

RAI Toolkit Current Features

Home Executive Summary Background & Guide RAI Toolkit Appendix Contact Us Sign in

Overview of RAI Activities Throughout the Product Life Cycle

SHIELD

Intake Ideation Assessment Development/Acquisition TEVV Integration & Deployment Use

SHIELD Navigation

Export Import

Clear Responses

View PDF

1. Intake

1.1 SET: Consider Previously I

1.2 SET: Determine Relevant I

1.3 SET: Identify and Engage

1.4 SET: Competitize the Use C

1.5 SET: Decide to Proceed to

2. Ideation

2.1 HONE: Define Requiremer

2.2 HONE: Identify Risks & Oj

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: Instr

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?

XAI Toolkit - Saliency Python Outlier Detection (PyOD)

Response...

8. Have you established a cadence and procedure through which new data will be retrained, and the system will be updated?

Response...

9. Revisit 2.4. How are you designing your system or leveraging capabilities to reduce the ethical/risk burden on operational users, decision makers, and impacted stakeholders that would otherwise be present in the system?

Response...

10. 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...

4.2 Update Documentation

Filters & RASCI

Update SOCs and data/model cards, as necessary. Have team consult and update DAGR to support

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

Export/Import function to save and share progress

Navigation by AI Product Lifecycle Stage

Export as PDF

Navigation by Type of RAI Activity

"GATE" filter (displays most essential assessment questions)

SHIELD Assessment identifies risk and opportunities

Filters assessment questions by Persona/Project Role, AI Ethical Principle, Discipline, etc.

Links to tools to address identified risks and opportunities



RAI Toolkit Web App



RAI Toolkit Web App:

<https://rai.tradewindai.com/>



Toolkit Way Ahead

- **Develop version of Toolkit to support 3000.09 Reviews**
 - Deconflict with other required processes and documentation to support creation of integrated template or documentation process
 - ‘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**
 - Collect, organize, and provide lessons learned as inputs to policy
- **Develop Acquisitions-focused version of Toolkit**
- **Integrate into DCWF Courses**
- **Continue to add functionality**
 - Develop UI
 - Add further tailorability features (data & model type, use case, risk profile, etc.)
 - Continue Dashboard development
 - Integrate with other tools (T&E, Cyber) and with DoD Platforms (ADVANA)
 - Integrate feedback
 - Create high-side versions



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 & Questions

Dr Matthew Kuan Johnson

Chief of Responsible AI (Acting)

US Department of Defense

CDR Michael Hanna

Director of Global Fleet Operations

Office of Naval Intelligence,

Hopper Global Communications
Center

Contact the RAI Team:

osd.pentagon.cdao.mbx.dod-rai-toolkit@mail.mil



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



Extra



Background on the CDAO RAI Team

RAI at DoD

- The DoD defines 'Responsible AI' (RAI) as a dynamic approach to the design, development, deployment, and use of artificial intelligence systems which implements the DoD AI Ethical Principles to advance the trustworthiness of such systems.
- RAI at DoD emphasizes technical maturity, organizational change, modernized governance structures, and an understanding of socio-technical risk.

CDAO RAI Team Role

- Primary technical advisor to the DoD on RAI
- Oversees execution of the RAI Strategy & 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 DoD