“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?’
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

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsible</td>
<td>DoD personnel will exercise <strong>appropriate levels of judgment and care</strong>, while remaining responsible for the development, deployment, and use of AI capabilities.</td>
</tr>
<tr>
<td>Equitable</td>
<td>The Department will take deliberate steps to <strong>minimize unintended bias</strong> in AI capabilities.</td>
</tr>
<tr>
<td>Traceable</td>
<td>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 <strong>transparent and auditable methodologies, data sources, and design procedure and documentation</strong>.</td>
</tr>
<tr>
<td>Reliable</td>
<td>The Department’s AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to <strong>testing and assurance</strong> within those defined uses across their entire life-cycles.</td>
</tr>
<tr>
<td>Governable</td>
<td>The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to <strong>detect and avoid unintended consequences</strong>, and the ability to <strong>disengage or deactivate deployed systems</strong> that demonstrate unintended behavior.</td>
</tr>
</tbody>
</table>
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
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.

<table>
<thead>
<tr>
<th>What</th>
<th>Function</th>
<th>Example Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical or Software-Based</td>
<td>Helps developers and testers to assess factors such as bias, reliability, and safety</td>
<td>Data Bias Detection Tools</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Explainability Tools</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T&amp;E Harness</td>
</tr>
<tr>
<td>Documentation and Artifacts</td>
<td>Provides traceability of data sources, model limitations, risk identification and mitigation efforts</td>
<td>Use Case/Harms Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data Cards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model Cards</td>
</tr>
<tr>
<td>Frameworks and Checklists</td>
<td>Provides prompts to guide users in creating muscle memory around new processes for risk assessment and ethical considerations</td>
<td>Common Failure/Mishap List</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Algorithmic Impact Assessments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ethics Maturity Assessments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Research and Design Tools</td>
</tr>
<tr>
<td>Knowledge Sharing</td>
<td>Provides centralization for information sharing, learning, and common lexicon, practices, etc.</td>
<td>Use Case Repositories</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information Management Systems</td>
</tr>
<tr>
<td>Executive Dashboards</td>
<td>Provides visibility into organizational compliance, status, and risk</td>
<td>Key Performance Metrics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Progress Tracking</td>
</tr>
</tbody>
</table>
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 and 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.)

---

- Ensures Coverage
- Establishes Common Foundation & Enables Tailorability of Toolkit for Partners
- Identifies Gaps
- Updates as New Policy / Tools / Frameworks Emerge
- Ensures Coverage
# 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.

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wide diversity of use cases and priorities across the DoD</td>
<td>Modular and Tailorable</td>
</tr>
<tr>
<td>Demonstrate alignment with DoD AI Ethical principles</td>
<td>Traceable</td>
</tr>
<tr>
<td>Existing assessment processes can overwhelm a small team</td>
<td>Lightweight</td>
</tr>
<tr>
<td>RAI process require coordination among diverse team roles and stakeholder</td>
<td>Holistic</td>
</tr>
<tr>
<td>considerations</td>
<td>Integrated</td>
</tr>
<tr>
<td>RAI activities should take place during all phases of AI development</td>
<td>Upskilling</td>
</tr>
<tr>
<td>Existing approaches assume expert RAI knowledge</td>
<td>Iterative (Living Document)</td>
</tr>
<tr>
<td>RAI research and practice is still evolving</td>
<td></td>
</tr>
</tbody>
</table>
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

Evaluation Tools
Evaluate progress against RAI plan

In Development

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.
  4. Risk Considerations – Facilitate effective and collaborative risk discussion.

## 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
# 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.
4. **Risk Considerations** – Facilitate effective and collaborative risk discussion.

## Benefits

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

---

**AI Risk Relationship Dynamics**

- **Environment**
- **Operational**
- **Economic**
- **Political**
- **Technological**
- **Social**
- **Sustainability**

**STOPES Analysis**

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

**DoD Ethical AI Foundation**

- Human Rights
- Fairness
- Accessibility
- Transparency
- Accountability
- Security
- Privacy
- Transparency
- Usability
- Reliability
- Robustness
- Safety
Tools and Resource Database MVP

• Searchable database (70+ items) of COTS/GOTS/open-source RAI tools:
  o 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:
  o Tailorable labels for ethical principles, development lifecycle phases, category names, roles, and disciplines
  o Interactive search and exploration
### RAI Role**

<table>
<thead>
<tr>
<th>Role **</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Responsible</strong></td>
<td>The person who does the work to complete the task or create the deliverable</td>
</tr>
<tr>
<td><strong>Accountable</strong></td>
<td>The person ultimately accountable for the work or decision being made; this person gives final approval.</td>
</tr>
<tr>
<td><strong>Supporting</strong></td>
<td>Support for those who are responsible or accountable; participates in doing the work of a task</td>
</tr>
<tr>
<td><strong>Consulted</strong></td>
<td>Anyone who must be consulted with or add input prior to a decision being made and/or the task being completed</td>
</tr>
<tr>
<td><strong>Informed</strong></td>
<td>The people who need to be updated on project status, or informed when a decision is made or work completed</td>
</tr>
</tbody>
</table>

---

**Individuals or Teams may be dual-hatted; Roles map to DoD Cyber Workforce (DCWF) roles – BLUE text indicates relevant DCWF Role**

<table>
<thead>
<tr>
<th>Role **</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users/Stakeholders</td>
<td></td>
</tr>
<tr>
<td>Mission Commanders</td>
<td></td>
</tr>
<tr>
<td>Senior Leader / AI Innovation Leader</td>
<td></td>
</tr>
<tr>
<td>Functional Requirements Owner</td>
<td></td>
</tr>
<tr>
<td>Program Manager</td>
<td></td>
</tr>
<tr>
<td>AI Ethics &amp; Risk Specialist</td>
<td></td>
</tr>
<tr>
<td>Relevant Legal, Ethical, or Policy Expert</td>
<td></td>
</tr>
<tr>
<td>UX/Design/HMT / AI Adoption Specialist</td>
<td></td>
</tr>
<tr>
<td>AI Development Team</td>
<td>System Architect</td>
</tr>
<tr>
<td></td>
<td>Data Architect</td>
</tr>
<tr>
<td></td>
<td>Data Operations Specialist</td>
</tr>
<tr>
<td></td>
<td>Data Analyst</td>
</tr>
<tr>
<td></td>
<td>Data Scientist</td>
</tr>
<tr>
<td></td>
<td>Data Officer</td>
</tr>
<tr>
<td></td>
<td>AI Engineer / AI/ML Specialist</td>
</tr>
<tr>
<td></td>
<td>Data Steward</td>
</tr>
<tr>
<td>AI Test &amp; Evaluation Specialist</td>
<td></td>
</tr>
<tr>
<td>IT / Cyber Expert</td>
<td></td>
</tr>
</tbody>
</table>

---

Who: RASCI and Personas List
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

- Export/Import function to save and share progress
- Export as PDF
- Navigation by Type of RAI Activity
- SHIELD Assessment identifies risk and opportunities
- Links to tools to address identified risks and opportunities

Navigation by AI Product Lifecycle Stage

“GATE” filter (displays most essential assessment questions)

Filters assessment questions by Persona/Project Role, AI Ethical Principle, Discipline, etc.
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
“...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](mailto:osd.pentagon.cdao.mbx.dod-rai-toolkit@mail.mil)

RAI Toolkit Web App:  
[https://rai.tradewindai.com/](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