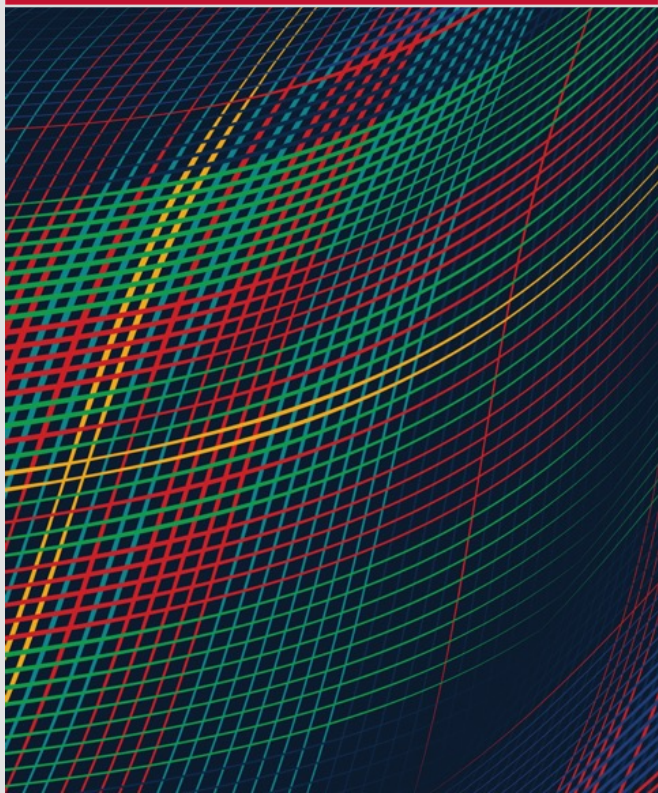
Responsible AI

11/30/2023

AI Division

Carnegie Mellon University
Software Engineering Institute



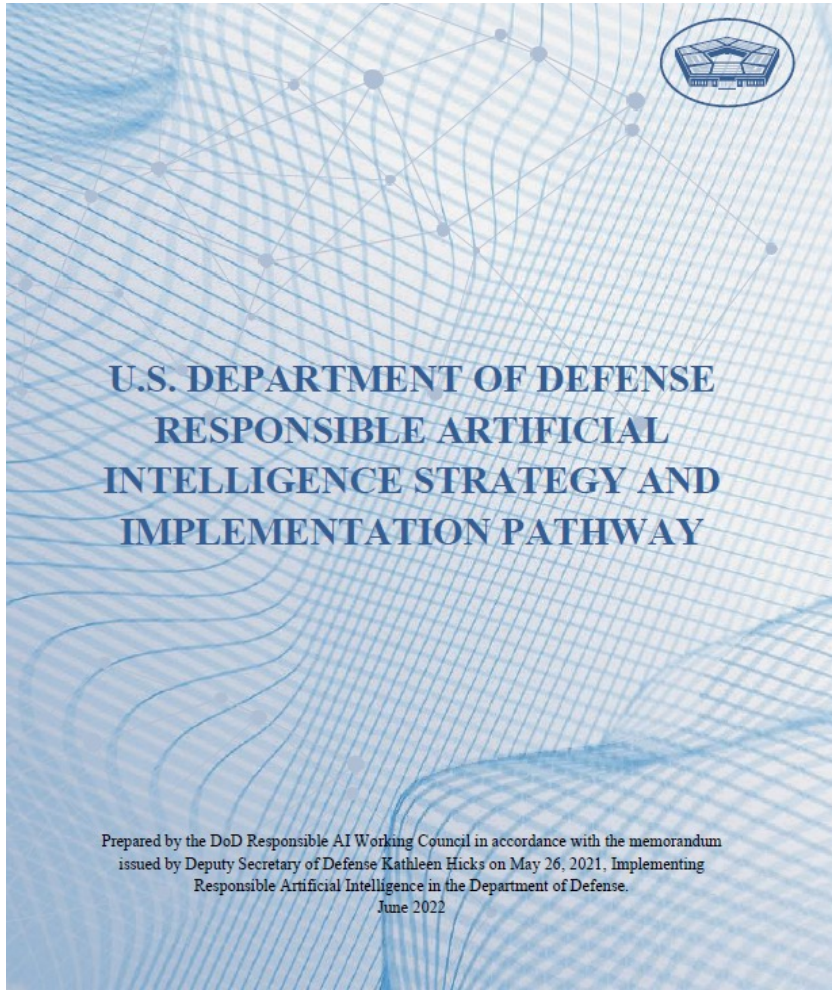
Document Markings

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-2260

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.

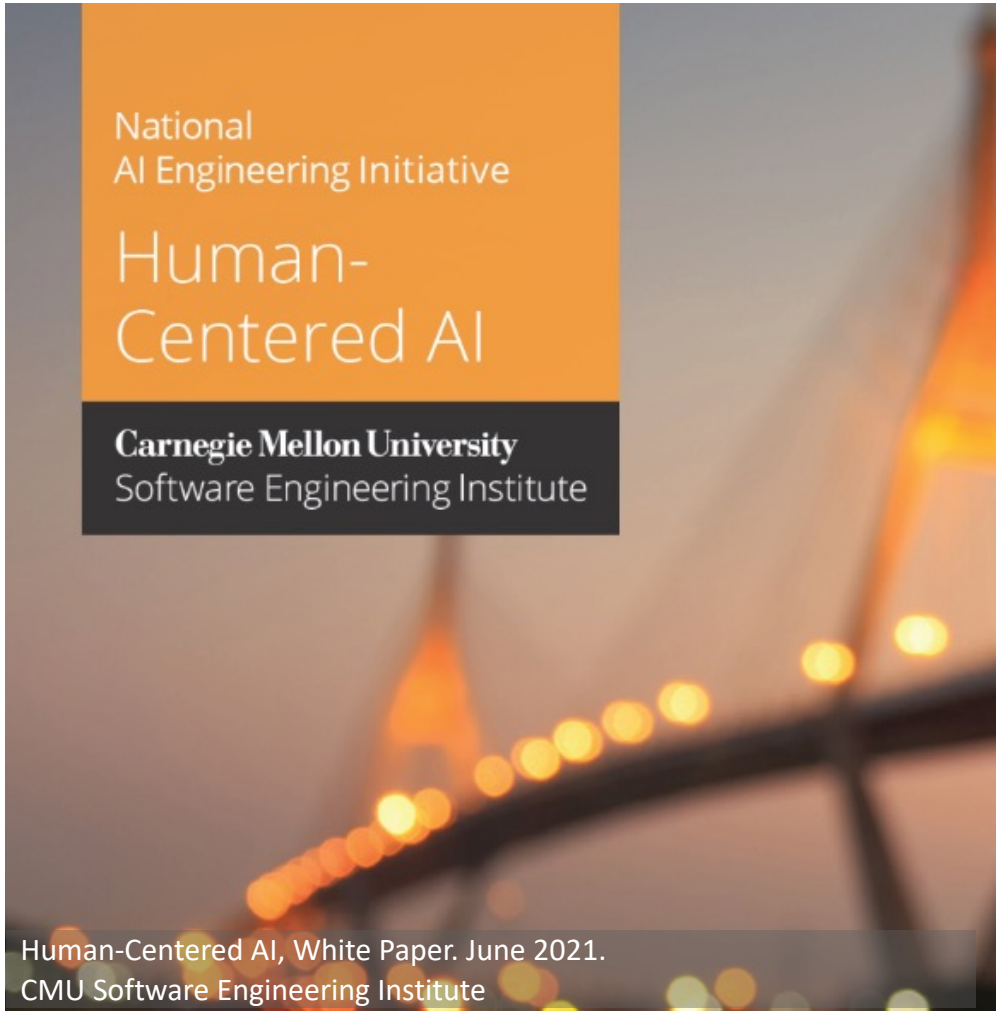


Responsible Artificial Intelligence (RAI) Strategy and Implementation Pathway (June 2022).
PDF: https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

RAI: Design to work with, and for, people

- Provide trustworthy interactions.

- We must design AI systems to:
  - be accountable to humans
  - identify and explain risks
  - be respectful, honest, and usable



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute

RAI Principles

- **Responsible** - Exercise appropriate judgment and care, while remaining responsible for development, deployment, and use of AI capabilities.
- **Equitable** – Minimize unintended and/or bias in AI capabilities.
- **Traceable** – Develop so relevant personnel posses and appropriate understanding of the technology.
- **Reliable** – Capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness will be subject to testing and assurance.
- **Governable** – design to fulfill intended functions while detecting and avoiding unintended consequences.

Defense AI Guide on Risk (DAGR)

Intended to provide DoD AI stakeholders with guiding principles, best practices, and other governing Federal and DoD guidance.

https://rai.tradewindai.com/appendix/dagr

A STOPES analysis examines the Social, Technological, Operational, Political, Economic, and Sustainability (STOPES) factors.

Site Demo Scenario - Session 1

You are bidding on an RFP to build an AI system for Job candidate identification, review, and screening system

Site Demo Scenario – Session 2

You are bidding on an RFP to build an AI system for Automatic recognition of hostile forces in active warzones.

Site Demo Scenario – Session 3

You are bidding on an RFP to build an AI system for Health insurance companies to identify risk factors when evaluating rates for customers.

Site Demo Scenario – Session 4

You are bidding on an RFP to build an AI system for automated responses to questions related to veterans benefits to assist warfighters.

Discussion (15 min)

Your group will discuss potential issues related to your AI system in one of the following topics (assigned by the moderators):

- Societal
- Political
- Economical
- Sustainability.

Consider Technical and Operational issues as well as time permits.

Additional Resources

CDAO has requested that you take a look at the RAI Toolkit site to make sure that you've accounted for any inherent biases, risks, or harms and that your budget / plan incorporate Responsible AI needs.

CDAO has specifically pointed out Section 4 of the SHIELD Assessment as sections of interest for you. Please walk through this sections to understand their impact.

https://rai.tradewindai.com/shield/development-acquisition

Feedback & Questions

- What are your first impressions of the SHIELD Assessment?
- What sticks out as MOST helpful from this site?
- What would prevent you from using a site like this?
- What types of things would you like to see incorporated into the RAI toolkit?
- What capabilities would you want to be developed for the toolkit?

Interested in Helping?

We are recruiting individuals that are willing to discuss their experiences building AI system. These interviews will help us identify effective strategies for improving AI system development and to help teams reduce risk throughout the process.

Please talk with your moderator (Katie) after the session if you are interested in more details.

Thank you

Carnegie Mellon University
Software Engineering Institute