CDAO

AI Assurance Toolkit Executive Summary

By: *Bryan McFadden*

What is the RAI Toolkit?

The AI Assurance Toolkit provides a centralized process that identifies, tracks, and improves the alignment of AI projects toward RAI best practices and the DoW AI Ethical Principles, while capitalizing on opportunities for innovation. The AI Assurance Toolkit provides an intuitive flow guiding the user through tailorable and modular assessments, tools, and artifacts throughout the AI product lifecycle. Using the Toolkit enables people to incorporate traceability and assurance concepts throughout their development cycle.

In 2020, the DoW officially adopted a series of ethical principles for the design, development, and deployment of AI-enabled capabilities. Two years later, the Deputy Secretary of War signed the **RAI Strategy and Implementation Pathway**, which outlined 64 lines of effort across the DoW for building the capability to support alignment with these Principles. The AI Assurance Toolkit is the cornerstone of the AI Assurance Strategy & Implementation Pathway.

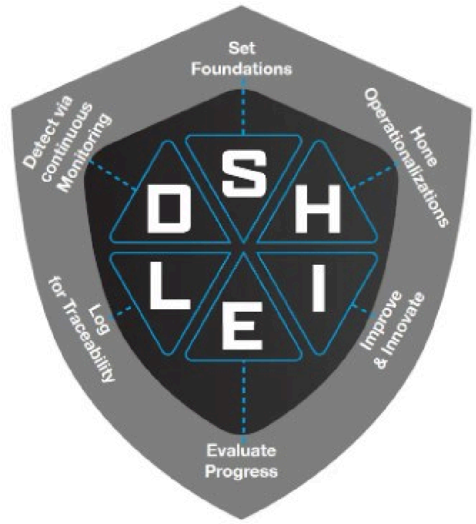
The RAI Toolkit is built upon the earlier **Responsible AI (RAI) Guidelines and Worksheets** developed by the Defense Innovation Unit (DIU), the **NIST AI Risk Management Framework and Toolkit**, and **IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design**, among other sources. This version of the RAI Toolkit is a Minimum Viable Product (MVP) and CDAO will continue to improve this Toolkit as the field evolves.

The CDAO developed the AI Assurance Toolkit to support the DoW community and it has five principals as its focus:

1. **Modular and Tailorable:** The AI Assurance Toolkit is built to be customized because every project differs in use case, context, or priorities. Toolkit users can tailor the content they see by tagging relevant labels. This means that rather than wrestling with irrelevant content, you only see what you need to see.
2. **Aligned to RASCI Matrix:** The AI Assurance Toolkit uses a RASCI (Responsible, Accountable, Supporting, Consulted, and Informed) Matrix. This makes it clear who is best situated to answer each question or execute on each action item and helps to more clearly delineate roles and responsibilities. This results in a more effective overall process, while promoting accountability. This alignment to a RASCI Matrix does not assume program managers (PMs) to have complete knowledge of a program and it does not require PMs to chase down appropriate experts for answers.
3. **Holistic:** The AI Assurance Toolkit provides resources for evaluating and improving AI systems — it's not just about engineering them. The Toolkit attempts to foster reliable development and use. This document starts with technological governance, but future versions will include resources for organizational governance and operational guidance.
4. **DoW AI Ethical Principles:** The AI Assurance Toolkit incorporates the DoW AI Ethical Principles. This allows a user to trace each element back to those Principles.
5. **Tools List:** The AI Assurance Toolkit offers a list of 70+ tools to assist in mitigating risks or improving development of AI systems. For this version of the Toolkit, the majority of the tools are industry-standard, open-source options. Many of these tools will eventually be replaced by DoW-specific tools and the inclusion of these open-source tools in the AI Assurance Tools List should not be seen as an endorsement by the DoW or CDAO. However, these open-source tools are available as a starting point for innovators and DoW personnel should still go through normal software deployment approval processes before using them.

The AI Assurance Toolkit is built around the SHIELD assessment, an acronym for the following six sequential activities:

- **S**et Foundations
- **H**one Operationalizations
- **I**mprove & Innovate
- **E**valuate Status
- **L**og for Traceability
- **D**etect via continuous Monitoring



A SHIELD assessment routes users to the relevant Tools List to conduct an assessment. The essence of SHIELD are short factual statements that identify risks projected in models called Statements of Concern (SOCs). These SOCs provide a starting point to personnel at all stages of the AI lifecycle to develop mitigation plans. They also provide a list of issues that innovators and users can continually reference as their products develop.

Underpinning the SHIELD assessment and the Statements of Concern is a DoW-specific risk identification resource: the Department of War AI Guide on Risk (DAGR). DAGR is a resource for identifying and evaluating sources of risk that will be built out into a DoW-specific AI Risk Management Framework in FY24, in collaboration with the National Institute of Standards and Technology (NIST). DAGR symbolizes how — contrary to a common objection — AI Assurance is not an impediment to warfighter effectiveness, but instead enables warfighter trust in AI capabilities — thereby bringing more capability to the warfighter and support for commander intent to the battlefield. Additionally, responsible AI practices within the DoW serve to reassure the assurance of the American public, industry, academia, allies, and the broader AI community, which will help the DoW sustain our technological and our interoperability with partners. In turn, RAI brings 'more fight' to our tactical edge.

The AI Assurance Toolkit is a living document. It will be continually updated with new capabilities, as they become available. The CDAO's AI Assurance team plans to add data/model/system card templates, RAI project management tools, harms and impact analysis templates, an acquisition toolkit with standardized contract language, and a human-machine teaming red-teaming guidebook in the near future.

We welcome feedback so that we can ensure this resource better tracks the needs and priorities of the DoW. The AI Assurance Toolkit Team can be contacted via **email**.