ACQUISITION INNOVATION RESEARCH CENTER

Digital Transformation in Test and Evaluation for AI/ML, Autonomous, and Evolving Systems – Option Year 1

EXECUTIVE SUMMARY

SEPTEMBER 2024

PRINCIPAL INVESTIGATOR

Dr. Laura Freeman, *Virginia Tech National Security Institute*

CO-PRINCIPAL INVESTIGATOR

Mr. Geoffrey Kerr, *Virginia Tech National Security Institute*

SPONSOR

Mr. Paul Lowe, Deputy Director, Strategic Initiatives, Policy and Emerging Technologies (Acting), Director, Operational Test & Evaluation (DOT&E)
Dr. Jeremy Werner, Chief Scientist, DOT&E
Mr. Nilo Thomas, DOT&E Software and AI Advisor

EXECUTIVE SUMMARY

This report is a culmination of research activities conducted across seven different research organizations in support of the Director, Operational Test and Evaluation's (DOT&E) Strategic Initiatives, Policy, and Emerging Technologies Directorate (SIPET). The methods developed in this research reflect emerging technologies, a changing threat landscape, and the need for efficient and effective test and evaluation (T&E) to ensure capabilities delivered to warfighters work as intended when called upon.

This report builds on the foundational work performed in the base year[1] by the Acquisition Innovation Research Center (AIRC) University Affiliated Research Center (UARC) research in partnership with the DOT&E. The research captures current best practices for improving T&E. The research looked at improving practices through policy, exemplar tools, training material, and the transition of emerging technology into practical application by T&E professionals. The research efforts were conducted as authorized by the execution of the option year contracts for WRT-1070: Test and Evaluation Methods for Middle Tier Acquisition (MTA) and WRT-1071: Digital Transformation in Test and Evaluation. The research team aligned their efforts to the DOT&E Implementation Plan (I-Plan) pillars. Specifically, the team supported the following 3 pillars.

- Pillar 1 – Test the Way We Fight
- Pillar 2 – Accelerate the Delivery of Weapons That Work
- Pillar 4 – Pioneer T&E of Weapon Systems Built to Change Over Time

The team organized their support to these pillars in the following lines of concentration/effort.

- Joint Test Concept (JTC) – Develop methodologies to test systems in order to support the operational assessment of system-of-system joint operations.
- Integrated Testing – Applying statistical methods and leveraging contractor testing, development testing, and operational testing (OT) to more efficiently perform operational assessments.
- Digital Engineering – Applying modern modeling techniques to integrate model-based test planning, modeling and simulation, test execution, model-based systems engineering, and digital product lifecycle management to ensure scientific rigor is ensured and efficient test planning and execution are realized on legacy platform development and born-digital weapon system development.
- Artificial Intelligence/Machine Learning (AI/ML) and T&E – Leverage state-of-the-art AI/ML techniques to accomplish T&E of Department of Defense (DoD) systems and develop ethical and responsible methods for performing evaluation of weapon systems that are enabled by AI/ML technologies.

The research team participated in wide engagement with industry, academia, and government partners authoring policy, developing tools, engaging in workshops, symposiums, conferences and working groups, and publishing findings.

1 Link to *Digital Transformation in Test and Evaluation for AI/ML, Autonomous, and Continuously Evolving Systems - Base Year Report*

Figure 1 (below) illustrates the products and events that the AIRC research team has participated in, lead, and produced.

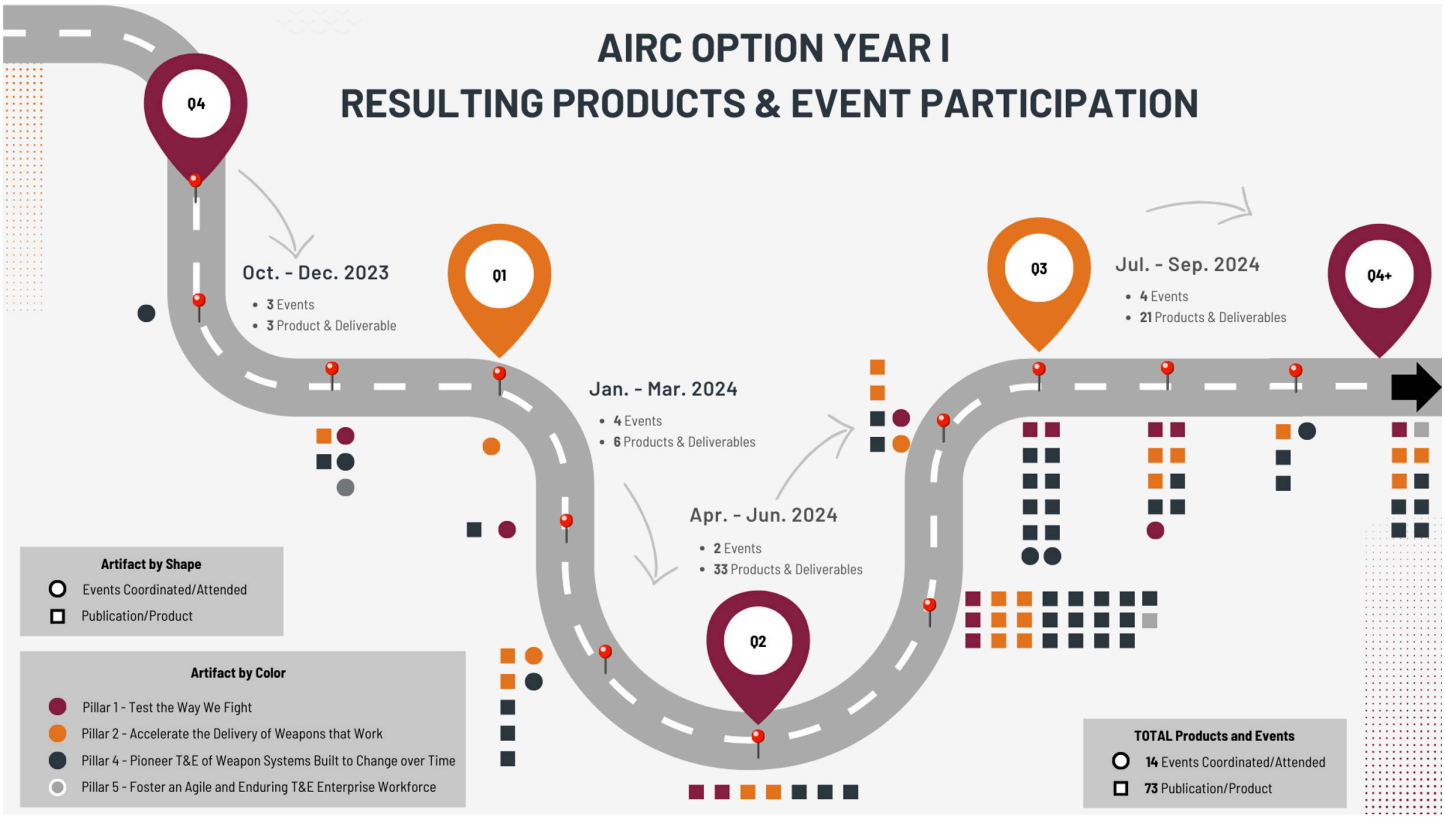

Figure 1: AIRC Option Year 1 Resulting Products and Event Participation

In support of Pillar 1, the AIRC researchers lead a series of three workshops to build and mature a Joint Test Concept (JTC) that enables the greater DoD community to assess joint forces capabilities. Fueled by an ever-expanding Community of Interest (COI), the research team's first workshop developed a test framework and reference architecture for planning and conducting joint test. Building on this framework and the fundamental JTC from year 1, the second workshop performed an actual simulation to evaluate the approaches and apply the framework. The final workshop focused on planning for a JTC Pilot that can further mature the methods. The research concludes that an execution roadmap coupled with pilots are necessary to continue to advance the communities' ability to conduct joint testing.

In support of Pillar 2, the research team matured integrated test concepts and supporting tooling and training materials. The research focused on Bayesian methods to integrate contractor test, development test, live fire test, and operational test to aid in shifting left operational assessments to enable more rapid fielding of capability to the warfighter. The team matured an R-Shiny application to aid the community in executing data analysis with Bayesian tools. In addition to the integrated test functions, the team expanded the original Integrated Development Support Key (IDSK) tooling with a multi-organization workshop that produced several IDSK exemplar tools. The IDSK is a critical element of a greater Test and Evaluation Master Plan (TEMP), which by policy outlines the test strategy for development of major capability for DoD platforms. Therefore, the research team matured an ontologically based toolset to enable the creation of TEMPs that are connected to real-time data for rapid and continual maturation of test planning artifacts to aid program offices and test organizations. The team intends to further mature learning aides and employ these capabilities in a pilot program in follow on research.

In support of Pillar 4, the team continued the maturation of methods to assess weapon systems that employ AI/ML capabilities and how to utilize tools such as Large Language Models (LLMs) to aid in assessing capabilities and operational suitability in DoD systems. The team, in partnership with industry and other government organizations, developed an AI/ML Test Hub environment to enable consistent and rapid assessment of these capabilities. The research team matured a test bed and framework to mature tooling and digital engineering practices for a more integrated environment that allows for maturing digital twins for system assessments. The team will continue maturing tools and training material as well as expand the test bed in follow on research efforts. Six research studies developed best practices, experimented with systematic approaches for learning important factors for AI model performance, developed new hierarchical scoring metrics, evaluated the impact of using systems theoretic process analysis (STPA) for evaluating AI ethics, evaluated how model-based test and evaluation could be used to test AI, and explored how multi-fidelity environments could be leveraged to build a body of evidence for testing AI. Two additional research studies looked at specific domains: optical sensors and cognitive electronic warfare. Finally, the research led to the development of one tool to assist testers in selecting test sets that cover a specified domain.

In addition to the above pillar-specific efforts, the research team partnered with DOT&E to revise the current T&E policy by providing technical consultation, reviewing and editing six policy documents, and planning for policy companion guide development for each of the new policy documents to help transition T&E practitioners to adapt their practices to be consistent with the new policy.

DISCLAIMER

The Acquisition Innovation Research Center (AIRC) is a multi-university partnership led and managed by the Stevens Institute of Technology and sponsored by the U.S. Department of Defense (DoD) through the Systems Engineering Research Center (SERC)—a DoD University-Affiliated Research Center (UARC).

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD(A&S)) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ0034-19-D-0003, TO#0510.

The views, findings, conclusions, and recommendations expressed in this material are solely those of the authors and do not necessarily reflect the views or positions of the United States Government (including the Department of Defense (DoD) and any government personnel), the Stevens Institute of Technology, or Virginia Tech National Security Institute.

No Warranty.

This Material is furnished on an "as-is" basis. The Stevens Institute of Technology and Virginia Tech National Security Institute make no warranties of any kind—either expressed or implied—as to any matter, including (but not limited to) warranty of fitness for purpose or merchantability, exclusivity, or results obtained from use of the material.

The Stevens Institute of Technology and Virginia Tech National Security Institute do not make any warranty of any kind with respect to freedom from patent, trademark, or copyright infringement.