ACQUISITION INNOVATION RESEARCH CENTER

# Test and Evaluation Methods for Middle-Tier Acquisition (MTA) – Option Year 1

EXECUTIVE SUMMARY AND REPORT

SEPTEMBER 2024

**PRINCIPAL INVESTIGATOR**

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

**CO-PRINCIPAL INVESTIGATOR**

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

VIRGINIA TECH

**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**

DISTRIBUTION STATEMENT A.
Approved for public release:
distribution unlimited.

## DISCLAIMER

*Copyright © 2024 Stevens Institute of Technology and Virginia Tech National Security Institute. All rights reserved.*

*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#0396.*

*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.*

# TABLE OF CONTENTS

DISCLAIMER ..... 2
TABLE OF CONTENTS ..... 3
LIST OF FIGURES ..... 7
LIST OF TABLES ..... 7
RESEARCH TEAM ..... 8
ACKNOWLEDGEMENTS ..... 10
ACRONYMS AND ABBREVIATIONS ..... 11
EXECUTIVE SUMMARY ..... 14
BACKGROUND ..... 17
PILLAR 1 – TEST THE WAY WE FIGHT ..... 18
JOINT TEST CONCEPTS ..... 18
RESEARCH OBJECTIVES ..... 18
METHODS ..... 18
RESULTS/FINDINGS ..... 19
RECOMMENDATIONS ..... 20
PILLAR 2 – ACCELERATE THE DELIVERY OF WEAPONS THAT WORK ..... 21
INTEGRATED TEST ..... 21
RESEARCH OBJECTIVES ..... 21
METHODS ..... 21
RESULTS/FINDINGS ..... 22
RECOMMENDATIONS ..... 22
DIGITAL TEST & EVALUATION MASTER PLAN (DTEMP) ..... 23
RESEARCH OBJECTIVE ..... 23
METHODS ..... 23
RESULTS/FINDINGS ..... 25
RECOMMENDATIONS ..... 27

OPERATION SAFE PASSAGE (OSP) ..... 27
RESEARCH OBJECTIVE ..... 27
METHODS ..... 28
RESULTS/FINDINGS ..... 28
RECOMMENDATIONS ..... 28
INTEGRATED DECISION SUPPORT KEY (IDSK) ..... 29
RESEARCH OBJECTIVES ..... 29
METHODS ..... 29
RESULTS/FINDINGS ..... 31
RECOMMENDATIONS ..... 31
VERIFICATION, VALIDATION, AND UNCERTAINTY QUANTIFICATION (VVUQ) ..... 32
RESEARCH OBJECTIVES ..... 32
METHODS ..... 32
RESULTS/FINDINGS ..... 33
RECOMMENDATIONS ..... 34
PILLAR 4 – PIONEER T&E OF WEAPON SYSTEMS BUILT TO CHANGE OVER TIME ..... 35
DIGITAL TWINS ..... 35
RESEARCH OBJECTIVE ..... 35
METHODS ..... 35
RESULTS/FINDINGS ..... 36
RECOMMENDATIONS ..... 36
UNCERTAINTY QUANTIFICATION FOR DIGITAL TWINS ..... 36
RESEARCH OBJECTIVES ..... 36
METHODS ..... 37
RESULTS/FINDINGS ..... 37
RECOMMENDATIONS ..... 39

T&E FOR AI/ML BEST PRACTICES GUIDE ..... 39
RESEARCH OBJECTIVES ..... 39
METHODS ..... 39
RESULTS/FINDINGS ..... 40
RECOMMENDATIONS ..... 41
SYSTEMATIC INCLUSION/EXCLUSION (SIE) RESEARCH FRAMEWORK ..... 42
RESEARCH OBJECTIVES ..... 42
METHODS ..... 42
RESULTS/FINDINGS ..... 43
RECOMMENDATIONS ..... 43
HIERARCHICAL SCORING ..... 43
RESEARCH OBJECTIVES ..... 43
METHODS ..... 44
RESULTS/FINDINGS ..... 44
RECOMMENDATIONS ..... 45
COGNITIVE ELECTRONIC WARFARE (COGEW) ..... 45
RESEARCH OBJECTIVES ..... 45
METHODS ..... 46
RESULTS/FINDINGS ..... 46
RECOMMENDATIONS ..... 47
MODEL-BASED TEST OF AI USING SYSML AND OPEN NEURAL NETWORK EXCHANGE (ONNX) ..... 47
RESEARCH OBJECTIVES ..... 47
METHODS ..... 48
RESULTS/FINDINGS ..... 49
RECOMMENDATIONS ..... 49
USE CASE: T&E FOR AI/ML PROTOTYPING T&E METHODS FOR ROBUSTNESS TO REAL-WORLD SITUATIONS FOR AN OPTICAL SENSOR ..... 49
RESEARCH OBJECTIVES ..... 49
METHODS ..... 49
RESULTS/FINDINGS ..... 50
RECOMMENDATIONS ..... 50

SYSTEMS THEORETIC PROCESS ANALYSIS (STPA) FOR AI ETHICS ASSESSMENT ..... 51
RESEARCH OBJECTIVES ..... 51
METHODS ..... 51
RESULTS/FINDINGS ..... 52
RECOMMENDATIONS ..... 52
COVERAGE OF DATA EXPLORER (CODEX) ..... 53
RESEARCH OBJECTIVES ..... 53
METHODS ..... 54
RESULTS/FINDINGS ..... 55
RECOMMENDATIONS ..... 55
TEST & EVALUATION FOR MULTI-FIDELITY AI MODELS ..... 55
RESEARCH OBJECTIVES ..... 55
METHODS ..... 55
RESULTS/FINDINGS ..... 56
RECOMMENDATIONS ..... 57
CONCLUSIONS ..... 58
APPENDIX A. DELIVERABLES AND PRODUCTS ..... 59
PILLAR 1 – TEST THE WAY WE FIGHT ..... 59
PILLAR 2 – ACCELERATE THE DELIVERY OF WEAPONS THAT WORK ..... 60
PILLAR 4 – PIONEER T&E OF WEAPON SYSTEMS BUILT TO CHANGE OVER TIME ..... 64
PILLAR 5 – FOSTER AN AGILE AND ENDURING T&E ENTERPRISE WORKFORCE ..... 67
APPENDIX B. LIST OF PUBLICATIONS RESULTED ..... 68
REFERENCES ..... 70

## LIST OF FIGURES

FIGURE 1: AIRC OPTION YEAR 1 RESULTING PRODUCTS AND EVENT PARTICIPATION ..... 14
FIGURE 2: TEST PRACTITIONER DASHBOARD ..... 25
FIGURE 3: THE DIGITAL IDSK, TIP OF THE ICEBERG OF THE T&E ENVIRONMENT FOR DECISION MAKING ..... 29
FIGURE 4: SOBOL' MAIN EFFECTS INDICES FOR QOI ..... 32
FIGURE 5: HIGH-LEVEL SYSML DESCRIPTION OF A LEARNING MODEL CAPTURED IN ONNX ..... 47
FIGURE 6: REDUCTION IN UNCERTAINTY ABOUT THE REQUIREMENTS, MEASURED USING ENTROPY ..... 56

## LIST OF TABLES

TABLE 1: PROPOSAL TASKS TO I-PLAN MAPPING ..... 16
TABLE 2: MB TEMP ONTOLOGIES ..... 23
TABLE 3: NON-EXHAUSTIVE SUMMARY OF SUGGESTED UQ TECHNIQUES THAT CAN BE APPLIED TO ADDRESS SPECIFIC UQ CHALLENGES ARISING IN COMPLEX SYSTEMS ANALYSIS (CORTES, WONG AND CORTES-MORALES). ..... 37
TABLE 4: COMPARISON OF THE NUMBER OF TESTS NEEDED AND THE CORRESPONDING COSTS ..... 56

## RESEARCH TEAM

| Name | Organization | Labor Category | Research Task |
|---|---|---|---|
| Laura Freeman | Virginia Tech National Security Institute (VTNSI) | Principal Investigator | WRT-1070; WRT-1071 |
| Geoffrey Kerr | VTNSI | Co-Principal Investigator | WRT-1070; WRT-1071 |
| Orlando Florez | VTNSI | Associate Director for Program Management | WRT-1070; WRT-1071 |
| Sanglin Chang | VTNSI | Project Manager | WRT-1070; WRT-1071 |
| Anna Flowers | VTNSI | Graduate Research Assistant | WRT-1070 |
| Brian Lee | VTNSI | Research Data Analyst | WRT-1070; WRT-1071 |
| Dan DeCollo | VTNSI | Research Data Analyst | WRT-1071 |
| Dylan Steburg | VTNSI | Graduate Research Assistant | WRT-1070 |
| Emma Meno | VTNSI | Research Associate | WRT-1070 |
| Erik Higgins | VTNSI | Research Scientist | WRT-1070; WRT-1071 |
| Erin Lanus | VTNSI | Research Assistant Professor | WRT-1070 |
| Jared Clark | VTNSI | Graduate Research Assistant | WRT-1070 |
| John Gilbert | VTNSI | Assistant Director, Mission Systems Division | WRT-1070; WRT-1071 |
| Justin Kauffman | VTNSI | Research Assistant Professor | WRT-1070; WRT-1071 |
| Justin Krometis | VTNSI | Research Assistant Professor | WRT-1070 |
| Kelli Esser | VTNSI | Associate Director, Intelligent Systems Division | WRT-1070; WRT-1071 |
| Kyle Risher | VTNSI | Undergraduate Research Intern | WRT-1070 |
| Nicola McCarthy | VTNSI | Research Assistant Professor | WRT-1071 |
| Paul Hess | VTNSI | Adjust professor / Senior Advisor, Naval Engineering | WRT-1071 |
| Paul Wach | VTNSI | Research Assistant Professor | WRT-1071 |
| Peter Beling | VTNSI | Director, Intelligent Systems Division | WRT-1070; WRT-1071 |

| Name | Organization | Labor Category | Research Task |
|---|---|---|---|
| Tim Sherburne | VTNSI | Research Associate | WRT-1071 |
| Tyler Cody | VTNSI | Research Assistant Professor | WRT-1070; WRT-1071 |
| Jaganmohan Chandrasekaran | Virginia Tech Sanghani Center for AI and Data Analytics | Research Assistant Professor | WRT-1071 |
| Brandon Stringfield | Georgia Tech Research Institute (GTRI) | Senior Research Scientist | WRT-1071 |
| Jennifer Sharpe | GTRI | Research Engineer II | WRT-1071 |
| Retonya Brinkley | GTRI | Research Scientist II | WRT-1071 |
| Alton (AJ) Schultheis | GTRI | Research Engineer I | WRT-1071 |
| John Rafferty | GTRI | Research Scientist II | WRT-1071 |
| Jonathon Giles | GTRI | Research Engineer I | WRT-1071 |
| David Narehood | Penn State Applied Research Laboratory (ARL) | Department Head, Model-Based Engineering Division | WRT-1070 |
| Alyssa Sharrar | Penn State ARL | Undergraduate Research Assistant | WRT-1070 |
| Andrew Hoskins | Penn State ARL | Research and Development Engineer | WRT-1070 |
| Andrew Shaffer | Penn State ARL | Research and Development Engineer | WRT-1070 |
| Domenic Nelson | Penn State ARL | Undergraduate Research Assistant | WRT-1070 |
| Justin Valenti | Penn State ARL | Researcher | WRT-1070 |
| Michael Warren | Penn State ARL | Undergraduate Research Assistant | WRT-1070 |
| Sheri Martinelli | Penn State ARL | Department Head, Simulation Software | WRT-1070 |
| William Laplante | Penn State ARL | Undergraduate Research Assistant | WRT-1070 |
| Jitesh Panchal | Purdue University | Associate Head of Undergraduate Programs – School of Mechanical Engineering | WRT-1070 |
| Karen Marais | Purdue University | Professor, Associate Head for Undergraduate Education | WRT-1070 |
| Jin-Suh Park | Purdue University | Graduate Research Assistant | WRT-1070 |
| Robert Seif | Purdue University | Graduate Research Assistant | WRT-1070 |
| Sanidhya Jain | Purdue University | Graduate Research Assistant | WRT-1070 |
| Zichong Yang | Purdue University | Graduate Research Assistant | WRT-1070 |
| Alejandro Salado | The University of Arizona (UoA) | Associate Professor | WRT-1071 |
| Bennett Jackson | UoA | Undergraduate Researcher | WRT-1071 |

| Name | Organization | Labor Category | Research Task |
|---|---|---|---|
| Joe Gregory | UoA | Postdoctoral Research Associate | WRT-1071 |
| Nate Bushong | UoA | Graduate Researcher | WRT-1071 |
| Samuel Cornejo | UoA | Graduate Researcher | WRT-1071 |
| Samuel Victor | UoA | Undergraduate Researcher | WRT-1071 |
| Sarthak Halder | UoA | Graduate Researcher | WRT-1071 |
| Visalakshi Iyer | UoA | Graduate Researcher | WRT-1071 |
| Maegen Nix | Virginia Tech Applied Research Corporation (VT-ARC) | Director, Decision Science Division | WRT-1070 |
| Christina Houfek | VT-ARC | Lead Project Manager | WRT-1070 |
| Natalie Wells | VT-ARC | Lead Project Manager and Systems Engineer | WRT-1070 |
| Daniel Wolodkin | VT-ARC | Staff Data Scientist | WRT-1070 |
| Grant Beanblossom | VT-ARC | Lead Data Scientist | WRT-1070 |
| Kobie Marsh | VT-ARC | Associate Data Scientist and Software Engineer | WRT-1070 |

## ACKNOWLEDGEMENTS

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 authors would like to thank DOT&E SIPET for their support and engagement in shaping new methods for the future of test and evaluation (T&E). The methods developed in this research reflect emerging technologies, a changing threat landscape, and the need for efficient and effective T&E to ensure capabilities delivered to warfighters work as intended when called upon.

# ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| **AFIT** | Air Force Institute of Technology |
| **AI** | Artificial Intelligence |
| **AIES** | Artificial Intelligence-Enabled Systems |
| **AI/ML** | Artificial Intelligence/Machine Learning |
| **AIRC** | Acquisition Innovation Research Center |
| **AIWG** | Artificial Intelligence Working Group |
| **ASME** | American Society of Mechanical Engineers |
| **AV** | Autonomous Vehicle |
| **CAD** | Computer Aided Design |
| **CAP** | Cyber Assessment Program |
| **CIL** | Capability Immersion Layer |
| **CODEX** | Coverage of Data Explorer |
| **CogEW** | Cognitive Electromagnetic Warfare |
| **COI** | Community of Interest |
| **CT** | Combinatorial Testing |
| **DARPA** | Defense Advanced Research Projects Agency |
| **DATAWorks** | Defense and Aerospace Test and Analysis (DATA) Workshop |
| **DE** | Digital Engineering |
| **DoD** | Department of Defense |
| **DoDI** | Department of Defense Instruction |
| **DoE** | Design of Experiments |
| **DOT&E** | Director, Operational Test and Evaluation |
| **DT** | Digital Twin |
| **dTEMP** | Digital Test and Evaluation Master Plan |
| **EA** | Ethics Assessment (of adherence to the EPs) |
| **EPs** | DoD's Ethical Principles |
| **EW** | Electromagnetic Warfare |
| **FY** | Fiscal Year |
| **GTRI** | Georgia Tech Research Institute |
| **GWEF** | Guided Weapons Evaluation Facility |

| | |
|---|---|
| **HTV** | Hypersonic Technology Vehicle |
| **HTV-2** | Hypersonic Technology Vehicle 2 |
| **IDA** | Institute for Defense Analyses |
| **IDR** | Interim Design Review |
| **IDSK** | Integrated Decision Support Key |
| **I-Plan** | Implementation Plan |
| **IP** | Intellectual Property |
| **JCDL** | Joint Capability Demonstration Layer |
| **JSON** | JavaScript Object Notation |
| **JTC** | Joint Test Concept |
| **J-TEST** | Joint Test and Evaluation Strategy Team |
| **LFT&E** | Live Fire Test & Evaluation |
| **LLM** | Large Language Model |
| **LVC** | Live, Virtual, Constructive |
| **M&S** | Modeling and Simulation |
| **MAE** | Mean Absolute Error |
| **MBSE** | Model-Based Systems Engineering |
| **MB TEMP** | Model-Based Test and Evaluation Master Plan |
| **ML** | Machine Learning |
| **MSE** | Mean Squared Error |
| **MTA** | Middle Tier of Acquisition |
| **NIST** | National Institute of Standards and Technology |
| **OML** | Ontological Modeling Language |
| **ONNX** | Open Neural Network Exchange |
| **OSP** | Operation Safe Passage |
| **OT** | Operational Testing |
| **OT&E** | Operational Test and Evaluation |
| **PCE** | Polynomial Chaos Expansion |
| **QOI** | Quantity (or Quantities) of Interest |
| **RF** | Radio Frequency |
| **RMSE** | Root Mean Squared Error |
| **SAR** | Synthetic Aperture Radar |

| | |
|---|---|
| **SIE** | Systematic Inclusion/Exclusion |
| **SIPET** | Strategic Initiatives, Policy, and Emerging Technologies |
| **SPL** | System Performance Layer |
| **STAT COE** | Scientific Test & Analysis Techniques Center of Excellence |
| **STPA** | Systems Theoretic Process Analysis |
| **SUT** | System Under Test |
| **SysML** | System Modeling Language |
| **T&E** | Test and Evaluation |
| **TEMP** | Test and Evaluation Master Plan |
| **TETRA** | Test and Evaluation Threat Resource Activity (under DOT&E) |
| **TRF** | Test Resource Facility (or Facilities) |
| **TRMC** | Test Resource Management Center |
| **UAOS** | University of Arizona Ontology Stack |
| **UARC** | University Affiliated Research Center |
| **UAV** | Unmanned Aerial Vehicle |
| **UGV** | Unmanned Ground Vehicle |
| **UoA** | University of Arizona |
| **UQ** | Uncertainty Quantification |
| **VT** | Virginia Tech |
| **VTNSI** | Virginia Tech National Security Institute |
| **VVUQ** | Verification, Validation, and Uncertainty Quantification |
| **XML** | Extensible Markup Language |

# 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 *Test and Evaluation Methods for Middle-Tier Acquisition (MTA) - Base Year Report*

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

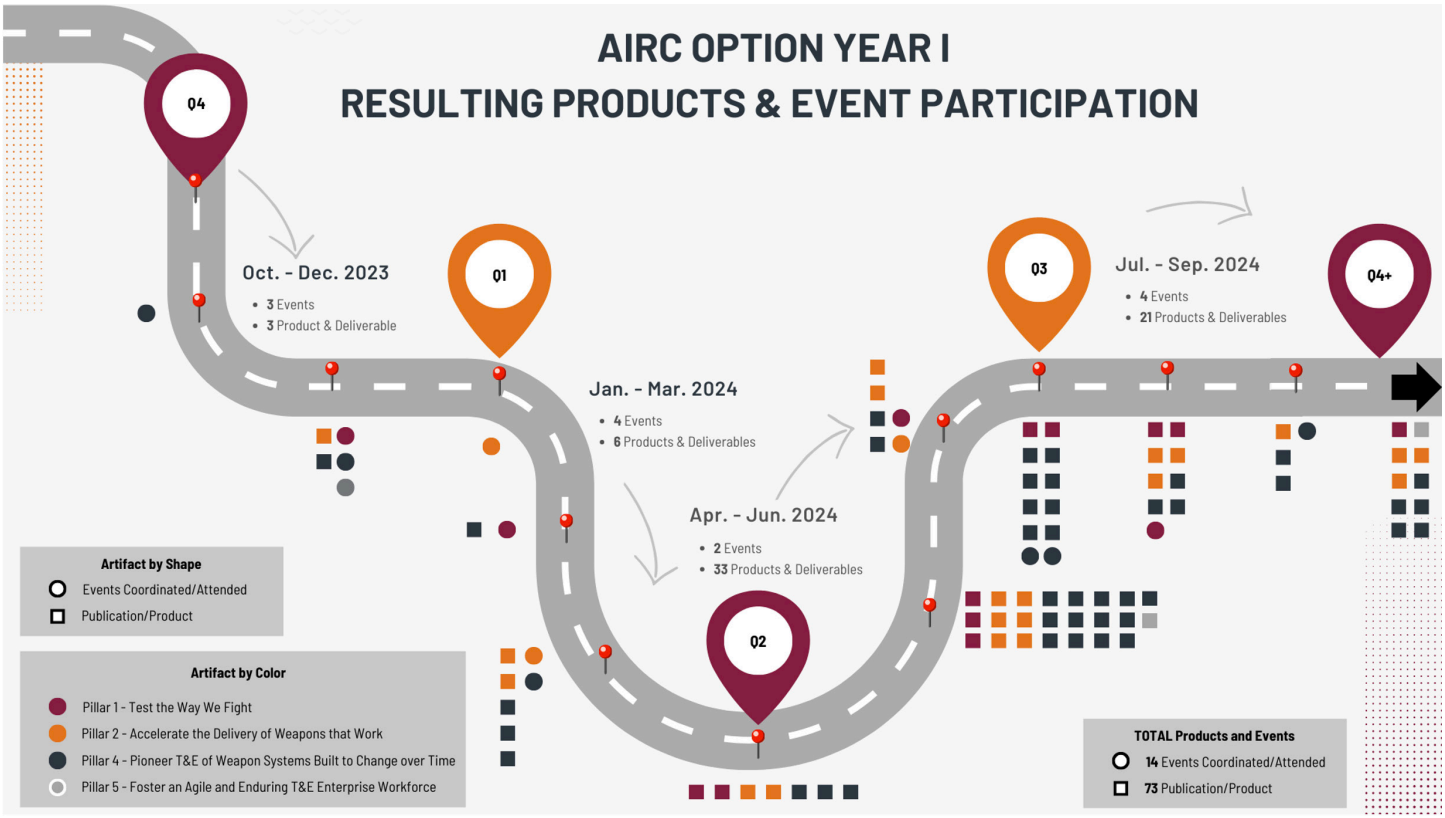


**Figure 1: AIRC Option Year 1 Resulting Products and Event Participation**

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.

# BACKGROUND

The Director, Operational Test and Evaluation (DOT&E) technical staff has engaged with the Acquisition Innovation Research Center (AIRC) University Affiliated Research Center (UARC) to advance tooling and processes for executing test and evaluation (T&E) on Department of Defense (DoD) systems. With a focus on advances in Digital Engineering (DE), Artificial Intelligence (AI), Integrated Testing, Joint operation evaluations, and other key components to improvements in DoD acquisition. The research team, in partnership with DOT&E, embarked on advancing T&E practices in September of 2022. This report addresses the accomplishments of the option-year effort as the research team focused on building on the year one foundation. In the option year, the team matured concepts and best practices to address new capabilities and methods for T&E.

The two base contracts were originally planned to focus specifically on Middle Tier of Acquisition (MTA) and Digital Transformation. After the contracts were issued, in close coordination with the DOT&E sponsor, the AIRC team integrated the two efforts to maximize the synergy of the contracted tasks and to best align with the DOT&E Implementation Plan (I-Plan). Please see Table 1 (below) for a mapping of the original proposed work to the lines of research and DOT&E I-Plan pillars.

**Table 1: Proposal Tasks to I-Plan Mapping**

| | | | Pillar 1 | Pillar 2 | | | Pillar 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Integrated Testing | | Digital Engineering | | T&E for AIES | |
| | | | Joint Test Concept | Data Strategy/Security | Bayesian Sequential | MBTEMP/IDSK | Digtal Twins/VVUQ&A | T&E for AIES | AI for T&E |
| WRT-1070 | MTA-1 | Strategic Planning for Int. T&E | | X | X | | | | |
| | MTA-2 | Integrated T&E Harness | | | X | | | | |
| | MTA-3 | Interoperability Testing in Complex, Evolving Network Centric Systems | X | | | | | | |
| | MTA-4 | Test Requirements for IP | | | | | | X | |
| | MTA-5 | Test-driven SE | | | | X | X | | |
| | MTA-6 | Automation to Support Penetration Testing | | | | | | | X |
| | MTA-7 | Workforce Development for Next Gen T&E | Unfunded Line of Effort | | | | | | |
| | MTA-8 | DOT&E Portfolio Coordination and Outreach | Cuts Across Pillars | | | | | | |
| | | | | | | | | | |
| WRT-1071 | DT-1 | Digital TEMPs | | | | X | | | |
| | DT-2 | Digital Engineering Enhanced T&E for AI Systems | | | | | X | X | |
| | DT-3 | M&S V&V | | | | | X | | |

Seventy-one distinct publications and products result from this effort. This report summarizes the key results, but all of the products are listed in the appendix and available for more information. For each pillar the research team breaks out the specific project, objectives, methods, results, and recommendations (if applicable).

# PILLAR 1 – TEST THE WAY WE FIGHT

## JOINT TEST CONCEPTS

### *RESEARCH OBJECTIVES*

The Year II Joint Test Concept (JTC) study design maintained the three-phase approach to achieve three distinct but interrelated goals:

- *Goal 1:* Create a JTC reference architecture that ensures data quality, accessibility, utility, and analytic value across existing and emergent Joint mission (kill/mission) webs for all systems under test throughout the entire capability lifecycle.
- *Goal 2:* Assess the reference architecture performance through an end-to-end capability lifecycle T&E architecture simulation.
- *Goal 3:* Develop a JTC Implementation Roadmap including quick win opportunities.

### *METHODS*

Relying on Year I findings and the expanding Community of Interest (COI) to provide input toward each goal, the study team executed a series of three workshops, each one primarily focused on one goal. The workshop outcomes were then analyzed, synthesized, and utilized to inform the following workshops and forthcoming Draft Joint Test Concept v1.0.

- Design the JTC Architecture (February 2024)
  - » Design an overarching JTC reference architecture that facilitates traceability, integration and interoperability across stakeholders and supports effective and efficient decision-making dependent on appropriate data collection, storage, and maintenance which enables JTC integrated architecture design and development.
- Evaluate the JTC Architecture (June 2024)
  - » Run a JTC Integrated Architecture Use Case through a series of simulations to examine JTC T&E assessments across all three JTC layers[2], identify the impacts of common challenges (particularly related to data), and inform the JTC implementation roadmap.
    - ◊ Investigate how the JTC application to both materiel and non-materiel assessments and kill web (mission web) vulnerabilities and performance.
    - ◊ Explore how programs of record might be impacted by the application of the JTC approach.
    - ◊ Examine the role of the Joint Test & Evaluation Strategy Team (J-TEST)
    - ◊ Evaluate participant learning through JTC application challenges.

[2] The JTC Framework consists of three overlapping non-hierarchical JTC Layers for evaluation areas to incorporate iteratively across the capability lifecycle. The layers include Joint Capability Demonstration Layer (JCDL), Capability Immersion Layer (CIL), and System Performance Layer (SPL).

- Develop a JTC Implementation Roadmap (August 2024)
  - » Expand on J-TEST roles and responsibilities, identify key stakeholder roles, and capture key stakeholder interactions critical for JTC implementation.
  - » Identify a three-stage pathway to JTC implementation with short-, mid-, and long-term milestone goals which create value propositions across the diverse stakeholder community.

### *RESULTS/FINDINGS*

*Goal 1:* The resultant overarching JTC Pilot reference architecture defines the JTC core functions and provides high-level architectural and process elements important to JTC implementation. This reference architecture will function as an authoritative information source that guides and constrains JTC T&E campaign of learning implementations and furthers the development of future detailed architectures that are derived, decomposed, and traceable to the overarching JTC reference architecture.

*Goal 2:* Workshop outcomes demonstrated that JTC implementation could:

- Generate multiple, efficiency-enhancing feedback loops, support Joint readiness assessments, help reduce technical and mission performance risks, and enable reliability, maintainability, and availability assessments early in the capability lifecycle.
- Encourage common data structures and sharing protocols, optimize schedule management, improve communication, and reduce barriers to leveraging training exercises, provide additional resourcing and information.
- Enable T&E strategy and system requirements updates, contribute to system obsolescence identification, identify system trends across services and joint interfaces between service systems and domains, assess supply chain stability, identify and foster joint Live, Virtual, Constructive (LVC) opportunities, and help overcome collaboration barriers between organizations.
- The workshop's iterative methodology is applicable to current practices and using a step-by-step iterative process could result in an ideal scenario for current programs of record.

*Goal 3:* The resultant outcomes provided a deeper understanding of key roles and responsibilities of the J-TEST, critical stakeholder interactions and offerings to JTC implementation, and implementation activities and deliverables needed for JTC adoption across the stakeholder community. The team gathered detailed COI recommendations for the implementation roadmap that are synthesized in the JTC Implementation Roadmap Workshop Outcomes Report and will be utilized in a forthcoming JTC implementation plan product early in the next year (Fiscal Year [FY] 25).

### *RECOMMENDATIONS*

Outcomes culminated across this study's three-phase approach illuminated critical needs surrounding JTC implementation. Findings captured throughout this study will be expanded on and utilized in the forthcoming JTC implementation plan product. A summary of key next steps required to advance the JTC into practice include:

- Develop detailed architectures derived, decomposed, and traceable to the overarching JTC reference architecture.
- Develop detailed documentation of stakeholder and J-TEST roles and responsibilities, an adequate cross-stakeholder governance structure with a lead integrator, and an organizational change management strategy to support JTC implementation and adoption.
- Develop a financial management and resourcing process critical to JTC implementation.
- Launch pathfinder initiatives to demonstrate short-term wins and value of JTC while also informing further process and architecture refinement.
- Accelerate JTC implementation and adoption through policy reform and alignment.
- Design and implement adequate training and education, including detailed guidebooks and knowledge repositories surrounding JTC stakeholder offerings, process elements, and architectures.
- Identify the right champions to advocate for and communicate the vision to maintain stakeholder buy-in, support, and momentum along the implementation phases.

# PILLAR 2 – ACCELERATE THE DELIVERY OF WEAPONS THAT WORK

## INTEGRATED TEST

### *RESEARCH OBJECTIVES*

The Integrated Testing research, which aligns with DOT&E's I-Plan, Pillar 2.2.1, "Accelerate the development of tools that enable adequate performance inference from a growing body of evidence" was done in collaboration between Virginia Tech and the Scientific Test and Analysis Techniques Center of Excellence (STAT COE) researchers and involved regular coordination with Metron on their work on Bayesian decision theory.

The team's primary objective was to better understand the characteristics of operational systems by developing and illustrating Bayesian inference techniques for leveraging all available data, including potentially dissimilar data from earlier in the program life cycle. Based on the research, leveraging all available data should enable more accurate and more precise estimates of system parameters and more efficient use of operational testing (OT) resources.

### *METHODS*

Bayesian inference is a technique for estimating unknown parameters from data and involves three key components:

- The *prior*, a probability distribution that describes what is known about the parameters before data is collected.
- The *likelihood*, a function that describes the data's relationship to the parameters.
- The *posterior*, a probability distribution that describes what is known about the parameters after data is incorporated into the model. The posterior is the "answer" in a Bayesian setting and has high probability for parameter values that agree with both the data and the prior and has lower probability for parameter values that exhibit significant mismatch with one or the other.

Much of the project work on Integrated Testing focused on how to use priors to carry information across the acquisition lifecycle. For example, early in the system acquisition, priors may be based on known constraints, subject matter expertise, or data from previous similar systems. As early tests are conducted, those initial priors can be combined with the resulting data to produce updated estimates of the system. The resulting posteriors can in turn be used as priors in updates using subsequent test data, thereby refining the understanding of the system as it develops. Should either the system or its environment change during development, the priors can be modified accordingly via downweighing, for example, to convey additional uncertainty reflecting the changes in the system or associated conditions.

This year's efforts involved applying Bayesian methods to DoD systems, systems involving artificial intelligence and machine learning (AI/ML), and model-based systems engineering (MBSE) programs. The research team communicated results via tutorials, conference presentations, journal articles, and software tools.

### *RESULTS/FINDINGS*

One key area the research team focused on was the effect of assumptions on the results of Bayesian inference. The team used the reliability of the Stryker family of vehicles as an example to conduct a systematic comparison of several Bayesian approaches for integrating data from developmental and operational testing. The analysis showed the benefit of using the Bayesian approach to integrate information, but also the importance of using careful and justifiable assumptions when developing priors and data models to ensure defensible results. The team submitted the paper *A Framework for Using Priors in a Continuum of Testing* to the Military Operations Research Journal during this period of performance; this paper builds on work done in the base year of the contract and will be published in the Fall 2024.

The research team collaborated with DOT&E leadership to write an article on Bayesian methods for T&E for the Naval Engineers Journal. The article focused on a notional torpedo example and leveraged the R-Shiny reliability app (see below) for analysis. The team is also working on Bayesian analysis for the Institute for Defense Analyses' (IDA) notional counterfire radar model and the integration of Bayesian methods into Operation Safe Passage to estimate physics model parameters from data.

Bayesian methods also show promise for understanding the behavior of AI/ML systems; the Integrated Testing research team has worked to illustrate how Bayesian uncertainty can be used to estimate confidence in the accuracy of AI/ML models in various circumstances and to determine the best training data to collect to augment that knowledge.

Finally, this line of effort has also developed presentations and software products to communicate Bayesian methods and best practices to stakeholders. Research team leads collaborated with Metron personnel to present a mini-tutorial on Bayesian Methods for Integrated Testing at the 2024 Defense and Aerospace Test and Analysis Workshop (DATAWorks). The presenters walked participants through the application of key concepts to T&E then guided the audience through a detailed example on a notional system.

The team also produced a series of web applications using the R-Shiny tool to allow users to conduct Bayesian inference for reliability and binary (pass/fail or hit/miss) data without having to write any code. The purpose of these prototypes was to help illuminate concepts and illustrate what software interfaces for developing priors and Bayesian inference could look like in the future.

### *RECOMMENDATIONS*

The papers and presentations generated by this research have increased awareness of Bayesian methods and illustrated their promise when applied to the test and evaluation of defense systems. The effort also helped generate a growing base of both front-end and back-end software for applying these methods. Future efforts should build out additional examples that represent relevant challenges for the DoD testing community. Additionally, foundational work is needed in Bayesian experimental design, to include the application of active learning and space-filling techniques for selecting test points to maximize the information gained. Assembling an increasingly broad set of exemplars would assist the DoD in the application of Bayesian methods to T&E.

## DIGITAL TEST & EVALUATION MASTER PLAN (DTEMP)

### *RESEARCH OBJECTIVE*

This research effort's objective was to digitalize the Test and Evaluation Master Plan (TEMP) (as defined in DoDI 5000.89) as a Model-Based TEMP (MB TEMP) to support automated consistency checking, issue detection, report and dashboard generation, and data collection with regards to test strategy planning and execution.

### *METHODS*

A possible approach to digitalize the TEMP is to leverage semantic web technologies. Ontologies and other semantic web technologies provide a means of structuring data in a way that offers potential in terms of reasoning and querying capabilities. This approach enables the inference of knowledge and is particularly well-suited for highly connected data by providing the opportunity to query the data more efficiently than relational databases. Other motivations for the use of semantic web technologies include the elimination of data heterogeneity/semantics problems, the improvement of interoperability, the facilitation of data integration, and the provision of semantic content for requirements and data analysis.

This work leveraged existing research on T&E ontologies and digital verification processes. The MB TEMP is modular and currently comprises nine individual ontologies with dependencies between them. A summary of the nine ontologies and the relevant section(s) in DoDI 5000.89 is listed in Table 2.

**Table 2: MB TEMP Ontologies**

| MB TEMP Ontology | Description | Section of DoDI 5000.89 |
|---|---|---|
| Organization | Defines the roles that organizations can play in terms of the TE strategy and their structure. | Section 1.1: Applicability<br>Section 2.1: Director of OTE<br>Section 2.2: USD: (R&E) |
| Test Program Policy | Defines the purpose of a Test Program and its relationship with a Defense Acquisition Program. | Section 1.2: Policy |
| Test Program Structure | Defines the types of Test Program and how they combine to form the overall Test Program. | Section 3.1: Overview |
| Test Program Frameworks | Defines the frameworks used to prescribe Test Programs and their structure. | Section 3.1: Overview |
| Execution and Reporting | Defines the types of output expected to be produced by the Test Programs, including reports and evaluation models. | Section 2.1: Director of OTE<br>Section 3.1: Overview<br>Section 3.3: TE Management |
| Test Program Teams | Defines the expected outputs of the functions that the organizations and individuals undertake. | Section 3.1: Overview<br>Section 3.4 TE Program Planning |
| Decision | Defines the terminology associated with decisions and the agents responsible for them. | Section 3.1: Overview<br>Section 3.2: TE Oversight List |
| Responsibility | Defines the responsibilities that organizations and individuals have. | Section 3.3: TE Management<br>Section 3.4 TE Program Planning |
| TE Oversight | Defines the technical oversight and the functions associated with this (e.g., notification of military). | Section 3.2: TE Oversight List |

The MB TEMP builds on ontology stack, which provides modularity, and is how fundamental concepts (such as 'Test') and relations (such as 'verifiedBy') are defined. The MB TEMP has been written in the Ontological Modeling Language (OML) and developed using OML Rosetta – both the language and tool are developed and maintained by the Jet Propulsion Laboratory as part of the OpenCAESAR initiative. As demonstrated in Table 2 (above), the MB TEMP provides a framework to capture T&E organization and responsibilities, approval structure, reporting, and decision-making. Its foundation on the University of Arizona Ontology Stack (UAOS) also enables the modeling of key measures (e.g., Measures of Effectiveness), mission and system descriptions, threat descriptions, relations between tests and requirements, and test results.

### *RESULTS/FINDINGS*

In this project, the research team has demonstrated that ontologies and modeling can be used to digitalize T&E planning and execution and has done so in a technology- and vendor-agnostic manner. Particularly:

- Test strategies and their documentation have been modeled in OML compliant to DoD instruction.
- Test strategy planning and execution data directly established in OML have been integrated with mission and system models established in System Modeling Language (SysML) v2.
- This comprehensive dataset was automatically checked (for consistency and completeness), reasoned across, and queried. Different use cases have shown the ability of the MB TEMP to report the status of the T&E strategy, anticipate issues and/or conflicts emerging from data related to the status of the different elements in the test strategy (e.g., systems, components, test equipment, personnel, facilities, etc.), and integrate requirements, constraints, and guidelines across different subject matter experts.
- Convenient representations of query results can be displayed on dashboards to aid practitioners. The example in Figure 2 is the 'Testing' dashboard, which displays key information relevant to a test practitioner including the number of defined tests, the test schedule, and an overview of results so far. In this example, the ontology has also inferred that there may be scheduling conflicts that the test practitioner should address.
- Test planning and execution data have been integrated with analytical models (including a Bayesian network) to enable quantitative evaluation of the value of different test activities.

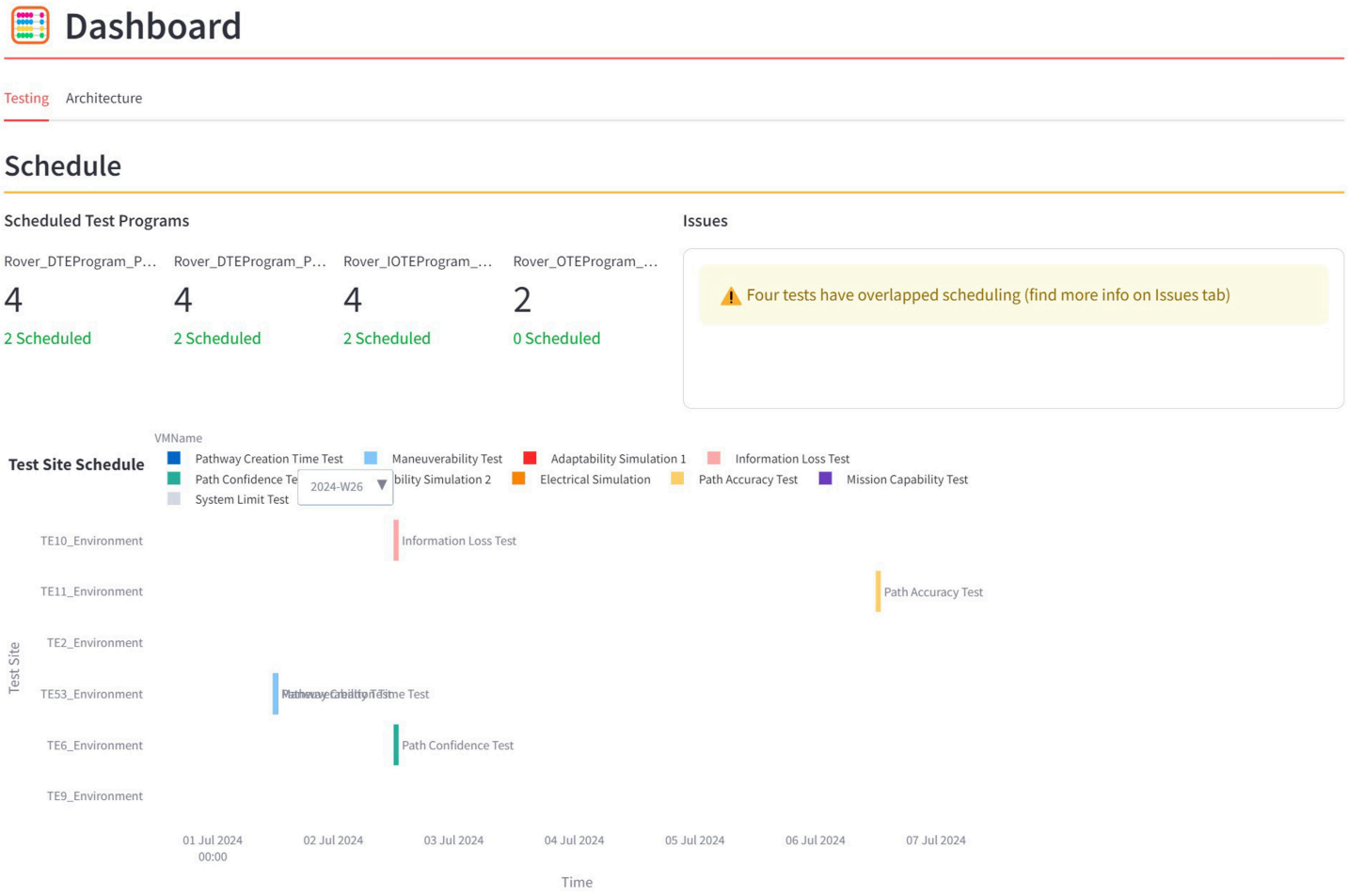


**Figure 2: Test Practitioner Dashboard**

### *RECOMMENDATIONS*

The research team recommends the following future work:

- Develop User Interfaces to support the input of test strategy data. Instead of modeling test strategy information as OML code, test engineers could model their test strategy as in SysML v2, for example, and the representative OML be automatically generated using the Violet tool.
- Develop a User Interface using Large Language Models (LLMs) or a similar technology to enable test modelers and other stakeholders to engage with the MB TEMP using natural language. The LLM on the backend could translate natural language into formal queries, decoupling workforce capabilities from technological evolution.
- Align the MB TEMP with the latest TEMP instruction (DoDM 5000.UX).
- Develop different classes of MB TEMPs for a notional mission with corresponding guidance to support programs at different stages of digital transformation transition towards MB TEMP.
- Transition the proposed solution into a real program to develop and test the scalability of the solution and develop corresponding guidance for adoption to a program.

## OPERATION SAFE PASSAGE (OSP)

### *RESEARCH OBJECTIVE*

The research team's focus for Operation Safe Passage (OSP) was to leverage a mock program to produce a test bed and framework for developing and integrating:

- Mission engineering models
- Systems engineering models
- Physical design
- Modeling and simulation
- Test execution
- Test planning
- Programmatic decision making

This mock or proxy program is unclassified and provides an inexpensive environment for maturing tooling and methods to achieve the tooling integration objective. Furthermore, the program makes it easy to iterate on tools and methods to accelerate research advancements.

### *METHODS*

To ease and enhance the digital transformation of T&E, the research team developed and matured an affordable open-source solution that can serve as a proxy for a DoD system acquisition. The solution consists of a scalable cyber-physical system to frame and test out tooling capability and T&E methods for educating and advancing acquisition offices and test professionals. Together, the team developed a framework and testbed for maturing DE tooling and methods in support of T&E of weapon systems.

The framework and testbed are representative of development and test environments. Starting with a notional mission definition (Operation Safe Passage) requiring an operational capability to move cargo and troops across an enemy minefield without the use of airlift, an Unmanned Ground Vehicle (UGV) was conceptualized with system specifications. Once the fictitious system was specified, the team created a proxy system utilizing the Lego Mindstorm ™ robot as a platform and generated a parallel mission and system to equate to the envisioned, "real" platform. The extensive Lego ecosystem consists of tangible digital and physical artifacts that enable controlled creation and growth of the framework and testbed.

The DE environment consists of SysML-based architecture models (e.g., mission and system), Computer Aided Design (CAD) models, physics-based analytical models, and decision dashboards. The research team generated a MB TEMP with a supporting Integrated Decision Support Key (IDSK) that links to the architecture models. A physical test environment/range has been established to execute physical testing of the UGV system. The research team has demonstrated that the thread between architecture models, test planning models, analytical models, and physical models enables rapid decisions based on the evaluation of accumulated results. These results inform test planning to highlight where test plan adjustments will be required or when results are secured. All the tools developed by the team are open-source and easily adaptable to multiple tool suites to enable ease of adoption in real acquisition.

### *RESULTS/FINDINGS*

The research team conducted a mock Interim Design Review (IDR) with DOT&E representatives to illustrate the system in the integrated environment. The team demonstrated how performance issues detected in tests or modeling and simulation can inform and guide real-time adjustments to system design, operational employment, and future test planning. Based on feedback from the DOT&E representatives, the research team was able to expand on the testbed and framework to integrate "threats" into the various system models and support Mission Based Risk Assessments.

### *RECOMMENDATIONS*

The team will continue to mature the framework and testbed to include more complex data sets and model types to determine best practices and inform data models and tooling interfaces for program applications. To meet the objectives of the DoD digital transformation and develop practices that field capability fit-for-purpose more quickly, the T&E community requires tools and methods to complement DE practices that are transforming DoD acquisition. The Operation Safe Passage framework and testbed enhance the realization of the transformative vision for T&E.

## INTEGRATED DECISION SUPPORT KEY (IDSK)

### *RESEARCH OBJECTIVES*

Building upon the base year Integrated Decision Support Key (IDSK) research that yielded policy guidance and an exemplar R-Shiny application of an IDSK, the research team's objectives for this option year were to:

- Further mature IDSK policy in collaboration with DOT&E, develop IDSK tooling that can easily be adopted by T&E practitioners.
- Further IDSK information dashboards to enable program stakeholder decision-making.
- Establish a T&E data element framework for consistent community nomenclature and dashboard population that is reusable across the DoD.

Key research questions the team aimed to address included:

- How do T&E contributions to decision-making evolve over the acquisition and sustainment lifecycle of a system?
- How can T&E practitioners re-imagine the use of current T&E processes and artifacts to enhance their contributions to end-to-end decision-making through the use of an IDSK?
- How can IDSK development and a standard IDSK data construct to account for early and iterative model development and Modeling and Simulation (M&S) data?

### *METHODS*

The research team pursued various efforts to improve understanding of the IDSK. Activities included leading and participating in greater DoD T&E community workshops to refine an understanding of MB TEMPs and the role of the IDSK within the MB TEMP. In these same engagements, the research team developed minimal viable products that can be utilized by program offices to meet DoD policy guidance. The AIRC team led a workshop that resulted in the development of an architecture and data table construct that allows for the maximum reuse of T&E data and tools across the DoD (Figure 3). The results of this effort were published in a Naval Engineers Journal Article (Werner, Esser and Arndt).

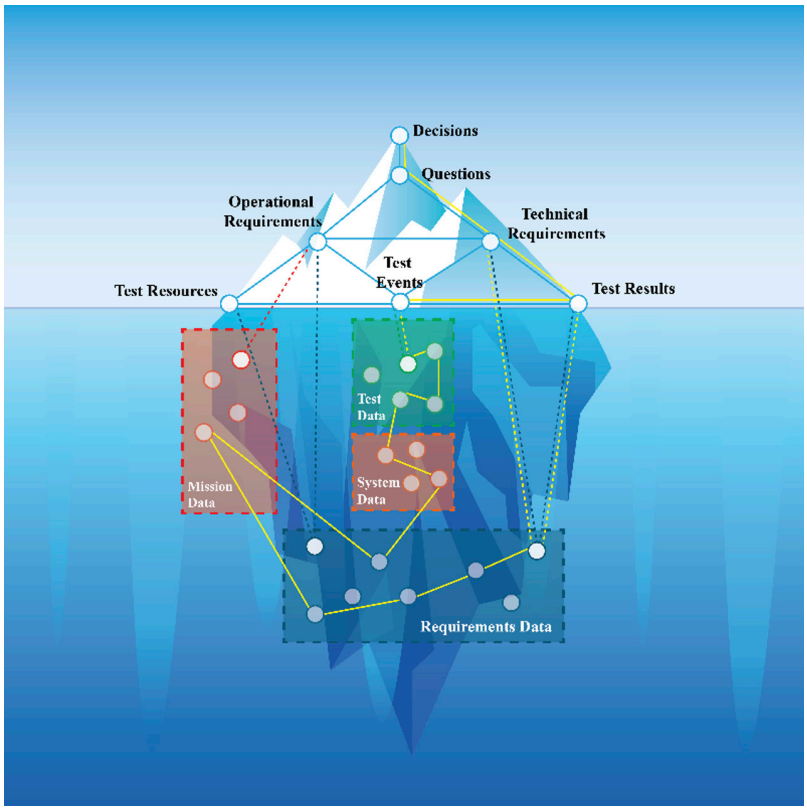


**Figure 3: The Digital IDSK, Tip of the Iceberg of the T&E Environment for Decision Making[3]**

The research team supported iterative review of upcoming DOT&E revised policy on the use of IDSK and how this policy should be rolled out to the T&E community. The rollout guidance will potentially continue in future research efforts. Awareness of the forthcoming policy changes shaped the research team's further development of IDSK digital tooling and the development of a hierarchical T&E data element framework to enable consistent reuse of data for decision-making across the DoD. To develop the framework, the researchers explored the full lifecycle of DoD systems to understand the key elements affecting traditional decision-making by key program stakeholders.

Other parts of the research team leveraged the workshops and the data element framework to iterate on the R-Shiny IDSK tooling to improve dashboards and T&E data connectivity. Several iterations of this tooling were created and are available for demonstration.

[3] We thank Dr. William Fisher of the MITRE Corporation, whose thought leadership and expertise in systems engineering contributed to the create of the IDSK "iceberg" presented in the report.

### *RESULTS/FINDINGS*

As noted above, there were several key results from this research effort. The team has provided updated IDSK policy guidance, which will be incorporated in upcoming DoD policy updates. The updated policy defines and supports a value proposition for the IDSK in key decision areas.

The team also established a T&E data element framework that is available for adoption for decision-making across the acquisition lifecycle for DoD systems. This framework is useful for data sharing and for integration with digital tooling for managing program execution and deployment, especially at major decision milestones.

Revised digital tooling that the team has provided enables programmatic decision-makers to query T&E data sources for informed decisions on operational suitability, lethality, survivability, and effectiveness. This tooling meets the minimum viable product constraints established by the T&E working groups and has demonstrated that dashboards can be used by program decision-makers who do not necessarily understand the technical design of the tools. The digital IDSK tooling has been integrated with the AIRC team's MB TEMP tooling as part of the Operation Safe Passage section of this report.

Finally, the AIRC team's involvement in the various model-based T&E communities and working groups has contributed significantly to the broader understanding and adoption of integrated decision tools to make informed programmatic decisions.

### *RECOMMENDATIONS*

Going forward, the research team recommends the following efforts:

- Further mature the IDSK tooling by integrating actual DoD program data and using dashboard visualization tools such as Tableau. The research team is already investigating the use of the Littoral Combat Ship development and test data and showing how the IDSK could have been utilized by program stakeholders/decision makers.
- Introduce a mission-based risk assessment as part of the IDSK to aid decision-makers in understanding various data-driven risks in the programmatic decision space. More broadly, the research team will investigate how uncertainty can be reflected and presented to decision-makers as part of the IDSK framework, from characterizing uncertainty in system estimates after the test to projecting the evolution of uncertainty during the acquisition life cycle to other uncertainty metrics that might be useful for decision-making.
- Assist DOT&E by maturing policy rollout products and presentation materials to help T&E practitioners adopt the use of a digital IDSK.
- Continue to advance IDSK understanding and adoption by participation in ongoing working groups, test conferences, and program assist engagements.

## VERIFICATION, VALIDATION, AND UNCERTAINTY QUANTIFICATION (VVUQ)

### *RESEARCH OBJECTIVES*

The research team conducted a study of a representative, generic, hypersonic weapon Digital Twin (DT) implemented in Simulink® to further understand how uncertainty from component models combines to affect the full DT system uncertainty. Note that the representative DT is not a true DT by definition (Banerjee, Chakravarthy and Fisher) as it lacks a physical counterpart – it was leveraged by the research as a demonstration tool. The research defined and implemented a workflow to perform an Uncertainty Quantification (UQ) study with a Simulink® model. Given a workflow, data can be collected and studied to perform UQ analysis.

### *METHODS*

The research team performed a literature search to assess the state of the art of UQ applied specifically to hierarchical or system-of-system complex models, such as DT. Topics that appeared frequently in the search included: addressing sparse data, mostly using multi-fidelity modeling, Bayesian analysis to integrate data into models, and graph theoretic considerations related to reliability theory. Reliability analysis (e.g., fault trees) stood out as a promisingly related application area as the Simulink® implementation of a digital twin evokes the graphical structure of a system-of-systems.

A representative model of a DT for a hypersonic missile system based on the Defense Advanced Research Projects Agency (DARPA)/Air Force Hypersonic Technology Vehicle (HTV) program was created in Simulink® under a related research task. This model was analyzed in terms of its member components (sub-models) and critical input data and output data were identified. Sandia's Dakota software (Adams, Bohnhoff and Dalbey) was selected for demonstrating the UQ analysis because it is fairly mature, up-to-date, open source, and well-documented. The constructed workflow generates a compatible interface that allows the DT to run repeatedly in a loop over parameter values.

UQ studies were performed to demonstrate the types of analysis enabled by interfacing Dakota with a Simulink® DT. The variables considered in the uncertainty analysis were initial glide altitude, vehicle body mass, vehicle ballistic coefficient (with mass factored out), and lift-to-drag ratio. All uncertain input parameters were endowed with uniform distributions informed to the extent possible by literature. The Quantities of Interest (QOI) for the hypersonic DT were taken to be:

- Terminal velocity,
- Terminal flight angle relative to local horizontal, and
- Total glide range.

The research team used Dakota to compute Sobol' indices (Sobol) for a global sensitivity analysis. Studies were performed with quadrature orders up to five to confirm the convergence of the results.

This work was presented at the American Society of Mechanical Engineers (ASME) VVUQ Symposium in May 2024 held in College Station, Texas.

### *RESULTS/FINDINGS*

- The Simulink®/Dakota workflow involved building a standalone executable from a Matlab® function that takes parameter arguments as input and formats the output response quantities into a Dakota-compatible text file. Template files were set up to inform the Dakota preprocessor what inputs and outputs to consume. A batch script (Windows) that runs the Dakota preprocessor followed by the packaged executable was invoked within the Dakota study input file.

- Sensitivity analysis on the hypersonic DT QOI showed that body mass and lift-to-drag ratio were the only significant contributors to the QOI uncertainty of the four uncertain parameters in Figure 4 (below). This should be verified with more thorough analysis. It was discovered that polynomial chaos expansions (PCE) (Xiu and Karniadakis) were necessary to achieve valid estimates of the Sobol' indices.

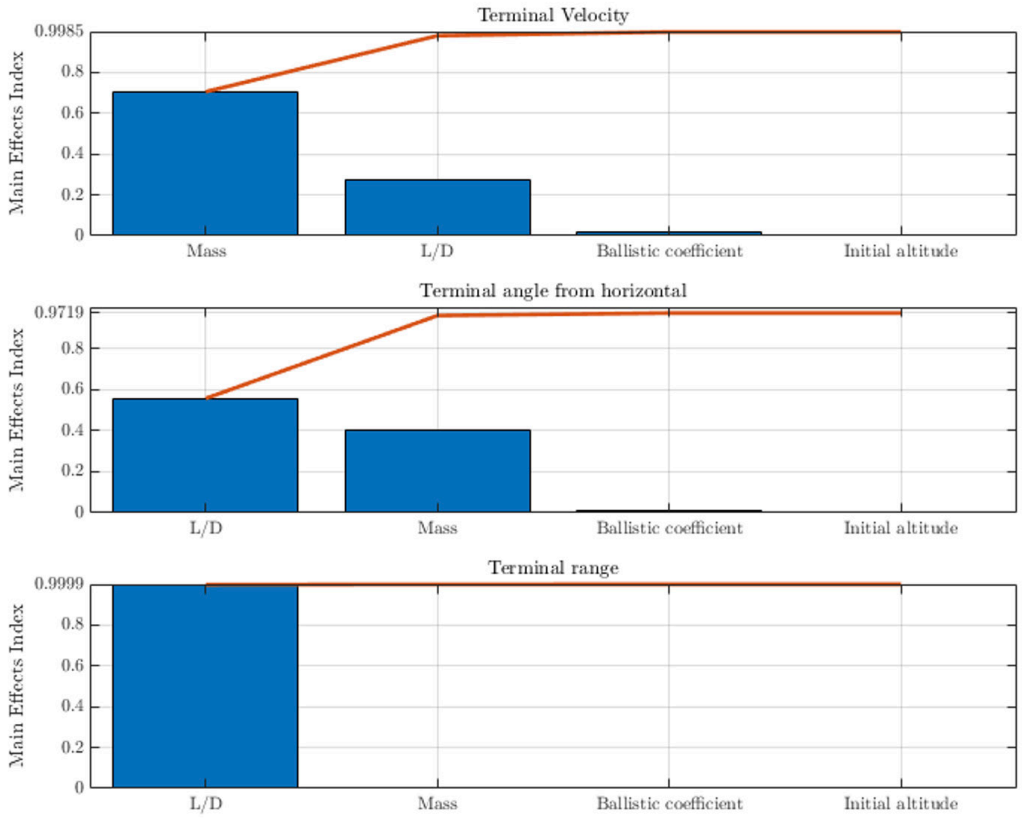


**Figure 4: Sobol' Main Effects indices for QOI**

- The fate of the HTV program inspired a demonstration quantitative risk assessment case study wherein the team enumerated "failure" scenarios and estimated likelihoods to answer the question, "How many physical systems do we need to build in order to achieve an 80% chance of completing a successful test mission?"

### *RECOMMENDATIONS*

Future research should include component model optimization, as well as a methodology for conducting quantitative risk analysis using information provided by the UQ toolchain. Furthermore:

- Dakota provides a useful tool for UQ, but Dakota documentation is not fully up-to-date and a newer python interface to Dakota may eliminate the need to create template files for the interface. Updates to the documentation would enable larger scale application of the tools in Dakota.
- Continue to develop the DT representative model to capture more sub-models and sources of uncertainty.
- Fundamental research is still needed to advance UQ for digital twins. A significant question in UQ continues to be how to assign probability distributions to uncertain parameters, and how to balance this against sparse data availability. Another open research question is how to perform UQ when the number of uncertain parameters is large (curse of dimensionality). It is also important to consider the interactions of uncertain quantities – standard UQ assumes parameters are statistically independent random variables. Finally, further exploration is needed to understand the application of Shapley value to identification of significant sub-model contributors to system uncertainty. A relationship has been described in UQ literature between Shapley value and Sobol' indices (Owen).
- This particular DT example would be further enhanced by perform the analysis using PCE to estimate probability of failure scenarios to fill out the proposed risk assessment and present risk curves for the DARPA/Air Force HTV program as a demonstration of how UQ can inform programmatic decision making.

# PILLAR 4 – PIONEER T&E OF WEAPON SYSTEMS BUILT TO CHANGE OVER TIME

## DIGITAL TWINS

### *RESEARCH OBJECTIVE*

The objective of this research was to construct a representative digital twin of a generalized hypersonic missile weapon system based on parameters extracted from open-source literature with the constraint of omitting proprietary data or classified features. Additionally, the model was intended to be based on a flexible and scalable architecture to allow for evolving capability and increasing levels of fidelity over time.

This research aimed to leverage Design of Experiments (DoE) approach for planning, conducting, and analyzing a weapon system test for a digital twin. The goal was to explore quantitative mission-focused measures across M&S and "real world" test events in the context of a mission-oriented evaluation.

### *METHODS*

A model of a boost-glide hypersonic missile was created in MATLAB Simulink® based on the work of Tracy and Wright, 2020 (Tracy and Wright). This model represents the HTV-2 system developed and flown by DARPA in the early 2010s (Acton). The model is representative of the type of performance model expected as part of a digital twin but lacks the data paths between the digital representation and the actual physical system per the definition of digital twin of (Banerjee, Chakravarthy and Fisher). In their definition, the digital representation of the system both evolves from the live infusion of data describing the physical system and also informs the evolution of the physical system by feeding its developers crucial information. Given the constraints of the current effort, the team settled on a model that could be envisioned as part of a digital twin. The model incorporates a proven architecture based on prior work on other systems including torpedoes and Unmanned Undersea Vehicles. The model is decomposed into representations of key subsystems of varying fidelity and is extensible to allow further subsystem model refinement and model decomposition. Additionally, the model included equations of motion for the glide phase of the mission as well as computations of the thermal conditions expected during operation. The model was extended to allow expression of vehicle motion in geodetic terms and standard aerospace flight parameters (e.g., latitude, longitude, roll, pitch, yaw, etc.). Model parameters were exposed for DoE exploration through input files and output data was analyzed using plotting routines.

### *RESULTS/FINDINGS*

- The Simulink® implementation developed for this study was shown to faithfully recreate the results published in (Tracy and Wright). Flight test data is not available for the system of interest so ensuring the model results were similar to those of the parent paper-built confidence in the implementation of the equations of motion in the model architecture chosen for the study.
- The model developed for this effort was successfully integrated into an Uncertainty Quantification workflow as described under the 2nd Pillar portion of this report, which describes Verification, Validation and Uncertainty Quantification in Modeling and Simulation.
- Analyses of model predictions for a range of flight conditions that were not explored in the parent paper (Tracy and Wright) showed that the model was well-behaved and was representative of how a model would be incorporated into a DoE-based preparation for system test.
- While the (Tracy and Wright) model captured the basic physics of glide vehicle flight and afforded the creation of a model that is based on open literature, this study was not included in many design details and was not included in open literature test data. This reality is likely common to all hypersonic systems given the lack of non-defense applications at present.

### *RECOMMENDATIONS*

The DoE for a system using a combination of M&S and real test data necessitates a balance between two fundamentally different activities: collecting test data with an actual physical system combined with developing and evaluating numerical approximations of the systems in their operating environment. Follow-on efforts should explore a framework where simple, yet comparatively exquisite models emulate experimentation while an intentionally rudimentary model of tunable fidelity represents M&S in exploring DoE. The goal of this future effort would be to quantify the relative return on investment for model refinement in an effort to identify a knee in the curve of model fidelity as a function of system assessment.

## UNCERTAINTY QUANTIFICATION FOR DIGITAL TWINS

### *RESEARCH OBJECTIVES*

The uncertainty quantification (UQ) for digital twins research focused on the different model categories along the mission engineering and digital engineering chain. The research objectives were:

- To develop best practices for propagating uncertainties within a digital twin.
- To aid in better operational assessments and to help shape test planning by closing the gaps on performance/capability assessments.

### *METHODS*

The research team employed Banerjee's definition of a digital twin, which is "a digital representation of a specific real-world system of interest that bi-directionally update each other at a frequency and fidelity befitting the use case" (Banerjee, Chakravarthy and Fisher).

The team developed best practices for propagating uncertainties within a digital twin. The practices were identified through an extensive literature review of the different models that make up the mission engineering and digital engineering chain. This literature review outlines the general model categories, providing background, use cases, and examples for each model category. Next the paper presents current UQ approaches and how these techniques are applied to each model category discussed. Any interfaces and links between the different model categories are presented with the underlying methods and techniques that can be used to propagate uncertainties across the chain. Finally, we provide updated procedures and model and test refinements.

### *RESULTS/FINDINGS*

The main model categories defined are empirical or data-driven models, physics-based models, MBSE, and LVC simulations. Each of these categories have their place within the acquisition cycle but overloaded terms like "models" and "modeling and simulation" need additional specificity, especially when these different modeling categories get coupled together. Example cases include data-driven models or physics-based models to MBSE or LVC simulations. Since the development of a digital twin is a multidisciplinary approach, the process requires proper communication across disciplines and among decision-makers for proper "bookkeeping" to take place. Proper communication enables tracing where or which models the uncertainties came from, identifies where refinement needs to happen, and illuminates where to focus further system testing.

Incorporating UQ into the digital twin environment is vital for accurately capturing the range of possible outcomes and ensuring that the models faithfully represent real-world systems. UQ plays a critical role in identifying, characterizing, and propagating uncertainties across complex models, which is key to enhancing model reliability, reducing reliance on physical testing, and ultimately improving decision-making in mission and digital engineering.

UQ faces several significant challenges in the context of mission engineering and digital engineering, namely: model complexity, non-linear interactions, high-dimensionality, correlated uncertainties between system components, model heterogeneity, sparse or noisy data, and many more. Various UQ techniques can be employed to address these challenges. Table 3 (below) provides suggested UQ techniques that may be effective in addressing the challenges described above.

**Table 3: Non-exhaustive summary of suggested UQ techniques that can be applied to address specific UQ challenges arising in complex systems analysis (Cortes, Wong and Cortes-Morales).**

| Challenge Category | Suggested UQ Technique |
|---|---|
| Model Complexity | • Surrogate Modeling (e.g., Gaussian Process Regression)<br>• Dimensionality Reduction Techniques (e.g., Principal Component Analysis) |
| Non-linear Interactions | • Polynomial Chaos Expansion<br>• Monte Carlo Simulation |
| Dimensionality | • Sparse Grid Quadrature<br>• Stochastic Collocation Methods |
| Correlations | • Copula Methods<br>• Canonical Correlation Analysis |
| Discrepancy between Models and Reality | • Bayesian Calibration<br>• Error Modeling |
| Scalability | • High-Performance Computing<br>• Parallel Computing Methods |
| Heterogeneity of Models | • Multi-fidelity Modeling<br>• Ensemble Modeling |
| Sparse or Noisy Data | • Bayesian Data Analysis<br>• Maximum Likelihood Estimation |

Integrating UQ into the digital twin framework enhances the reliability of simulation models and facilitates more informed decision-making by providing a clearer understanding of the uncertainties involved. The reliance on physical testing can be significantly reduced by addressing the outlined challenges through the strategic use of advanced UQ techniques thereby accelerating development cycles and increasing the likelihood of systems meeting performance requirements upon deployment.

### *RECOMMENDATIONS*

The resulting best practices document (Kauffman, Higgins and Gilbert) highlights the functional form of the interfaces and links between model categories and how (generic) model outcomes or performance parameter distributions get mapped, bi-directionally, and tracked from one piece of the mission engineering chain to the next. It is also necessary to highlight any tools, methods, or techniques used to map and track uncertainties between model categories. Lastly, the document defines procedures in which (generic) model or test data are injected into the chain for refinement and where that data gets injected. The guidance developed in (Kauffman, Higgins and Gilbert) can be used to mature the best practices methodologies, tools, and techniques for the propagation of uncertainties within digital twins. Furthermore, these approaches can be demonstrated through the implementation of digital twins for model integration.

## T&E FOR AI/ML BEST PRACTICES GUIDE

### *RESEARCH OBJECTIVES*

The rapid advancement in machine learning (ML) and widespread adoption of ML-enabled software systems have made it increasingly critical to ensure the quality of these systems so that they can be deployed and operated with confidence. The primary objective of this research was to study the current best practices in the T&E of ML-enabled software systems. While the ML model is considered the core component of a ML-enabled software system, it is just one of the many elements of a larger software system. Therefore, to harness the potential of a ML-enabled software system, it is necessary to guarantee that the ML-enabled software system exhibits the desired performance across its lifecycle. Thus, the project team took a holistic view of the current T&E best practices, challenges, and gaps across the lifecycle of a ML-enabled software system.

### *METHODS*

To achieve the research objective of gaining a comprehensive understanding of the current T&E practices across the ML system's lifecycle, the study divided the lifecycle of a ML-enabled system into three phases:

- *Component phase:* the ML model is the core component of a ML-enabled software system. This phase focused on T&E activities during ML model development. The research team examined the current methods and approaches for test generation, evaluation (metrics), and adequacy in ML model T&E.
- *Integration and deployment phase:* this phase explored the T&E practices and challenges for integrating ML model(s) into the broader software system and deploying the ML-enabled software system into production environments. Specifically, the research investigated the T&E challenges in model compression (reducing model size for deployment) and model conversion (facilitating cross-platform interoperability) activities.
- *Post-deployment phase:* the study investigated key T&E challenges in maintaining the performance of operationalized ML-enabled software systems and explored existing T&E practices for monitoring, maintaining, and re-engineering ML-enabled software systems.

### *RESULTS/FINDINGS*

A comprehensive literature review was conducted to identify T&E practices across the three phases. The review revealed the following key findings:

*Component phase*

- Practitioners have successfully adapted traditional testing methods to the T&E of ML models. The study found that traditional test methods such as metamorphic testing, differential testing, combinatorial testing, and fuzz testing are some of the widely employed test methods in the T&E of ML models. While practitioners have successfully adapted existing test methods in testing ML models, the generation of valid and meaningful test inputs remains a challenge.

- Unlike traditional software, the overall performance of a ML model on a test suite is an indicator of its quality rather than the outcome of individual test cases. Metrics for assessing model performance vary based on the model type. Practitioners often rely on accuracy, precision, recall, and F1-score for classification models. In the case of regression models, practitioners use metrics like root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE). However, one of the key limitations is that the current set of metrics may not adequately capture the model's learning, rather overemphasizing memorization capabilities. Furthermore, they do not adequately evaluate model capabilities based on the domain or environment in which it will be operated.

- Traditional structural coverage adequacy metrics (e.g., branch coverage) are not applicable to ML models. While neuron coverage has been proposed as an adequacy criterion for deep neural networks, data-centric adequacy metrics like surprise adequacy and combinatorial coverage were also explored to assess test suite adequacy in ML models. However, test adequacy for ML models is nascent and needs further exploration.

*Integration and deployment phase*

- Model compression and model conversion are the two most common techniques practitioners use to integrate and deploy ML-enabled software systems. While variations in the ML system's performance post-conversion or post-compression are expected, distinguishing between acceptable deviations and critical errors remains challenging. Model compression or model conversion processes can introduce defects leading to non-failure performance degradation such as increased inference time or resource consumption, which are challenging to identify through traditional T&E methods.

- T&E research in this phase is notably limited compared to the component level. Addressing these challenges requires a systematic approach to T&E, including developing new methods and metrics. Future research should also investigate the impact of model conversion or model compression on critical AI quality attributes like explainability, fairness, and security.

*Post-deployment phase*

- Unlike traditional software systems, maintenance activities in ML-enabled systems span across its lifecycle. Maintenance of ML-enabled systems is complex due to factors such as versioning challenges, evolving behavior, and the need for frequent updates. The evolving behavior – the ability of an ML-enabled software system to learn and adapt over time once deployed in the operational environment – highlights the need for live monitoring of the ML-enabled system's performance to detect and mitigate unanticipated failures of ML-enabled systems.
- Another significant challenge arises from the rapid evolution of ML libraries and frameworks, leading to versioning-related issues. These incompatibilities can introduce bugs, increase maintenance costs, and disrupt system stability. These challenges highlight the need for proactive T&E methods to detect and mitigate the impact of drifts, standardized re-engineering processes, and effective regression testing approaches to ensure the operational reliability of ML-enabled systems and to prevent any undesirable behavior or harmful failures.

### *RECOMMENDATIONS*

*Component phase*

- Develop novel metrics to comprehensively assess the learning capabilities of ML models.
- Create application-centric metrics to accurately evaluate the operational performance, safety, and reliability of ML-enabled systems on which they will be operated upon on (real-world environments).
- Test adequacy measurements for ML systems remain underexplored and require the exploration of new approaches.

*Integration and deployment phase*

- Expand research on T&E methodologies to systematically evaluate the performance of ML-enabled software systems in the integration and deployment phase including developing new test generation methods and metrics to address the T&E challenges in this phase.
- Develop test frameworks to evaluate the impact of transformed ML models (compressed or converted ML models) on AI assurance characteristics like explainability, fairness, and security.

*Post-deployment phase*

- Develop proactive maintenance practices to ensure the operational reliability and safety of ML-enabled systems.
- Standardize re-engineering processes to improve efficiency.
- Implement effective regression testing strategies during the re-engineering phase.

## SYSTEMATIC INCLUSION/EXCLUSION (SIE) RESEARCH FRAMEWORK

### *RESEARCH OBJECTIVES*

Specifying the conditions under which a ML model was trained is crucial to defining the operating envelope which, in turn, is important for understanding where the model has known and unknown performance. Metrics such as combinatorial coverage applied over (metadata) features provide a mechanism for defining the envelope for computer vision algorithms, but not all (metadata) features impact performance. Systematic inclusion/exclusion (SIE) is an experimental framework that draws on practices from design of experiments and combinatorial interaction testing to identify the critical (metadata) features that define the dimensions of the operating envelope. SIE supports the "reliable" ethical principle; defining the dimensions of a model's operating envelope is critical for achieving "well-defined uses."

### *METHODS*

The SIE framework performs the following five steps:

1. *Build a universal test set*: create a universal test set that covers each interaction approximately the same number of times. For each interaction, split the universal test set into two subsets: set with all samples containing the interaction and complement.

2. *Build a collection of training sets*: for each interaction, generate a training dataset where the interaction is systematically excluded. Generate one baseline dataset that excludes nothing. All training sets have the same number of samples; when there are more candidate samples than required, down-select randomly.

3. *Build a collection of models*: for each interaction, train a model on the training set that excludes that interaction. Use the same hyperparameters and architecture for all.

4. *Evaluate the models*: for each interaction, test the model on both test sets for that interaction. For baseline, test the model on universal test.

5. *Evaluate the results*: statistical analysis of contrasts compares whether being excluded from training and whether a particular combination being excluded has an impact on model performance.

SIE utilizes coverage metrics implemented in the Coverage of Data Explorer (CODEX) Python package. The SIE code in CODEX previously performed steps 1 and 2. In this effort, the team developed code for a rudimentary test harness to conduct automatic training and testing of models for steps 3 and 4. The statistical analysis was previously performed in JMP statistical software and efforts to replicate it using existing Python statistics packages failed as Python has historically lagged behind tools like JMP and special purpose languages like R. In this effort, the team first replicated the behavior of JMP in the R programming language and then wrote custom code to conduct validated analysis in Python, which resulted in an end-to-end SIE implementation within CODEX.

The SIE framework was developed for Project Maven and demonstrated on restricted data for identifying critical factors in synthetic aperture radar (SAR). The experiment needed to be replicated on open data for publication and dissemination to the broader community. The RarePlanes dataset was selected because it includes several types of metadata.

### *RESULTS/FINDINGS*

The research team identified five metadata factors for experimentation in the RarePlanes dataset – average pan resolution, biome, hour of day, off nadir max, and season – and performed the experiment on the real (non-synthetic) data. The team also included a randomly generated control variable. The team trained 31 You only look once (Yolo)V8(nano) object detection models, excluding one metadata factor level from each. ML models were evaluated for precision, recall, and F1. Two logistic regression models were fit over the ML model outputs: one for the metadata factors and one for the control factor. Inclusion of the metadata factors was found to be significant while inclusion/exclusion of the control factor was not. Inclusion/exclusion of average pan resolution, biome, and hour of day were found to be significant for all three performance metrics while season was found to be significant only for the recall metric.

### *RECOMMENDATIONS*

The algorithms developed for building balanced test and training sets should be improved. Additionally, the framework uses an experimental design that withholds one potential critical factor (combinatorial interaction of features) at a time to build a model. As the number of (metadata) features becomes large or the size of feature interaction considered increases beyond one-way interactions, the number of models required becomes intractable. Aliasing is a concept from DoE for creating test sets where a factor or interaction appears frequently or always with another, making it difficult or impossible to distinguish effects between the aliased factors and interactions. On the other hand, aliasing reduces the size of the constructed test set. Some designs from combinatorial testing, specifically covering arrays, allow aliasing and generally result in smaller test sets as the number of rows in a covering array grows logarithmically in the number of factors to enable identification of critical factors in larger spaces. DoE or combinatorial testing may be useful for constructing a design over the factors withheld as the number of factors or size of interaction increases.

## HIERARCHICAL SCORING

### *RESEARCH OBJECTIVES*

One of the primary uses of AI models is classification or predicting the class of a sample. Object detection is an extension of classification that includes localization of the object via a bounding box within the sample. Classification (and by extension object detection) is typically evaluated by counting a prediction as incorrect if the predicted label does not match the ground truth label. This pass/fail scoring treats all misclassifications as equivalent. In many cases, class labels can be organized into a class taxonomy with a hierarchical structure to either reflect relationships among the data or operator valuation of misclassifications. When such a hierarchical structure exists, hierarchical scoring algorithms can be developed that return the model performance of a given prediction related to the distance between the prediction and the ground truth label. These hierarchical scoring algorithms can be viewed as giving partial credit to predictions instead of pass/fail. The hierarchical scoring approach enables a finer-grained understanding of the impact of misclassifications, enabling the ethical principle of reliable AI.

The research team previously developed hierarchical scoring algorithms that utilize a scoring tree to encode these relationships between class labels and produce metrics that reflect distance in the scoring tree. Five hierarchical scoring algorithms and three types of scoring trees have been developed, and combinations of scoring algorithms and trees have been demonstrated on abstract examples. The experiment results need to be documented and disseminated to the academic community along with a thorough literature review.

### *METHODS*

The team conducted a literature review and is writing an academic paper titled *Hierarchical Scoring for Evaluating Machine Learning*. To demonstrate and compare the algorithms' behavior, Python code executes the various algorithms on a scoring tree and a batch of inputs. All scoring algorithms take in a prediction and a ground truth label as nodes in the tree and produce a score between 0 and 1 with higher scores indicating preferred performance. The scoring algorithms range from simple to complex:

1. *Path length (PL)*: computing the number of edges between nodes,
2. *Lowest common ancestor (L):* computing the reward edge weight up to the lowest common ancestor,
3. *Lowest common ancestor with path penalty (LPP):* adding the penalty in edge weights from the lowest common ancestor to both nodes to the reward, and
4. & 5. *Lowest common ancestor true path standardization (LTPS)* and *Lowest common ancestor with path penalty true path standardization (LPPTPS):* standardizing the computation (reward only or reward plus penalty) given the edge weight to the truth, which is important when predictions or ground truth labels may not always reside at the leaves.

The edge weights in the scoring trees encode distance between the class labels which are represented as nodes. For the PL scoring algorithm, all weights are 1; for all other algorithms, continuous-valued weights imply an infinite number of possible scoring trees. To demonstrate the impact of scoring tree on how the algorithms differentiate models, three general weighting schemes were chosen to produce three representative scoring trees. The weighting schemes describe how weights change from root to leaf and include decreasing, increasing, and non-increasing (being balanced where possible).

To understand how the hierarchical scoring algorithms rank models that make different types of misclassification errors, four abstract models were characterized by behavior: 1) as correct as possible (predicting the same node as the ground truth), 2) as incorrect as possible (predicting the node furthest in the scoring tree from the ground truth), 3) conservative (predicting a node at or above the ground truth), and 4) aggressive (predicting a node at or below the ground truth). Given the same set of ground truth labels, synthetic outputs were generated for each of the abstract models' predictions. The pairs of predictions and ground truth labels were used to "stub" an AI classifier to evaluate how the five scoring algorithms would rank the performance of the four models and the impact of the three weighting schemes on the ranking.

### *RESULTS/FINDINGS*

The simpler scoring algorithms fail to capture granularity of misclassification sufficiently. The path length (PL) algorithm slightly favors conservativism but is not capable of expressing more complicated preferences. The lowest common ancestor (L) and lowest common ancestor with true path standardization (LTPS) algorithm rewards stopping higher in the tree (conservative over aggressive). The weighting scheme greatly affects the overall scores; scores tend to be much higher when higher order paths have higher weights, and the preference for aggression over conservatism correlates with increasing weights. The algorithm that adds a path penalty (LPP) and the version with true path standardization (LPPTPS) are both very sensitive to the weighting scheme. The decreasing weights scoring tree prefers the aggressive model that predicts at or below the ground truth as the penalty for going further down the tree past the ground truth is reduced as the weights decrease. The increasing weights scoring tree prefers the conservative model as the penalty for predicting lower than the ground truth is increased with the increasing weights, and the non-increasing weight scheme yields almost the same scores for aggressive and conservative classifiers.

### *RECOMMENDATIONS*

This work demonstrates an approach to evaluating model performance that can rank models not only by how many errors are made but by the degree or impact of errors. To ensure this work is useful for operational testing, metrics should be validated on real data sets that have hierarchical structures, ideally on real AI-enabled systems (AIES). Most classifier architectures assume a flat class hierarchy and are trained without knowledge of the class structure even when it is hierarchical. However, classifiers that are trained to be "hierarchy aware" also exist. Experimentation is needed to determine if the hierarchical scoring metrics distinguish hierarchy-aware vs. hierarchy-agnostic classifiers. Additionally, we hypothesize that hierarchical scoring may be used as part of a loss function to produce models with the preferred behavior. Future work is needed to determine if implementation is feasible and compare performance against other hierarchy-aware classifiers.

## COGNITIVE ELECTRONIC WARFARE (COGEW)

### *RESEARCH OBJECTIVES*

The introduction of cognition i.e., AI and ML driven capabilities, into Electromagnetic Warfare (EW) systems will necessitate changes in acquisition, development, T&E, and risk management. AI/ML-driven cognition will introduce new attributes to EW systems, such as learning, autonomy, and higher sensitivity to electromagnetic effects that will challenge the existing T&E practices in EW. The Test Resource Facility (TRF) infrastructure, which is heavily involved in verifying and validating EW systems, will be significantly impacted by these new EW capabilities. The new challenges that Cognitive EW (CogEW) systems present will force TRFs to re-evaluate their physical infrastructure, rethink how they test EW systems, and potentially alter their business models. This change will likely be a substantial task and TRFs will need support to make their evolution as painless as possible.

Support is necessary because TRFs are critical to the path of deployment for EW systems, and failure to evolve before CogEW systems reach their stage of production may severely delay the safe adoption of these new capabilities. Public law requires that every EW system go through multiple categories of test resource facilities before they are allowed to enter initial production. This mandate ensures that EW systems utilize the nation's well-established resources for system integration, exposure to validated environmental effects and threats, and the execution of credible testing at the fidelity appropriate for their pre-production stage. However, these are all features that will be pushed beyond their limits by CogEW systems that are complex, learn, change, and are highly sensitive to integration and radio frequency (RF) environment effects. This research aims to support the TRF community by characterizing these new challenges and facilitating academic research to overcome them, particularly in the test process space. Research support will be critical in order to enable and facilitate the adoption and safe deployment of CogEW systems onto the battlefield at the speed of relevance.

### *METHODS*

The team's research methods consisted of a literature review and community engagement. In collaboration with the Test and Evaluation Threat Resource Activity (TETRA), the team was able to piggyback onto problem domain discussions and interviews with the following test resource stakeholders from around the country: Electronic Combat Simulation and Evaluation Laboratory, Joint Preflight Integration of Munitions and Electronic Systems Facility, The National Cyber Test Range, Guided Weapons Evaluation Facility, Integrated Defensive Avionics Laboratory, Electronic Warfare Avionics Integration Support Facility, and the Benefield Anechoic Chamber. Additionally, the team had more informal discussions with stakeholders from three programs of record, Test Resource Management Center (TRMC), two intelligence organizations, and IDA.

These engagements were the primary drivers for the research findings. The team looks to academia to further substantiate and characterize the threads that were gleaned from the community discussions and investigate any possible recommendations for further research and development.

### *RESULTS/FINDINGS*

Some CogEW developers do not have confidence in the TRF infrastructure's ability to subject their systems to a sufficiently representative environment for accurate and credible T&E. All three of the programs of record we talked to are in the process of or have already pushed their systems into some form of flight testing for evaluation. Flight testing, while sufficiently representative, will not subject systems to a level of rigor TRFs can provide. To earn stakeholder trust and credibly, the following challenges must be addressed:

- Scale
- Accessibility/Availability
- Efficient Test Design and Execution

The average TRF must be able to scale up to a more representative environment in terms of density, variety, and reactivity. Representativeness is currently distributed across the TRF categories. More environmental effects need to be uniformly distributed across the infrastructure. The further left in a product's lifecycle, the better. Optionally, research into identifying minimally viable levels of fidelity in EW test environments would also be practical and cheaper.

TRFs must become more accessible and, ideally, more available for persistent testing. This shift is necessary for the new agile needs and requirements that rely on evolving testing and training. Currently, facilities are booked year-round, leaving limited bandwidth for follow-up T&E when new issues are found or the system changes. The inefficiencies driving this situation are numerous, ranging from high integration time to limited automation. Implementing data, interface, and other test standards would help streamline the use of TRFs within the EW community. Early integration of TRFs into the development cycle is also beneficial.

Finally, TRFs must be able to plan and execute efficient tests that get the most informative data out of the limited runs typically completed. Most test activities can only be done across a few weeks at great expense and for just a couple thousand runs. With CogEW systems being more complex than traditional EW systems, new test design and execution techniques are required to characterize these systems in an iteration-constrained environment. Adaptive experimental design would be an effective tool that balances familiarity with the new characterization needs of CogEW systems. With an adaptive design framework, sparse designs can be enhanced by intelligent sampling driven by the responses from the system under test (SUT). New test points are added or removed based on how they impact the SUT. The higher the impact, which can be in terms of variance or error, the more informative the point is. This framework can complicate the testing process, but it allows stakeholders to use this process to start small and build up to a satisfactory endpoint.

### *RECOMMENDATIONS*

DOT&E should invest in further research into the level of fidelity needed to adequately and credibly test CogEW systems. The burden of creating highly representative environments in Human-In-The-Loop or even purely digital form is too great for the average TRF or Development Test stakeholder. If a minimally viable subset of environmental effects or characteristics can be identified, it can save millions in modernization expenses for TRFs and earn back confidence in the TRF infrastructure for the CogEW task.

Guidance and policy should be developed encouraging TRFs to move towards a more agile model and support TRFs and the EW community in the development of infrastructure and process consistency that promote a revolving door of systems (and algorithms) coming in, hooking in and running thousands of iterations with easy scenario tear-down and set-up where appropriate. For some TRFs (such as Installed System Test Facilities), only a little could be done, but any progress could unlock thousands of additional test hours for iteration-hungry systems and programs.

Within that agile model, adaptive test design and execution will be required until pre-test analysis techniques can produce more informative static designs. DOT&E should produce guidance for the execution of adaptive designs in TRFs. Further research is needed to identify the appropriate adaptive design parameters for the EW domain, such as sampling off of error or variance and determining when to stop. If adaptive experimental design becomes too cumbersome for TRFs, it is recommended that additional research be put into pre-test analysis for informative static designs suitable for CogEW systems. For example, this research investigated Conformal prediction frameworks as a path to pre-test analysis and test design. For parametric models, it is relatively simple to quantify uncertainty across the test space and build a balanced, informative design. However, it is not as intuitive for integrated AI/ML-driven systems that would be tested at a TRF. Well-informed static designs would simplify and add more credit to operational and risk management test activities.

## MODEL-BASED TEST OF AI USING SYSML AND OPEN NEURAL NETWORK EXCHANGE (ONNX)

### *RESEARCH OBJECTIVES*

This research examines the feasibility of using a standard modeling language, SysML, to capture AI/ML systems from a T&E perspective. In previous work, the Open Neural Network Exchange (ONNX) was identified as a way to capture general learning systems. This effort builds upon that by 1) exploring translation of the ONNX learning system into SysML, in order to develop a system-agnostic AI/ML metamodel that captures the essential components and relationships of AI/ML systems, 2) evaluating candidate systems to effectively test the SysML/learning language integration, and 3) investigating a path forward to integrate the Cameo-based learning model with the previously proposed model-based T&E methodology, in order to facilitate efficient test case generation, execution, and analysis for an AI/ML system.

### *METHODS*

The research methodology included a thorough investigation into expressing learning systems in SysML, while simultaneously developing an AI/ML metamodel. An appropriately sized example ONNX model – only large enough to test a Python translator script (Onnx2json) and provide sufficient coverage of ONNX model elements – was selected and translated in order to determine how the ONNX elements could be represented in SysML. Initially, efforts were made to determine how to convert a learning system model from an ONNX file into an extensible markup language (XML) file. However, an open-source package (Onnx2json) that allowed conversion from ONNX files to JavaScript Object Notation (JSON) files was discovered and used to convert a handwritten digit recognition ONNX model into a JSON file and dictionary. The general learning system concepts present within the JSON were captured in a SysML library. These SysML constructs, which describe structure and behavior (i.e., Blocks and Activities), are sufficient to describe the structure and associated behavior of a learning system. The library was populated based on high-level concepts identified in ONNX (e.g., Model Layer, Activation Function, etc.). High-level concepts were specialized and further defined with SysML notation to describe subtypes of each concept (e.g., Convolutional Layer, Pooling Layer, etc.).

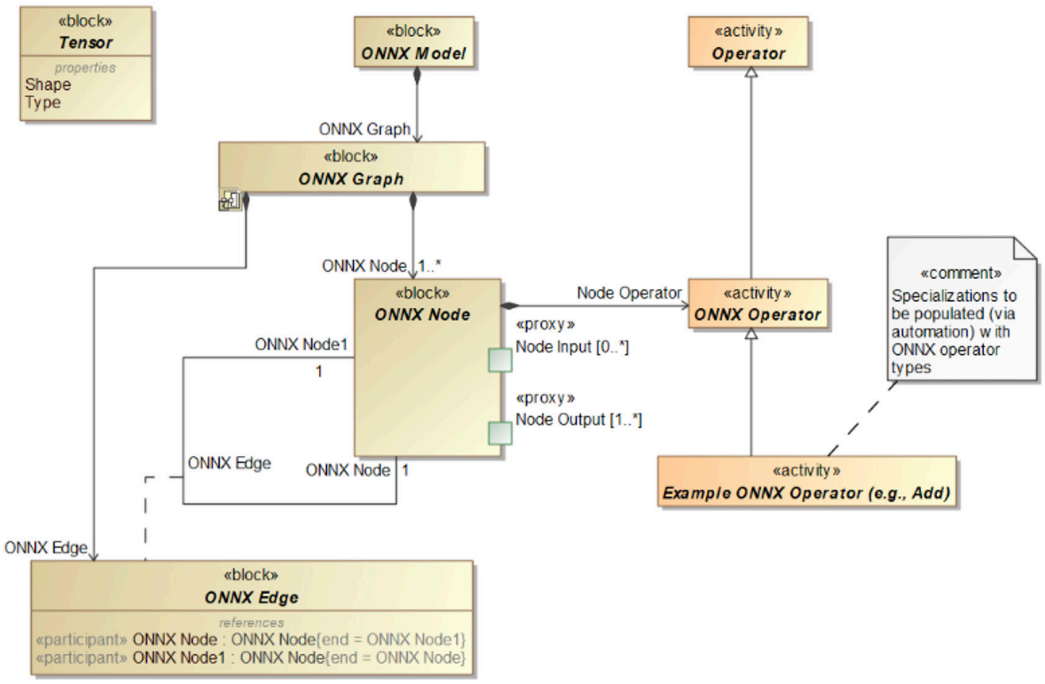


**Figure 5: High-level SysML Description of a Learning Model Captured in ONNX**

With these SysML constructs in place, development of a metamodel in the AI/ML context began. These efforts collectively established the foundation for capturing the intricacies of learning systems and integrating them with MBSE best practices using Cameo Systems Modeler.

### *RESULTS/FINDINGS*

ONNX captures essential concepts needed to produce a useful system model with SysML to support test and evaluation activities. Further, ONNX-SysML translation is shown to be a suitable starting point for applying robust test and evaluation methods to learning systems. In the future, completing a SysML library that aligns with ONNX will enable automated translation from ONNX-compatible learning system models into general representations of those systems in SysML. Developing a process to generate SysML models containing all parameters of a learning system that are relevant to test and evaluation would enable the use of previously developed model-based testing frameworks to be modified to automatically produce complete sets of test cases necessary for robust evaluation of learning systems.

### *RECOMMENDATIONS*

The recommended next steps are to finalize a SysML library that captures all of the parameters of a learning system pertinent to the test and evaluation of the learning system, evaluate candidate systems that can effectively test the SysML/learning language integration, and explore the modifications needed to the proposed T&E methodology, from this period of performance, to make the approach adaptable to an AI/ML system.

## USE CASE: T&E FOR AI/ML PROTOTYPING T&E METHODS FOR ROBUSTNESS TO REAL-WORLD SITUATIONS FOR AN OPTICAL SENSOR

### *RESEARCH OBJECTIVES*

The objective of this task was to develop a proof-of-concept for a virtual testbed that enables researchers to evaluate T&E methods for cyber-physical intelligent systems without requiring an actual, cyber-physical intelligent system.

### *METHODS*

The virtual testbed is intended to replicate behaviors of systems that exhibit any sort of intelligent behavior. Specifically, the testbed represents a system that has an intelligent component, but also has additional, system-level components that affect the input or output data of the intelligent component. In other words, the testbed is designed to manipulate the combined effect of intelligent components with unintelligent components.

The testbed is a surrogate of an intelligent detection system composed of an optical lens, a camera, and a classification system. While the intelligent algorithm is implemented as part of the classification system, the results of intelligence are a function of the performance of the lens and the camera. The models of the three components are designed to enable modifications that represent those that an actual optical detection system may exhibit in real life, including manufacturing variations (such as lens transmission or pixel defects), technology upgrades or obsolescence (such as increase of camera resolution), or operational wear (such as lens transmission deterioration or pixel cracks). Such a design allows for (1) inducing variations that a real system would exhibit and (2) variations that enable the assessment of the effectiveness of systems engineering methods for intelligent systems.

The *actual, real* image that the detection system observes is simulated by a bank of digital images. The detection system can be easily replicated with different settings in their parameters, which enables the simulation of different manufacturing and operational conditions, and hence the evaluation of different test and evaluation strategies for the intelligent system (e.g., use the same training data for all detection systems vs training each detection system individually).

The following use cases were tested:

- Manufacturing variations expressed through pixel damaging,
- Operational wear expressed through brightness reduction,
- Technological change (refresh) expressed resolution doubling, and
- Technological change (obsolescence) expressed resolution halving

The implementation of the testbed was performed on Python. The default dataset was extracted from the Kaggle repository. Data was prepared in three steps:

1. Specifying the manipulations that will be undertaken on the imagery to simulate the intelligent system.
2. Training the object detection algorithm Using the Yolov5 algorithm.
3. Testing the performance during the inference phase of the process. The training step along with the testing step constitutes the object detection algorithm component.

### *RESULTS/FINDINGS*

The study showed that changes in the physical components represented performance changes of the "intelligent component." Thus, demonstrating the testbed can emulate performance of the intelligent system beyond the performance of its intelligent component.

### *RECOMMENDATIONS*

The different parameters in the testbed must be calibrated, ideally with an actual detection system. Additional use cases may be explored to evaluate the extent of the testbed's capabilities.

## SYSTEMS THEORETIC PROCESS ANALYSIS (STPA) FOR AI ETHICS ASSESSMENT

### *RESEARCH OBJECTIVES*

In February of 2020, the DoD published an article stating that the DoD has "officially adopted a series of ethical principles (EPs) for the use of AI." Given the broad scope of the EPs, it seems that much of the assessment of the ethics of AI-enabled systems (AIES) can only occur at the level of an integrated AIES and the programs that develop and operate them. That is, component-level assessment scoped to learning algorithms or learned models will fail to address the fundamentally systems nature of assessing adequate adherence to the EPs. Therefore, given the emphasis of T&E on integrated system testing, it is especially critical to address capability gaps in operational T&E (OT&E) and the central role it will play in the implementation of the EPs.

In this research, the team considered gaps in ethics assessment methodology for the ethics of AIES. The team refers to ethics assessment (EA) as the task of using T&E to assess adherence of an AIES, its development, and its operation to the EPs, considers the role systems theoretic approaches like systems theoretic process analysis (STPA) have in EA, and considers a case study in AI disengagements, a topic closely related to the "governance" EP.

### *METHODS*

Part of the difficulty in developing rigorous EA methodologies for AIES is the idiosyncratic nature of AI, the AIES they embody, and the (evolving) operating conditions where they are deployed. The *narrow* scope of EA methodologies in the machine learning literature reflects this difficulty. Systems theory, by being highly abstracted, offers the chance for EA methodologies with a *general* scope that are, in a sense, "flexible" enough to fit the varying specifics of AI, AIES, and their operating conditions.

The basic intuition for using an STPA-like approach is that like safety (a common STPA use case), ethics is a fundamentally systems-level, not component-level concept, and therefore requires systems-level methodology. The pragmatic perspective is that systems-oriented approaches to EA can leverage systems models to make guarantees of ethical outcomes (in terms of consequences) without strong assumptions regarding the uncertain, stochastic performance of AI components (like learning algorithms and learned models).

The ability to disengage AI capabilities is a focal point of DoD's EPs. If the AI capability of an AIES cannot be disengaged, then it seems that most, if not all, ethical dilemmas have as difficult or more difficult trade-offs than when an AIES's AI capability can be disengaged. For example, if a human operator is always available to serve as a substitute for the AI capability, then decision-makers do not have to worry about mission success hinging on an AI capability that cannot be disengaged, and therefore will not have to trade-off adherence to EPs with mission success. Thus, disengagement is an important capability for AIES. The team considered a systems modeling case study in disengagement for an unmanned aerial vehicle (UAV).

### *RESULTS/FINDINGS*

The research team's STPA case study emphasized identifying adversity chains that lead to losses in terms of the UAV's mission and the DoD's EPs. Specifically, the team focused on the UAV's ability to disengage AI control safely in various scenarios, such as electronic warfare attacks, sensor deception by enemy forces, and communication failures. By mapping out these scenarios, the team identified several ways in which the system can evolve, through a series of system state transitions, to a loss. The team also identified and proposed possible mitigations that EA testers could provide as feedback to the AIES development program, including redundancy in systems, enhanced operator interfaces, and rigorous adversary simulation testing.

The results of STPA can be used in EA activities. The insights gained from using STPA can lead to the development of testable requirements for AI ethics. These requirements can be assessed as functional requirements of the integrated, AI-enabled system. Also, adversity chains and their associated state transitions can be used to identify initial states for EA scenario-based testing, for both virtual T&E and live-fire T&E. Similarly, the adversity chains can be used to identify points in the system for EA instrumentation.

This case study is a rough sketch of a loss-driven analysis. The analysis procedure can be more formalized, the scope can be expanded, and significantly more depth can be added to each consideration. While this case study was not model-based, STPA can make use of system documentation, system models, and many other system artifacts. Ultimately, the team's objective for this notional case study was to illustrate the stark differences between systems-theoretic, loss-driven approaches, and the socio-technical or algorithmic approaches of the machine learning literature in hopes to motivate further study of techniques like STPA in the field of EA.

### *RECOMMENDATIONS*

AI ethics is a broad topic. Narrowing ethics to assessment and narrowing ethics to DoD's EPs helps provide scope but the existing methods in the literature are not well-scoped to the task. While there are many frameworks that are meant to help address and smooth the gaps between available methods, there are few, if any, top-down systems methodologies. Loss-driven systems analysis, such as STPA, offers a promising approach to assessing AI ethics that is focused on consequences over likelihood estimation, focused on modeling the system over modeling the environment, and focused on integrated systems (the subject of OT&E) over individual AI components. As such, it should be explored both as a tool and basis for EA programs.

## COVERAGE OF DATA EXPLORER (CODEX)

### *RESEARCH OBJECTIVES*

The Coverage of Data Explorer (CODEX) tool is a Python package that implements data coverage metrics and algorithms for machine learning T&E applications. CODEX's metrics are based on the theory of combinatorial testing (CT) adapted from software testing to AI/ML T&E with a data-centric focus. The metrics implemented in CODEX, combinatorial coverage and set difference combinatorial coverage, compute the coverage of the universe of possible inputs to an ML model represented by datasets of samples.

As a data assurance tool, CODEX requires datasets for all functionalities. Some functionalities additionally require ML training algorithms and a mechanism for automatically training models and evaluating them on a test set (a test harness). CODEX provides functionalities such as:

- Evaluation of the coverage of a defined universe by a dataset, which may be used for applications such as selecting a model for deployment domain.
- Between dataset coverage, which may be used for applications such as computing the difference between a test and training set to generate representative and challenging test sets.
- Data splitting algorithms, which may be used to construct balanced data sets for training or testing. The SIE framework uses the data splitting algorithm for constructing a balanced universal test set. In the future, this framework could use the same algorithm to produce training sets that are not only covering but also balanced.
- Prioritization of samples to best cover a space, which may be used to support high information gain in resource restricted scenarios such as labeling and retraining.
- SIE, which is used for identification of critical metadata factors.

CODEX produces visualizations to characterize the data and the resulting model, such as:

- *Binary coverage:* the interactions that are present/absent.
- *Frequency coverage:* the proportion of dataset samples containing each interaction.
- *Performance by interaction:* how the model performs on data containing each interaction.
- *Performance by frequency:* model performance as a function of frequency coverage.
- *Performance by coverage:* how models trained on datasets with different degrees of coverage perform and how performance differs as a function of distance between the training and testing dataset.

Academic/research code existed for these functionalities as they were developed as proof of concept for Project Maven, but the code is not well documented and lacks rigor of good software development practices, such as modularity. This effort's main objective was to mature the tool for practical use outside the academic setting. This requires code cleaning and documentation, as well as training material. Additional research is needed for new metrics. Combinatorial coverage metrics in CODEX are currently limited to binary coverage of discretized variables. In this effort, the research team explored frequency coverage to understand how over- or under-representation impacts performance. CODEX functions enable equitable AI (e.g., performance by interaction, training data adequacy) as well as reliable AI (e.g., test data adequacy). New metrics enable equitable AI (e.g., frequency coverage may be useful for detecting representation bias).

Development of code that became CODEX has been funded by multiple sponsors beginning in 2019. DOT&E Cyber Assessment Program (CAP) substantially funded the development of CODEX and provided insight into tool applications in assessments. Another objective of this effort was to maintain involvement with the CAP AI Working Group.

### *METHODS*

The CODEX codebase was matured under this effort with code cleaning and documentation. Logging controls and visualization were added to assist developers in creating new modes and enhancing interpretability for testers. The proportion frequency map is a new coverage map variant designed to convey the relative dominance of a given feature interaction within the overall dataset. One version measures representation as a simple proportion of counts out of all samples in the set, while another standardizes the counts to express overrepresentation and underrepresentation. Another visualization plots the standardized proportion frequency of an interaction against its corresponding performance as a cluster of points, as well as displaying regression lines and statistical significance correlation between frequency of feature interactions among samples and ML model performance.

In addition, this effort produced code to enable CODEX's interoperability with a ML harness that, given a model and a dataset, can automatically train and test without manual user input for every ML instance. The model provided is intended to operate on datasets as created by the native data-splitting algorithms offered by CODEX, effectively linking combinatorial and ML testing processes to explore data. The first working version of this integration was demonstrated through the SIE framework. Besides present constraints of model-specificity, users can automatically train many models from scratch and store the results and performance, all from a single mode.

Under this effort, the research team also hosted one day of the DOT&E CAP AI Working Group (AIWG) at Virginia Tech Arlington in March 2024, participated in the joint CAP-SIPET AIWG in July 2024, and co-hosted a CT for AI workshop with National Institute of Standards and Technology (NIST) for DOT&E and related government test community in September 2024. The team developed tutorial material to accompany CODEX, which will be piloted at the workshop.

### *RESULTS/FINDINGS*

The new visualizations enabled new analysis within CODEX modes. The performance by interaction mode was most improved by the new visualizations and results, which computes the aggregate performance of samples that possess a certain interaction for each t-way interaction in the training data. Through the significance testing plot, the correlation between representation and performance is no longer subjective. This enables determining when data representation bias may lead to differential performance, which may be a violation of fairness in certain use cases and is almost always a violation of reliability.

The creation of a combinatorial coverage-ML pipeline established the foundation for efficient test and evaluation. The benefits of an automated model training capability have been demonstrated in its application to the SIE framework. New visuals added to the codebase show that models exhibit greater performance variability on the test samples containing the interactions not covered in the training data compared to the test samples containing interactions that were covered. This supports the hypothesis that models may behave erratically when outside of their operating envelope. The construction of a rudimentary test harness gave CODEX access to model weights and outputs, enabling future development of custom performance analysis functions.

### *RECOMMENDATIONS*

CT has been demonstrated to show promise for ML T&E but adaptation from software testing is not direct. As one example, algorithms that construct a covering array for software testing assume that rows can be written down symbol by symbol and then executed. In ML, these rows must correspond to the features of a sample, so algorithms must instead perform row selection instead of construction. CT requires discretization of factor levels, but it is not always clear how to determine the correct binning scheme. The team's work on bias showed that there may be a correlation between performance and frequency coverage, and it is stronger when the data is more imbalanced, but the correlation is not clear. The team's hypothesis is that there is a threshold where the model gains enough information to perform equivalently across groups. Future work should identify the gaps between testing software and ML, explore applying CT to LLMs and reinforcement learning, explore coverage metrics for continuous variables, explore how to determine the threshold for frequency coverage where performance differences disappear, and pair coverage with distance metrics for post-deployment monitoring of ML systems.

## TEST & EVALUATION FOR MULTI-FIDELITY AI MODELS

### *RESEARCH OBJECTIVES*

The goal was to establish an efficient operational test and evaluation method for AI-enabled systems that can integrate evidence generated at various stages during the acquisition process. The specific objectives during the reporting period were to 1) implement the method for sequential test generation, and 2) demonstrate the method using an autonomous driving use case.

### *METHODS*

The team developed a sequential multi-fidelity method that integrates information from multiple levels of representativeness of the system and the environment to support T&E. In this method, the T&E problem is formulated as a decision problem where the objective functions are (i) minimizing the uncertainty in achieving a set of requirements and (ii) minimizing the cost (C) of T&E. The uncertainty in the requirements is quantified using Entropy (H), and the trade-off between the minimization of entropy and the cost incurred in the T&E process is quantified using a utility function U(H, C).

The steps in the multi-fidelity testing method are as follows:

1. *Identify requirements and the corresponding failure modes:* the requirements are typically developed early in the systems engineering process. Only a subset of test requirements may be relevant to a specific testing phase or the part of the system being tested.

2. *Characterize the space of multi-fidelity representations:* this step includes identifying the dimensions along which the representations of the system and the environment are different.

3. *Characterize the representativeness:* in this step, the cost of development of the representations is quantified, along with the probability with which the outcomes of a test represent the real performance of the system.

4. *Design the test plan:* this step consists of identifying the set of representations to use to test a given set of requirements and identifying the smallest set of tests using each model to maximize the T&E utility. This optimization step can be carried out by using single-step optimization or sequential optimization techniques. In this research, the focus was on sequential optimization.

During this reporting period, the approach was demonstrated using an example of a visual perception system in an autonomous vehicle (AV) use case. A list of requirements was created, with examples including: 1) system must identify vehicles within a certain distance of itself with a certain level of accuracy, 2) system must operate on a 4k/1080p/480 HD video format, 3) system must operate in direct overhead sunlight, and 4) system must operate under 5% random-occlusion. The tests were carried out using driving simulations with levels of realism of the environmental conditions. The driving simulations were created in Unity. For the cost function, the team estimated the total computational cost required in three aspects: total pixels generated, layers required to simulate the weather condition, and whether the video needed to be post-processed as occluded videos. A sequential test plan was then developed considering the cost of running the simulations and the corresponding value in reducing the uncertainty of satisfying the requirements.

### *RESULTS/FINDINGS*

The effect of the tests on the overall information entropy is shown in Figure 6 (below). It was observed that the proposed method resulted in a gradual decrease in entropy over time. Additional testing after Iteration 25 did not result in a change in entropy.

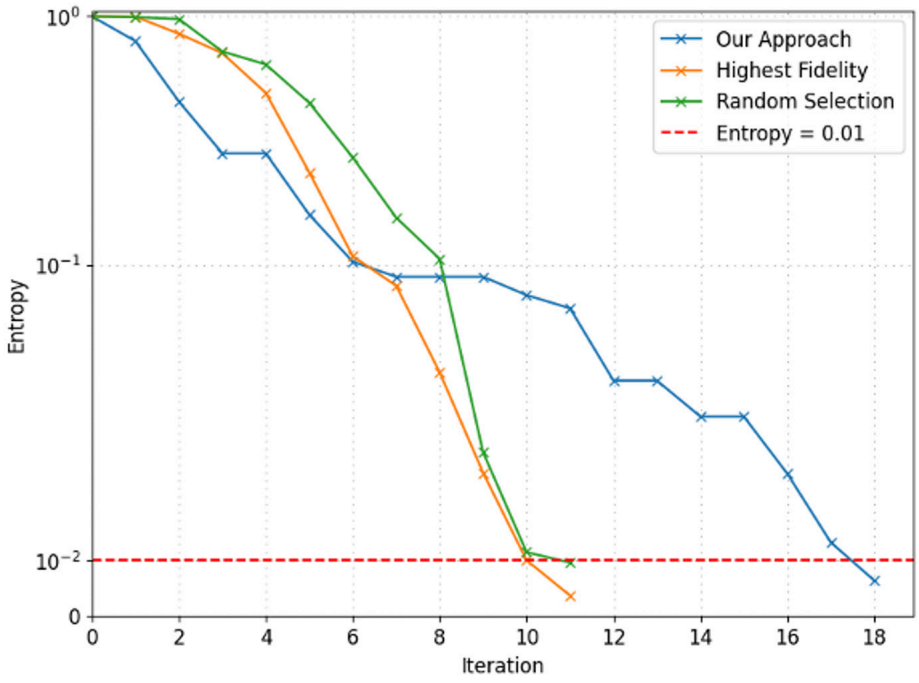


**Figure 6: Reduction in uncertainty about the requirements, measured using Entropy**

To compare the effectiveness of the proposed method, the number of tests needed and the corresponding cost to achieve a threshold of Entropy (H=0.1) are compared with (a) test plan using highest fidelity model only, and (b) random selection of the tests (see Table 4 below). The team found that the proposed approach can be used to develop a test-sequence with lower cumulative costs compared to using high-fidelity models alone. The cost is lower despite the higher number of tests required, as majority of the tests are carried out using low-fidelity models. In summary, the sequential decision-making method is cost-effective for T&E.

**Table 4: Comparison of the number of tests needed and the corresponding costs**

| | Proposed Method | Highest Fidelity | Random Selection |
|---|---|---|---|
| *Number of tests needed for H<0.1* | 17 | 10 | 12 |
| *Cost* | 1.61 | 4.34 | 3.07 |

### *RECOMMENDATIONS*

The method was tested for multi-fidelity representations of the environmental conditions, and shows promise for designing operational test plans and should be tested for a realistic acquisition scenario. The proposed method naturally accounts for the cost of data and testing and can be extended further to make decisions about acquiring intellectual property (IP) related to training data. In future work, this method should be extended to multiple fidelities of the AI/ML system.

## CONCLUSIONS

The research team has completed the option year of critical research in support of DOT&E's Implementation Plan and strategy for advancing technology for Test in Evaluation to continual needs of the warfighter. The team has advanced the JTC with simulations and the development of a reference architecture to aid in the assessment of joint capabilities to ensure mission success in the DoD complex kill webs. The research advanced Integrated test capabilities with further application of Bayesian methods and the maturation of an R-Shiny application, which is available to T&E practitioners. The research advanced DE methods by leading a community-wide effort to deliver IDSK and MB TEMP tooling that can be readily adopted by DoD personnel. The team also developed a T&E testbed to mature methods and tooling interfaces for the future. The research advanced the application of AI/ML in T&E by establishing best practices, conducting novel research on T&E of AI, developing tooling, and conducting numerous workshops. The research team has developed methods for T&E of multi-fidelity AI applications and recommended methods for ensuring training data and AI models will perform as expected once fielded. Over the last year, the research team has also helped shape the T&E policy changes that DOT&E has developed and has made initial contributions into how best to deploy the revised policy.

Though this initial research contract has come to closure with tremendous accomplishments, there is much work still to be done. Many of the research recommendations have been incorporated into forthcoming guidance by DOT&E for the OT&E and Live Fire Test & Evaluation (LFT&E) community. However, more effort is needed in translating the research papers and products developed into capabilities for the T&E community. There is a need for training materials, testbeds, tooling, and recommendations for T&E process implementation. DOT&E should continue to emphasize the need for exemplars to help translate conceptual recommendations into practice.

# APPENDIX A. DELIVERABLES AND PRODUCTS

## PILLAR 1 – TEST THE WAY WE FIGHT

### *JOINT TEST CONCEPT*

**Joint Test Concept (JTC) DOT&E Sponsor Briefing I**

- **Date/Location:** 12/11/2023; Virtual
- **Presenters:** Ms. Christina Houfek
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies and DOT&E action officers

**Host Joint Test Concept Workshop I**

- **Date/Location:** 2/7/2024; Arlington, VA
- **Audience:** JTC Defense and T&E community of interest (COI) contributing to JTC design and concept validation

**Joint Test Concept Overview Presentation**

- **Event:** DATAWorks 2024
- **Date/Location:** 4/16 - 4/18/2024, Alexandria, VA
- **Presenter:** Ms. Christina Houfek
- **Audience:** Defense and aerospace community members attending DATAWorks

**Joint Test Concept Pilot Year II Update Final Report**

- **Authors:** Dr. Maegen Nix, Ms. Christina Houfek
- **Publication/Release Date:** 4/16 – 4/18/2024
- **Delivered to:** Participants of DATAWorks as supporting material to the JTC Overview presentation (listed above); JTC defense and T&E COI; DOT&E Pillar 1 Sponsor

**Joint Test Concept DOT&E Sponsor Briefing II**

- **Event:** May DOT&E Monthly Technical Meeting
- **Date/Location:** 5/6/2024; Virtual
- **Presenter:** Ms. Christina Houfek
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies and DOT&E action officers

**Host Joint Test Concept Workshop II: JTC T&E Simulation Event**

- **Date/Location:** 6/12/2024; Arlington, VA
- **Audience:** JTC Defense and T&E community of interest (COI) contributing to JTC design and concept validation

**Reimagining T&E in the Joint Test Concept: Interactive Presentation Session**

- **Event:** ITEA MDO Conference
- **Date/Location:** 7/18/2024, Alexandria, VA
- **Presenters:** Dr. Maegen Nix, Ms. Natalie Wells
- **Audience:** Defense and T&E community members attending ITEA MDO conference

**JTC Simulation Workshop: T&E Simulation Outcomes (JTC Workshop II Outcomes Report)**

- **Authors:** Dr. Maegen Nix, Mrs. Christina Houfek, Mrs. Natalie Wells, Mr. Grant Beanblossom, Mr. Daniel Wolodkin, Mr. Kobie Marsh
- **Publication/Release Date:** 7/22/2024
- **Delivered to:** JTC defense and T&E COI; DOT&E Pillar 1 Sponsor

**Host Joint Test Concept Workshop III: JTC Implementation Roadmap Development**

- **Date/Location:** 8/7/2024; Arlington, VA
- **Audience:** JTC Defense and T&E community of interest (COI) contributing to JTC design and concept validation

**Joint Test Concept Implementation Roadmap Workshop Outcomes Report (JTC Workshop III Outcomes Report)**

- **Authors:** Dr. Maegen Nix, Ms. Christina Houfek, Ms. Natalie Wells, Mr. Grant Beanblossom
- **Publication/Release Date:** 8/30/2024
- **Delivered to:** JTC defense and T&E COI; DOT&E Pillar 1 Sponsor

**JTC Reference Architecture Report**

- **Authors:** Dr. Maegen Nix, Ms. Christina Houfek, Ms. Natalie Wells
- **Publication/Release Date:** 8/30/2024
- **Delivered to:** JTC defense and T&E COI; DOT&E Pillar 1 Sponsor

## PILLAR 2 – ACCELERATE THE DELIVERY OF WEAPONS THAT WORK

### *INTEGRATED TESTING*

**Integrated Test DOT&E Sponsor Briefing**

- **Event:** March DOT&E Monthly Technical Meeting
- **Date/Location:** 3/4/2024; Virtual
- **Presenter:** Dr. Justin Krometis
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies, DOT&E action officers

**Automated Tools to Improve the Accessibility of Bayesian Methods Poster Presentation**

- **Event:** Virginia Tech 11th Annual Hume Center Colloquium
- **Date/Location:** 4/9/2024, Blacksburg, VA
- **Presenter:** Mr. Kyle Risher
- **Audience:** Colloquium attendees

**Bayesian Methods for Integrated Testing Mini Tutorial**

- **Event:** DATAWorks 2024
- **Date/Location:** 4/16 - 4/18/2024, Alexandria, VA
- **Presenter:** Dr. Justin Krometis
- **Audience:** Defense and aerospace communities

**R-Shiny App: Integrated Testing for Reliability ZIP File & Web Link**

- **Developers:** Mr. Kyle Risher, Dr. Justin Krometis
- **Purpose:** This R Shiny app illustrates a Bayesian approach to integrated reliability testing. It generalizes the Stryker app delivered during the base year by allowing a user to upload their own reliability data and includes some other convenience features, like figure downloading.
- **Delivery date:** 8/12/2024
- **Recipient:** DOT&E SIPET Chief Scientist

**R-Shiny App: Integrated Testing for Binary Data ZIP File & Web Link**

- **Developers:** Mr. Jared Clark, Dr. Justin Krometis
- **Purpose:** This R Shiny app illustrates a Bayesian approach to integrated testing with binary data.
- **Delivery date:** 8/12/2024
- **Recipient:** DOT&E SIPET Chief Scientist

**R-Shiny App: Integrated Testing for Binary Data ZIP File & Web link**

- **Developers:** Mr. Jared Clark, Dr. Justin Krometis
- **Purpose:** This R Shiny app was built from the previous Developmental Testing/OT one for binary data by imagining what an interface might look like for a Bayesian approach to integrated testing across an arbitrary number of phases of test.
- **Delivery date:** 8/12/2024
- **Recipient:** DOT&E SIPET Chief Scientist

### *DIGITAL ENGINEERING*

**Digital Test and Evaluation Master Plan (dTEMP)**

**Use Cases of the Digital Test and Evaluation Master Plan (dTEMP) Report**

- **Authors:** Dr. Joe Gregory, Dr. Alejandro Salado
- **Delivery date:** 5/21/2024
- **Recipient:** DOT&E Action Officers, DOT&E Science Advisor

**User Guide (V1)**

- **Authors:** Dr. Joe Gregory, Dr. Alejandro Salado
- **Delivery date:** 6/3/2024
- **Recipient:** DOT&E Science Advisor

**Digital Engineering Sponsor Briefing**

- **Event:** June DOT&E Monthly Technical Meeting
- **Date/Location:** 6/10/2024; Virtual
- **Presenters:** Multiple
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies, DOT&E action officers, DOT&E Science Advisor

**Planned/Hosted FY 2025 MB/TEMP Planning Workshop**

- **Date/Location:** 6/13/2024; Arlington, VA
- **Audience:** DOT&E Action Officers, DOT&E Science Advisor, AIRC research team, OSD (R&E) DTE&A, Space Force

**Model-Based Integration and Test Planning: Automating the Propagation and Verification of Expert Knowledge using Ontologies**

- **Authors:** Dr. Joe Gregory, Dr. Alejandro Salado
- **Event:** 11th International Systems & Concurrent Engineering for Space Applications Conference (SECESA 2024)
- **Date/Location:** 9/25 - 9/27/2024; Strasbourg, France

**Integrated Decision Support Key (IDSK)**

**IDSK Sponsor Briefing**

- **Event:** December DOT&E Monthly Technical Meeting
- **Date/Location:** 12/11/2023; Virtual
- **Presenter:** Dr. Kelli Esser
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies, DOT&E action officers, DOT&E Science Advisor

**Host IDSK JUC Working Group Workshop**

- **Date/Location:** 1/17 - 1/18/2024; Arlington, VA
- **Audience:** IDSK JUC WG members

**Attend IDSK JUC Working Group Workshop**

- **Date/Location:** 3/19 - 3/21/2024; Smyrna, GA
- **Audience:** IDSK JUC WG members

**Integrated Decision Support Key (IDSK) Presentation**

- **Event:** DOT&E Lunch & Learn Series
- **Date/Location:** 5/8/2024; Virtual
- **Presenter:** Dr. Kelli Esser
- **Audience:** Director of DOT&E, DOT&E Action Officers

**Attend Digital Twin/Digital Engineering/Mission Engineering for T&E Small/Focused Workshop**

- **Date/Location:** 7/24 - 7/25/2024; Arlington, VA
- **Purpose:** showcase of participants' current work on developing technology specific to Test and Evaluation in a Digital environment

**Operation Safe Passage (OSP)**

**Operation Safe Passage Mock Interim Design Review**

- **Presenters:** Dr. Peter Beling, Dr. John Gilbert, Dr. Joe Gregory, Mr. Geoffrey Kerr, Dr. Alejandro Salado, Mr. Tim Sherburne, Mr. Daniel Wolodkin
- **Date/Location:** 6/25/2024; Virtual
- **Audience:** DOT&E action officers

**Operation Safe Passage Delta Interim Design Review**

- **Presenters:** Dr. Peter Beling, Dr. John Gilbert, Dr. Joe Gregory, Mr. Geoffrey Kerr, Dr. Alejandro Salado, Mr. Tim Sherburne, Mr. Daniel Wolodkin
- **Date/Location:** 8/27/2024; Virtual
- **Audience:** DOT&E action officers

**Verification, Validation, and Uncertainty Quantification (VVUQ)**

**Uncertainty Quantification Analysis Workflow Training Guide**

- **Developer:** Dr. Sheri Martinelli
- **Audience:** Undergraduate interns
- **Date:** March 2024

**Uncertainty Quantification in Digital Twins Sponsor Briefing**

- **Event:** May DOT&E Monthly Technical Meeting
- **Date/Location:** 5/7/2024; Virtual
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies and DOT&E action officers

**VVUQ Overview Presentation to TETRA and NSIC**

- **Event:** VT NSI Presentation to TETRA and NSIC
- **Date/Location:** 5/13/2024; Blacksburg, VA
- **Presenter:** Dr. John Gilbert
- **Audience:** Representatives from Tetra and NSIC

**Uncertainty Quantification in Digital Twins Presentation**

- **Event:** ASME VVUQ 2024
- **Date/Location:** 5/15/ - 5/17/2024; College Station, TX
- **Presenter:** Dr. Sheri Martinelli
- **Audience:** VVUQ 2024 Conference attendees

**Uncertainty Quantification in Digital Twins Technical Paper Submission**

- **Authors:** Dr. Sheri Martinelli, Justin Valenti, Chris Rogan, Michael Warren
- **Event:** ASME VVUQ 2024
- **Date/Location:** 5/15/ - 5/17/2024; College Station, TX

## PILLAR 4 – PIONEER T&E OF WEAPON SYSTEMS BUILT TO CHANGE OVER TIME

### *T&E FOR AI/ML*

**Assured Autonomy, Artificial Intelligence, and Machine Learning: A Roundtable Discussion**

- **Event:** Second IEEE Workshop on Assured Autonomy, AI, and Machine Learning
- **Date/Location:** 11/2/2023, Atlanta, GA
- **AIRC Panelist:** Dr. Jaganmohan Chandrasekaran, Dr. Erin Lanus

**Testing Cognitive Systems: The Big Challenges Panel Discussion**

- **Event:** 2023 ITEA Annual T&E Symposium
- **Date/Location:** 12/6/2023, Destin, FL
- **AIRC Panelist:** Dr. Laura Freeman

**TRMC JTEX-08 Conference Participation**

- **Date/Location:** 12/12 – 12/14/2024, Orlando, FL
- **AIRC Participant:** Dr. Tyler Cody

**T&E AI/ML Sponsor Briefing**

- **Event:** February DOT&E Monthly Technical Meeting
- **Date/Location:** 2/5/2024; Virtual
- **Audience:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies and DOT&E action officers

**Host DOT&E AIWG Workshop**

- **Date/Location:** 3/5/2024; Arlington, VA
- **Audience:** DOT&E Cyber Assessment Program (CAP)

**Coverage of Data Explorer (CODEX) Update Presentation**

- **Event:** DOT&E CAP AIWG
- **Date/Location:** 3/5/2024; Arlington, VA
- **Presenter:** Dr. Erin Lanus
- **Audience:** DOT&E Cyber Assessment Program (CAP)

**DoDM Technical Review/Adjudication of CDAO Comments**

- **Technical Reviewer:** Dr. Laura Freeman
- **Delivery Date:** 4/2/2024
- **Recipient:** Deputy Director, Operational Test and Evaluation for Strategic Initiatives, Policy, and Emerging Technologies and DOT&E AI action officer

**On the Role of Loss-Driven Analysis in Assessing AI Ethics Conference Presentation**

- **Event:** SPIE Defense + Commercial Sensing Conference
- **Date/Location:** 4/12 - 4/25/2024, National Harbor, MD
- **Presenter:** Dr. Tyler Cody

**DoDI 5000.XF, DoDM 5000.UX (TEMP/TES), DoDM 5000.UW (M&S V&V) Companion Guide Plans**

- **Technical Reviewer:** Dr. Laura Freeman
- **Delivery Date:** 5/14/2024
- **Recipient:** DOT&E Science Advisor

**Coverage of Data Explorer (CODEX) Status Update and Deep Dive Presentation**

- **Event:** DOT&E CAP/SIPET Technical Exchange Meeting
- **Date/Location:** 5/14/2024; Virtual
- **Presenter:** Dr. Erin Lanus
- **Audience:** DOT&E Cyber Assessment Program (CAP)

**Coverage for Identifying Critical Metadata in Machine Learning Operating Envelopes Conference Presentation**

- **Event:** IEEE International Conference on Software Testing
- **Date/Location:** 5/17 - 5/31/2024; Toronto, Canada
- **Presenter:** Dr. Erin Lanus
- **Audience:** the international software testing, verification, and validation community

**STPA Deep Dive Presentation**

- **Event:** AI AIRC Bi-Weekly Meeting
- **Date/Location:** 6/20/2024; Virtual
- **Presenter:** Dr. Tyler Cody
- **Audience:** Pillar 4 DOT&E action officer

**Support Organization of the Director, Operational Test and Evaluation Artificial Intelligence Working Group Event**

- **Date/Location:** 7/22 – 7/24/2024; Alexandria, VA
- **Audience:** DOT&E CAP and DOT&E SIPET Pillar 4 community

**T&E Considerations for Large Language Models (LLMs) Presentation**

- **Event:** Director, Operational Test and Evaluation Artificial Intelligence Working Group
- **Date/Location:** 7/23/2024; Alexandria, VA
- **Presenter:** Dr. Laura Freeman
- **Audience:** DOT&E Pillar 4 SIPET lead and contributors

**Coverage for Testing Machine Learning Presentation**

- **Event:** DOT&E AIWG
- **Date/Location:** 7/23/2024; Alexandria, VA
- **Presenter:** Dr. Erin Lanus
- **Audience:** DOT&E Pillar 4 SIPET lead and contributors

**Coverage for Testing Machine Learning Lighting Talk Presentation**

- **Event:** DOT&E AIWG
- **Date/Location:** 7/23/2024; Alexandria, VA
- **Presenter:** Dr. Erin Lanus
- **Audience:** DOT&E Pillar 4 SIPET lead and contributors

**Multi-Fidelity Testing and Evaluation of AI-Enabled Systems Presentation**

- **Event:** Director, Operational Test and Evaluation Artificial Intelligence Working Group
- **Date/Location:** 7/23/2024; Alexandria, VA
- **Presenter:** Dr. Jitesh Panchal
- **Audience:** DOT&E Pillar 4 SIPET lead and contributors

**STPA & AI Ethics Presentation**

- **Event:** Director, Operational Test and Evaluation Artificial Intelligence Working Group
- **Date/Location:** 7/23/2024; Alexandria, VA
- **Presenter:** Dr. Tyler Cody
- **Audience:** DOT&E Pillar 4 SIPET lead and contributors

**A Multi-Fidelity Approach to Testing and Evaluation of AI-Enabled Systems Master's Thesis Defense**

- **Date/Location:** 7/27/2024; West Lafayette, IN
- **Presenter:** Mr. Robert Seif
- **Audience:** Master's thesis defense committee

**A Multi-Fidelity Approach to Testing and Evaluation of AI-Enabled Systems Conference Presentation**

- **Event:** ASME IDETC-CIE 2024
- **Date/Location:** 8/25 - 8/28, 2024; Washington, DC
- **Presenter:** Mr. Zichong Yang
- **Audience:** the design and manufacturing engineering community

**Host Combinatorial Testing of Artificial Intelligence Enabled Systems Workshop**

- **Date/Location:** 9/4/2024; Arlington, VA
- **Audience:** T&E practitioners

**Data Assurance and Combinatorial Coverage Presentation**

- **Event:** Combinatorial Test of Artificial Intelligence Enabled Systems Workshop
- **Date/Location:** 9/4/2024; Arlington, VA
- **Presenter:** Dr. Erin Lanus
- **Audience:** T&E practitioners

**Application of Combinatorial Testing in Testing ML Systems**

- **Event:** Combinatorial Test of Artificial Intelligence Enabled Systems Workshop
- **Date/Location:** 9/4/2024; Arlington, VA
- **Presenter:** Dr. Jaganmohan Chandrasekaran
- **Audience:** T&E practitioners

**CODEX Python Software Package**

- **Developers:** Mr. Brian Lee, Dr. Erin Lanus, Mr. Dylan Steburg
- **Delivery date:** 9/4/2024
- **Recipient:** Delivered during DOT&E Combinatorial Testing of Artificial Intelligence-Enabled Systems Workshop

**CODEX Tutorials**

- **Developers:** Mr. Brian Lee, Dr. Erin Lanus
- **Purpose:** Accompanying training material for the delivered CODEX Python package
- **Delivery date:** 9/4/2024
- **Recipient:** Delivered during the DOT&E Combinatorial Testing of Artificial Intelligence-Enabled Systems Workshop

***PILLAR 5 – FOSTER AN AGILE AND ENDURING T&E ENTERPRISE WORKFORCE*** *(Note: although the research team was not contracted to support Pillar 5 during the option year, the team has engaged in the following workforce development activities)*

**Engaging the Next Generation T&E Workforce ITEA Panel Discussion**

- **Event:** 2023 ITEA Annual T&E Symposium
- **Date/Location:** 12/6/2023, Destin, FL
- **AIRC Panelist:** Dr. Laura Freeman

**Support Pathfinders program / DCTC Synergies**

- **AIRC Contributors:** Dr. Laura Freeman, Mr. Geoffrey Kerr
- **Period of support:** 9/2023 – Current

## APPENDIX B. LIST OF PUBLICATIONS RESULTED

Chandrasekaran, Jaganmohan, et al. "Leveraging Combinatorial Coverage in the Machine Learning Product Lifecycle." *Computer* 57.7 (2024): 16-26. Digital. <https://ieeexplore.ieee.org/abstract/document/10574393>.

—. "Testing Machine Learning: Best Practices for the Lifecycle." *Naval Engineers Journal* 136.1&2 (2024): 249-263. Digital. <https://bonotom.com/flipbook/2096/#2096/250>.

Cody, Tyler, Laura Freeman and Peter Beling. "On the Role of Loss-Driven Analysis in Assessing AI Ethics." *Assurance and Security for AI-Enabled Systems*. National Harbor: SPIE, 2024. <https://proceedings.spiedigitallibrary.org/conference-proceedings-of-spie/13054/1305403/On-the-role-of-loss-driven-systems-analysis-in-assessing/10.1117/12.3012385.full#_=_>.

Freeman, Laura, et al. "Accelerating Implementation of Critical Joint Warfighting Concepts and Capabilities." *Naval Engineers Journal* 136.1&2 (2024): 41-49. Digital. <https://bonotom.com/flipbook/2096/#2096/42>.

Gregory, Joe and Alejandro Salado. "A Semantic Approach to Spacecraft Verification Planning Using Bayesian Networks." *IEEE Aerospace Conference Proceedings*. Big Sky, 2024. 1-12. Digital. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10521072&isnumber=10520936>.

—. "An ontology-based digital test and evaluation master plan (dTEMP) compliant with DoD policy." *Systems Engineering Journal* (2024). Digital. <https://incose.onlinelibrary.wiley.com/doi/epdf/10.1002/sys.21769>.

—. "dTEMP: From Digitizing to Modeling the Test and Evaluation Master Plan." *Naval Engineers Journal* 136.1&2 (2024): 134-146. Digital. <https://bonotom.com/flipbook/2096/#2096/136>.

Jain, Sanidhya, Jin-Suh Park and Karen Marais. "Digital Twins for Early Resolution of Human Factors Issues: A Review of the State of the Art." *Naval Engineers Journal* 136.1&2 (2024): 271-279. Digital. <https://bonotom.com/flipbook/2096/#2096/272>.

Kauffman, Justin, et al. "Math-Theoretic Considerations for Accelerating the Implementation of Combined Joint All-Domain Command and Control Solutions." *Naval Engineers Journal* 136.1&2 (2024): 83-92. Digital. <https://bonotom.com/flipbook/2096/#2096/84>.

Kuhn, D. Richard, et al. "Assured Autonomy through Combinatorial Methods." *Computer* 57.5 (2024): 86-90. Digital. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10517690>.

Lanus, Erin, et al. "Coverage for Identifying Critical Metadata in Machine Learning Operating Envelopes." *2024 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. Toronto, 2024. 217-226.

McCarthy, Nicola, et al. "Key Steps to Fielding Combat Credible AI-Enabled Systems." *Naval Engineers Journal* 136.1&2 (2024): 237-248. Digital. <https://bonotom.com/flipbook/2096/#2096/238>.

Nix, Maegen, et al. "Capability Neutralization Warfare: : A Systems-Informed Theory of Victory." *Naval Engineers Journal* 136.1&2 (2024): 33-40. Digital. <https://bonotom.com/flipbook/2096/#2096/34>.

O'Toole, Raymond, Nilo Thomas and Jeff Upton. "Targeted Education and Training at Universities— Key to the Development of the Department of Defense Workforce of the Future." *Naval Engineers Journal* 136.1&2 (2024): 305-312. Digital. <https://bonotom.com/flipbook/2096/#2096/306>.

Risher, Kyle, et al. "Maximizing the Use of Data to Make Winning Decisions in the Face of Finite Resources." *Naval Engineers Journal* 136.1&2 (2024): 111-120. Digital. <https://bonotom.com/flipbook/2096/#2096/112>.

Seif, Robert J. "A Multi-Fidelity Approach to Testing and Evaluation of AI-Enabled Systems." Purdue University Graduate School, 27 July 2024. <https://doi.org/10.25394/PGS.26364277.v1>.

Seif, Robert J., et al. "Towards Multi-Fidelity Test and Evaluation of Artificial Intelligence and Machine Learning-Based Systems." *The ITEA Journal of Test and Evaluation* 45.1 (2024). Digital. <https://itea.org/journals/volume-45-1/towards-multi-fidelity-test-and-evaluation-of-artificial-intelligence-and-machine-learning-based-systems/>.

Seif, Robert, et al. "A Multi-Fidelity Approach to Testing and Evaluation of AI-Enabled Systems." *ASME 2024 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2024)*. Washington, DC, 2024.

Werner, Jeremy, et al. "Integrated Decision Support Key: Advancing Acquisition Decisions with Data Models and Tools." *Naval Engineers Journal* 136.1&2 (2024): 121-133. Digital. <https://bonotom.com/flipbook/2096/#2096/122>.

**Pending Publications**

Chandrasekaran, Jaganmohan, et al. "Test & Evaluation Best Practices for Machine Learning-Enabled Systems." *Journal of Systems and Software* (2024): TBD.

Kauffman, Justin A, et al. "Best Practices for Uncertainty Quantification of Digital Twins." *TBD* (n.d.).

Krometis, Justin, et al. "A Comparison of Bayesian Methods for Integrated Test and Evaluation." *Military Operations Research Journal* (n.d.).

Nix, Maegen, Christina Houfek and Natalie Wells. "Reimagining T&E for the Modern Joint Environment: The Joint Test Concept." *ITEA Journal of Test and Evaluation* (2024).

Sieck, Victoria R.C, Justin Krometis and Steven Thorsen. "A Framework for Using Priors in a Continuum of Testing." *Military Operations Research* (2024).

Wolodkin, Daniel, Erin Lanus and Laura Freeman. "Hierarchical Scoring for Evaluating Machine Learning Models." n.d.

## REFERENCES

Acton, J. M. "Hypersonic Boost-Glide Weapons." *Science & Global Security* 23 (2015): 191-219.

Adams, B. M., et al. *Dakota 6.19.0 documentation.* Technical Report SAND2023-133920. Albuquerque, NM: Sandia National Laboratories, 2023. Available online from http://snl-dakota.github.io. <http://snl-dakota.github.io.>.

Ao, Dan, Zhen Hu and Sankaran Mahadevan. "Dynamics Model Validation Using Time-Domain Metrics." *Journal of Verification, Validation and Uncertainty Quantification* 2.1 (2017): 011004. <https://doi.org/10.1115/1.4036182>.

Banerjee, Bonny, et al. "Digital Twin: A Quick Overview." *The ITEA Journal of Test and Evaluation* 45.1 (2024).

Cortes, Luis, et al. *Advance M&S in Acquisition T&E*. Technical Report MTR210454. McLean, VA: The MITRE Corporation, 2021.

Cruse, T. A., et al. "Mechanical System Reliability and Risk Assessment." *AIAA Journal* 32.11 (1994): 2249-2259. <https://doi.org/10.2514/3.12284>.

Du, Yuxian, et al. "A new method of identifying influential nodes in complex networks based on TOPSIS." *Physica A: Statistical Mechanics and its Applications* 399 (2014): 57-69. <https://www.sciencedirect.com/science/article/pii/S0378437113011552>.

Kauffman, Justin A, et al. "Best Practices for Uncertainty Quantification of Digital Twins." *In Prep* (2024).

Kerleguer, Baptiste. "Multifidelity Surrogate Modeling for Time-Series Outputs." *SIAM/ASA Journal on Uncertainty Quantification* 11.2 (2023): 514-539. <https://doi.org/10.1137/20M1386694>.

Mahadevan, Sankaran and Prakash Raghothamachar. "Adaptive simulation for system reliability analysis of large structures." *Computers & Structures* 77.6 (2000): 725-734. <https://www.sciencedirect.com/science/article/pii/S0045794900000134>.

Miller, S. W., M. A. Yukish and T. W. Simpson. "Design as a sequential decision process." *Struct. Multidisc. Optim.* 57 (2018): 305-324. <https://doi.org/10.1007/s00158-017-1756-7>.

National Research Council. *Assessing the Reliability of Complex Models Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. Washington, DC: National Academy Press, 2012.

*ONNX Models*. 21 December 2023. https://github.com/onnx/models/tree/main/validated/vision/classification/mnist/model. July 2024.

*Onnx2json*. January 2023. https://pypi.org/project/onnx2json/. July 2024.

Owen, Art B. "Sobol' Indices and Shapley Value." *SIAM/ASA J. Uncertainty Quantification* 2 (2014): 245-251.

Rebba, Ramesh, Sankaran Mahadevan and Shuping Huang. "Validation and error estimation of computational models." *Reliability Engineering & System Safety* 91.10 (2006): 1390-1397. <https://www.sciencedirect.com/science/article/pii/S0951832005002486>.

Sobol, I. M. "Sensitivity estimates for nonlinear mathematical models." *Math. Model. Comp. Exp.* 1.4 (1993): 407-414.

Thelen, Adam, et al. "A comprehensive review of digital twin -- part 1: modeling and twinning enabling technologies." *Structural and Multidisciplinary Optimization* 65.12 (2022): 354. <https://doi.org/10.1007/s00158-022-03425-4>.

—. "A comprehensive review of digital twin—part 2: roles of uncertainty quantification and optimization, a battery digital twin, and perspectives." *Structural and Multidisciplinary Optimization* 66 (2022): 1. <https://doi.org/10.1007/s00158-022-03410-x>.

Tracy, Cameron L and David Wright. "Modeling the Performance of Hypersonic." *Science & Global Security* 28.3 (2020): 135-170.

Urbina, Angel, Sankaran Mahadevan and Thomas L. Paez. "Quantification of margins and uncertainties of complex systems in the presence of aleatoric and epistemic uncertainty." *Reliability Engineering & System Safety* 96.9 (2011): 1114-1125. <https://www.sciencedirect.com/science/article/pii/S0951832011000640>.

Werner, Jeremy, et al. "Integrated Decision Support Key: Advancing Acquisition Decisions with Data Models and Tools." *Naval Engineers Journal* 136.1&2 (2024): 121-133. Digital. <https://bonotom.com/flipbook/2096/#2096/122>.

Wilson, Amy L., Michael Goldstein and Chris J. Dent. "Varying Coefficient Models and Design Choice for Bayes Linear Emulation of Complex Computer Models with Limited Model Evaluations." *SIAM/ASA Journal on Uncertainty Quantification* 10.1 (2022): 350-378. <https://doi.org/10.1137/20M1318560>.

Xiu, D. and G. E. Karniadakis. "The Wiener-Askey polynomial chaos for stochastic differential equations." *SIAM J. Sci. Comput.* 24.2 (2002): 619-644.

Zarghami, Seyed Ashkan, Indra Gunawan and Frank Schultmann. "Exact reliability evaluation of infrastructure networks using graph theory." *Qual. Reliab. Engng. Int.* 36 (2020): 498-510. <https://doi.org/10.1002/qre.2574>.