ACQUISITION INNOVATION RESEARCH CENTER

Test and Evaluation Methods for Middle-Tier Acquisition (MTA) – Base Year

EXECUTIVE SUMMARY AND REPORT

SEPTEMBER 2023

PRINCIPAL INVESTIGATOR:
Dr. Laura Freeman, *Virginia Tech National Security Institute*

CO-PRINCIPAL INVESTIGATOR:
Mr. Geoff Kerr, *Virginia Tech National Security Institute*

VIRGINIA TECH

SPONSORS:

Dr. Sandra Hobson, Deputy Director, Strategic Initiatives, Policy, and Emerging Technologies, Office of the Director, Operational Test and Evaluation (DOT&E)

Dr. Kristen Alexander, Chief Learning and Artificial Intelligence Officer, DOT&E

Dr. Jeremy Werner, Chief Scientist, DOT&E

Dr. Tyler Englestad, Emerging Technologies Advisor, DOT&E

Mr. Paul Lowe, Executive Officer for DOT&E Strategic Initiatives, Policy and Emerging Technologies, DOT&E

RESEARCH TEAM

NAME	ORG.	LABOR CATEGORY
Laura Freeman	Virginia Tech National Security Institute (VTNSI)	Principal Investigator
Geoffrey Kerr	VTNSI	Co-Principal Investigator
Orlando Florez	VTNSI	Associate Director for Program Management
Sanglin Chang	VTNSI	Project Manager
Jewel Pike	VTNSI	Project Manager
Andrew Wapperom	VTNSU	Undergraduate Research Intern
Anika Thatavarthy	VTNSI	Undergraduate Research Intern
Brian Lee	VTNSI	Research Data Analyst / Undergraduate Research Intern
Christian Frederiksen	VTNSI	Graduate Research Assistant
Daniel Sobien	VTNSI	Research Associate
Diya Khilnani	VTNSI	Undergraduate Research Assistant
Emma Meno	VTNSI	Research Associate
Erin Lanus	VTNSI	Research Assistant Professor
Jaganmohan Chandrasekaran	VTNSI	Postdoctoral Associate
John Gilbert	VTNSI	Research Assistant Professor
Justin Kauffman	VTNSI	Research Assistant Professor
Justin Krometis	VTNSI	Research Assistant Professor
Kelli Esser	VTNSI	Associate Director, Intelligent Systems Division
Kenny Becerra	VTNSI	Undergraduate Research Intern
Kyle Risher	VTNSI	Undergraduate Research Intern
Luis Pol	VTNSI	Graduate Research Assistant
Nicola McCarthy	VTNSI	Research Assistant Professor
Padmaksha Roy	VTNSI	Graduate Research Assistant
Paul Hess	VTNSI	Adjunct Professor / Senior Advisor, Naval Engineering
Paul Wach	VTNSI	Research Assistant Professor
Peter Beling	VTNSI	Director, Intelligent Systems Division
Tyler Cody	VTNSI	Research Assistant Professor
Victoria Sieck	Air Force Institute of Technology (AFIT), STAT Center of Excellence (COE)	Deputy Director, STAT COE
Corey Thrush	AFIT/STAT COE	CTR – STAT Expert
Corinne Weeks	AFIT/STAT COE	CTR – STAT Expert

Cory Natoli	AFIT/STAT COE	CTR – STAT Expert
Kyle Provost	AFIT/STAT COE	CTR – STAT Expert
Steve Oimoen	AFIT/STAT COE	Applied Research Lead
Wayne Adams	AFIT/STAT COE	CRT – STAT Expert
Douglas Montgomery	Arizona State University (ASU)	Regent Professor
Dustin Taylor	ASU	Graduate Student
Rong Pan	ASU	Professor, Program Chair of Data Science, Analytics, and Engineering
Bruce Einfalt	Penn State Applied Research Laboratory (ARL)	Division Autonomous & Systems Head
David Narehood	Penn State ARL	Department Head, Research Engineer
Andrew Shaffer	Penn State ARL	Research and Development Engineer
Marsha (Marcy) Perini	Penn State ARL	Research and Development Engineer
Owen Cramp	Penn State ARL	Research and Development Engineer
Sheri Martinelli	Penn State ARL	Research and Development Engineer
Jitesh Panchal	Purdue University	Professor of Mechanical Engineering
Karen Marais	Purdue University	Professor, Associate Head for Undergraduate Education
Atharva Sonanis	Purdue University	Graduate Research Assistant
Robert Seif	Purdue University	Undergraduate Research Assistant
Maegen Nix	Virginia Tech Applied Research Corporation (VT-ARC)	Director of Information Sciences & Analytics Division
Christina Houfek	VT-ARC	Lead Project Manager
Daniel Wolodkin	VT-ARC	Staff Data Scientist
Grant Beanblossom	VT-ARC	Lead Data Scientist
Kobie Marsh	VT-ARC	Staff Data Scientist
Timothy Crone	VT ARC	Program Manager

ACKNOWLEDGEMENTS

This report is a culmination of research activities conducted across 10 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) in the context of emerging technologies, rapidly changing threat landscape, and the need to accelerate T&E processes to meet the needs of warfighters.

ACRONYMS AND ABBREVIATIONS

ACM	Association for Computing Machinery
AFIT	Air Force Institute of Technology
AI	Artificial Intelligence
AIRC	Acquisition Innovation Research Center
ARL	Applied Research Laboratory
ARL-PSU	Applied Research Laboratory - Pennsylvania State University
ASA	American Statistical Association
COI	Community of Interest
CSER	Conference on Systems Engineering Research
CT	Contractor Test
DATAWorks	Defense and Aerospace Test and Analysis Workshop
DC	District of Columbia
DCL	Detection, Classification, and Localization
DE	Digital Engineering
DoD	Department of Defense
DoDI	Department of Defense Instruction
DOE	Design of Experiments
DOT&E	Director, Operational Test and Evaluation
DT	Developmental Testing
DTE&A	Office of the Director for Development Test, Evaluation, and Assessments
ECE	Electrical and Computer Engineering
FGSM	Fast Gradient Sign Method
GW	George Washington University
HPCC	DoD High Performance Computing Centers
HPP	Homogeneous Poisson Process
HW/SW	Hardware/Software
ICST	International Conference on Software Testing
ICSTW	International Conference on Software Testing, Verification and Validation Workshops
ID	Identify
IDA	Institute for Defense Analyses
IDSK	Integrated Decision Support Key
IEEE	Institute of Electrical and Electronic Engineers
IMS	Institute of Mathematical Statistics
I-Plan	Implementation Plan
IRB	Institutional Review Board
ISO 26262	International Organization for Standardization - Road vehicles - functional safety
IT	Information Technology
ITEA	International Test and Evaluation Association
IWCT	International Workshop on Combinatorial Testing
JCIDS	Joint Capabilities Integration and Development System
JTC	Joint Test Concept
JWC	Joint Warfighting Concept

LFT&E	Live Fire Test and Evaluation
LVC	Live, Virtual, Constructive
M&S	Modeling & Simulation
MB	Model-based
MBSE	Model-Based Systems Engineering
MBTEMP	Model-Based Test and Evaluation Master Plan
MCMC	Marvok Chain Monte Carlo
ML	Machine Learning
MOR	Military Operations Research
MORS	Military Operations Research Society
MORSS	Military Operations Research Society Symposium
MSM	Mistake Structure Matrix
NASA	National Aeronautics and Space Administration
NHPP	nonhomogeneous Poisson process
NJ	New Jersey
NPP	Normalized Power Prior
NPS	Naval Postgraduate School
NSI	National Security Institute
NY	New York
OT	Operational Testing
OT&E	Operational Test & Evaluation
OUSD	Office of the Undersecretary of Defense
OUSD (R&E)	Office of the Undersecretary of Defense for Research and Engineering
PGD	Projected Gradient Descent
PI	Principal Investigator
PSU	Pennsylvania State University
RL	Reinforcement Learning
SAE G-34	Society of Automotive Engineers Artificial Intelligence in Aviation
SAS	Synthetic Aperture Sonar
SDNS	Statistics in Defense and National Security
SEPTAR	Systems Engineering Process to Test Artificial Intelligence (AI) Right
SERC	Systems Engineering Research Center
SERDP	Strategic Environmental Research and Development Program
SIPET	Strategic Initiatives, Policy, and Emerging Technologies Directorate
SME	Subject Matter Expert
SoA	State-of-the-Art
STAT COE	Scientific Test & Analysis Techniques Center of Excellence
SVSS	Sediment Volume Search Sonar
SysML	Systems modeling language
T&E	Test and Evaluation
TEA-LEAF	Test, Evaluation, and Assurance of Learning Framework
TEMP	Test and Evaluation Master Plan
TRMC	Test Resource Management Center

t-SNE	t-Distributed Stochastic Neighbour Embedding
UL 4600	Underwriter Laboratories Standard for Safety for the Evaluation of Autonomous Products
UQ	Uncertainty Quantification
US	United States
USA	United States of America
USAF	United States Air Force
USC	University of Southern California
UXO	Unexploded Ordinance
V&V	Verification and Validation
VA	Virginia
VT	Virginia Tech
VT-ARC	Virginia Tech Applied Research Corporation
VTNSI	Virginia Tech National Security Institute
VV&A	Verification, Validation, and Accreditation
VVUQ	Verification, Validation, and Uncertainty Quantification
VVUQ&A	Verification, Validation, Uncertainty Quantification, and Accreditation

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The Department of Defense (DoD) 2022 National Defense Strategy has realized the need for advanced technology and more rapid development and fielding of that technology to sustain dominance against peer/near-peer threats. In support of these objectives, the Director, Operational Test and Evaluation (DOT&E) engaged the Acquisition Innovation Research Center (AIRC) University Affiliated Research Center (UARC) in a multi-year contract to advance test and evaluation (T&E) methods within the DoD, and the research is concluding the initial years' worth of effort. The multi-university AIRC research team that has partnered with DOT&E has focused research efforts over the past year in support of the DOT&E Implementation Plan. Though established as two contracts under WRT-1070 Test and Evaluation Methods for Middle Tier Acquisition (MTA) and WRT-1071 Digital Transformation (DT) in Test and Evaluation for AI/ML, Autonomous, and Continuously Evolving System, the research team has united the efforts for efficiency and for alignment to the Implementation Plan (I-Plan). The research team focused on, under advisement from DOT&E technical leadership, supporting three of the five key pillars within the Implementation Plan.

- 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

As the greater research team attacked this multifaceted space with focused efforts, there were commonalities in the basic approach. First, the team researched the current best practices, challenges, and strategies within industry and the government through detailed literature reviews and subject matter expert engagements. Second, the team summarized their findings and initial conclusions in a series of publications and presentations, as seen in Figure 1 below. Finally, the team explored test cases to evaluate the initial findings and recommendations. In the upcoming year, the research team will expand on these foundational efforts.

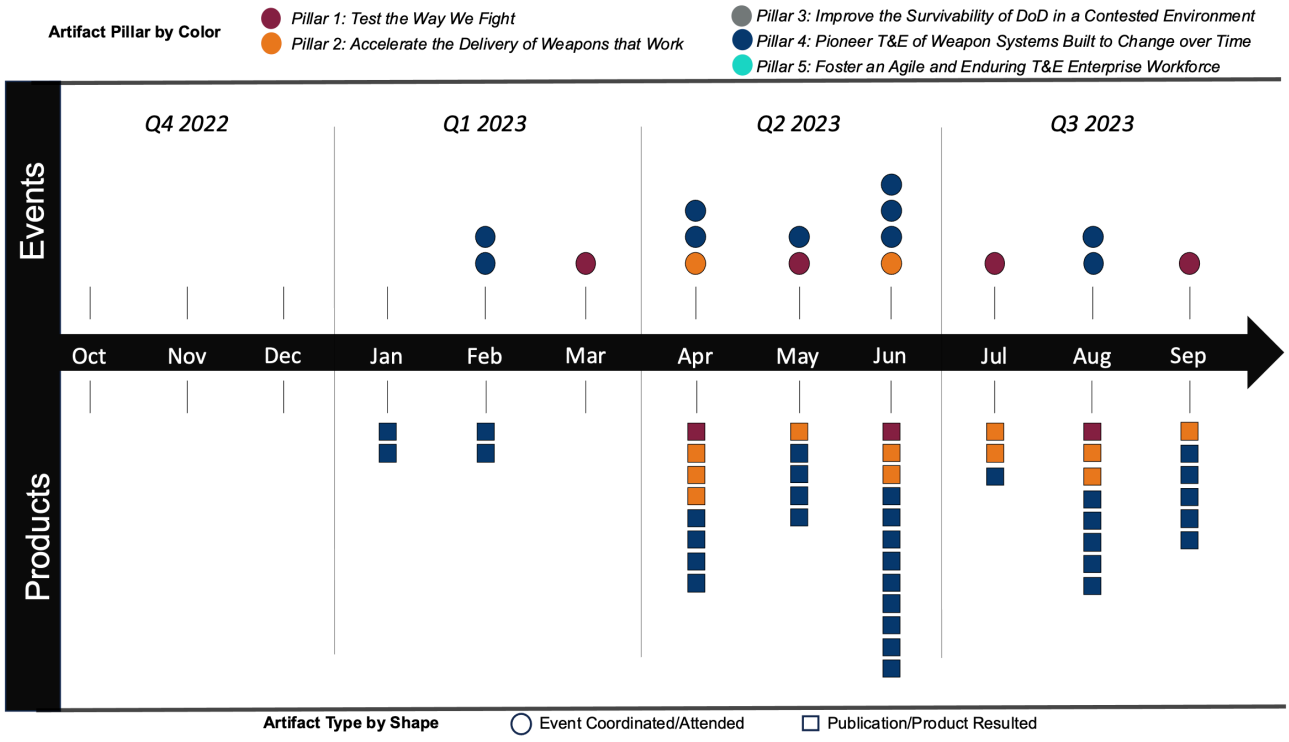


Figure 1: AIRC Year 1 Resulting Products and Event Participation

In support of Pillar 1, the AIRC research team focused on maturing evaluation methodologies for Joint Warfighting Concepts with the development of a fundamental Joint Test Concept (JTC). The team established a broad Community of Interest (COI) to explore numerous aspects and challenges around the test and evaluation of Joint Warfighting Concepts. The team conducted a three-phased research approach, each of which concluded in a workshop that first shaped and bound the focus of the research team, developed initial JTC priorities, and concluded with a tabletop simulation of joint operation assessment.

In support of Pillar 2, the research team explored current threats to DoD data and how these threats can be mitigated. The team responded with a Data Security paper that expands upon the findings. In addition to the data security aspects of accelerating Weapons that Work, the research team developed Bayesian-based approaches to leverage contractor testing, development testing, live fire testing, modeling and simulation results, and operational test to ensure more rapid test and evaluation of weapon systems to expedite fielding of warfighting capabilities. In the future efforts, the research team will be working on a Model Based Test and Evaluation Plan (MB TEMP) and supporting the advancement of the Integrated Decision Support Key (IDSK) beyond the initial efforts described in the Digital Engineering for T&E section of this report.

In support of Pillar 4, the research team investigated improved methods for test and evaluation of Artificial Intelligence (AI) /Machine Learning (ML) -enabled DoD systems. The team has developed recommendations on how to ensure AI/ML systems have proper training data while continuing to evaluate and inform system performance while in use. The team further explored methods to evaluate AI/ML systems when the T&E community has limited or no detailed understanding of the underlying AI/ML algorithms within a system. Also, in support of Pillar 4, the multi-university team helped further the application of digital engineering practices in support of Test and Evaluation. With a realistic look at tooling challenges and current cultural challenges, the team made practical recommendations on the use of Model Based Systems Engineering (MBSE) methods for test planning and execution, in addition to digital linkage from mission requirements through operational assessment. Lastly, Pillar 4 efforts also included maturing a framework to automate security Penetration Testing. After conducting cyber security research, the team developed early test software methods that enable cyber physical system penetration testing to be automated in concert with a continuous innovation, continuous deployment product development environment. In the next year, the team intends to test these capabilities with case study DoD products.

Finally, in support of DOT&E emerging priorities, the research team supported numerous tasks that included providing policy reviews and recommendations, offering workforce planning concepts, developing exemplar IDSK, and delivering technical support to T&E meetings, workshops, and conferences. The research impact has been wide reaching. The AIRC team hosted 5 different workshops (2 under WRT-1070), delivered 14 presentations/webinars/briefings (10 under WRT-1070), authored 11 reports/publications (8 under WRT-1070), and developed 3 test related tools/applications (2 under WRT-1070).

Based on the foundational work performed in this initial year, the AIRC research team looks forward to building on the year 1 recommendations and tooling by performing case study evaluations on actual DoD programs of record and on conceptual programs. The team also intends to introduce workforce development efforts in the follow-on year in support of DOT&E's fifth pillar, Foster an Agile and Enduring T&E Enterprise Workforce.

BACKGROUND

The Director, Operational Test and Evaluation (DOT&E) technical staff has engaged with the AIRC UARC to advance tooling and processes for executing test and evaluation on 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. Two contracts have been issued and are being worked together in a base plus option year contract arrangement. This final report addresses the accomplishments of the base year effort as the research team focuses on building on this foundation for the research to be performed in the upcoming option year.

The two base contracts were originally planned to focus specifically on Middle Tier Acquisition 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 synergy of the contracted tasks and to best align with the DOT&E I-Plan. Please see Table 1 below for a mapping of the original proposed work to the lines of research and DOT&E Implementation Plan (I-Plan) pillars.

			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		

Table 1: Proposal Tasks to I-Plan Mapping

ALIGNMENT TO I-PLAN PILLARS

Pillar 1: Test the Way We Fight

JOINT TEST CONCEPTS

Research Objective

The AIRC research team's objective was to develop a pilot Joint Test Concept (JTC) that supports Pillar 1 of the DOT&E I-Plan: Test the Way We Fight. This included identifying existing processes or articulating new processes, resources, and capabilities to evaluate joint warfighting capabilities and mission threads, kill webs, and system-of-systems performance. The intention was to provide a T&E framework to assess systems in an accurate representation of the joint, distributed, non-contiguous, multi-domain operating environment.

Methods

Following an extensive review of strategic guidance and documents related to the Joint Warfare Concept (JWC), AIRC researchers from the VT-ARC and VTNSI designed a three-phase plan to develop the pilot JTC.

Phase one included:

- Literature review
- Study framework design
- Workshop I (design and execution)

Phase two included:

- Workshop I report
- Workshop II (design and execution)

Phase three included:

- Workshop II report
- Workshop III (design and execution)
- Drafting and delivering a pilot JTC

The series of workshops depicted in Figure 2 allowed for a flexible, agile-inspired concept development. The Joint Test Concept Workshop III Report: JTC Pilot Final Report can be referenced for detailed descriptions of the workshop outcomes and the JTC Pilot framework.

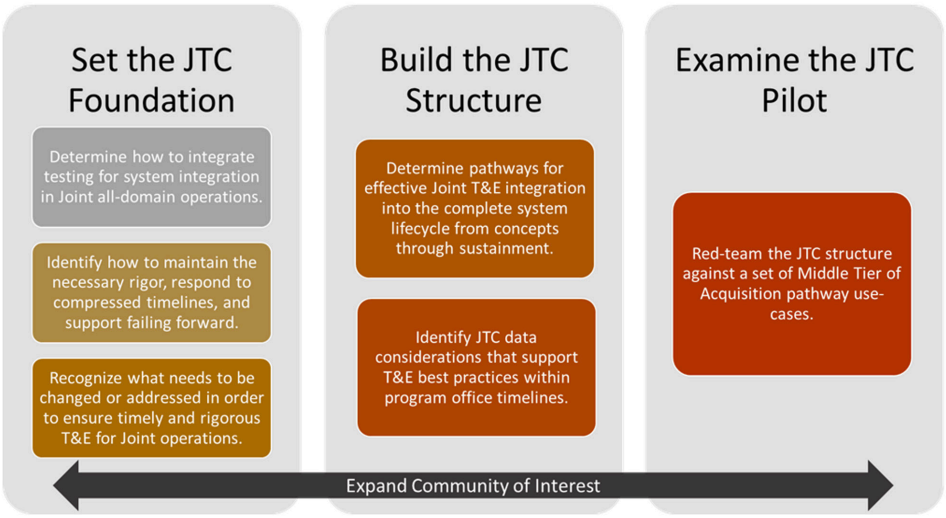


Figure 2: JTC Workshop Focus Areas

Findings

Over the series of three workshops, the Community of Interest (COI) expanded from 10 to 26 organizations. Representatives from six COI organizations provided more active support by providing workshop design input and report feedback. Additionally, COI members invited the VT-ARC study team to present and participate in multiple T&E-related workshops. While not listed as deliverable, there was a notable growth in interest and support for the JTC development.

The COI determined that while the existing T&E structure is executed in a linear and unidirectional pathway where an assessment failure can be catastrophic to the program, the complexity of the Joint operating environment, rapid innovation timelines, and emergent capability gaps require a shift away from a binary pass/fail construct and a shift toward a non-linear and multi-directional workflow that assesses appropriate field-ability within a complex kill web and supports mission engineering threat construct.

Analysis of the workshop outcomes and smaller COI working group feedback resulted in a Joint Test Concept framework pilot. The JTC pilot framework consists of a foundation comprised of the three JTC layers and a structure featuring eight critical elements (organization, training and education, authorities and policy, resourcing, end-to-end lifecycle (E2EL) continuity flow, data strategy, the T&E environment, and the capability or system specific E2EL T&E strategy). The JTC foundation is based upon the assumption that the current practice of assessing systems will continue with the current construct of contractor testing, developmental testing, and operational testing (live, virtual, and constructive-LVC) although structural changes may occur in the future. Therefore, the JTC pilot was founded on the principle that it must work both within and, in some cases, despite the current system and be flexible enough to absorb changes regardless of whether the changes are JTC-recommended or otherwise. To accomplish this flexibility, the JTC does not use the current vernacular (e.g., developmental testing (DT), operational testing (OT), contractor test (CT)) but instead recognizes three overlapping layers (system performance, capability immersion, and joint capability demonstration) that ensure the system meets performance requirements in isolation, within the pre-defined system of systems utilizing a mission engineering thread construct, and in a Joint multi-domain environment with realistic adversary representations and accurate kill webs.

The current acquisition structure does not naturally flow from a service-specific perspective to a Joint interoperability requirements perspective. This is in part due to funding (the amount and the type) and part to existing authorities and policies. The shift to a Joint, distributed, multi-domain, systems-of-systems approach to warfare requires a relook at the existing organizational structure to identify the most efficient and effective way to implement the Joint Test Concept.

The COI concluded that it is necessary to develop a Joint-Test & Evaluation Strategy Team (J-TEST) office that has the requisite authorities and funding to ensure Joint T&E is executed to inform system fielding, modernization decisions, and system integration into plans. The COI determined that the J-TEST should foster radical innovative thinking (or creative, critical, and open thinking) to ensure J-TEST strategies are efficient, resilient, and able to integrate updates to technology, analytic methodologies, security, and requirements. The COI envisioned six separate but overlapping J-TEST lines of effort, each featuring a director: environment, security, strategy, integration, communication, and information.

The rise of Peer/Near-Peer (P/N-P) actors suggest the Joint force could be contested in all domains during the execution of distributed, potentially non-contiguous, combat operations. Ensuring the advantage will stretch traditional T&E capabilities further than ever before. T&E must be re-imagined, placing increased emphasis on the operational and mission context in which the system under test is expected to perform. Therefore, the way in which we think about validation and how we test in support of validation must result in high levels of confidence in execution of complex kill webs, featuring multiple (Joint) systems of systems. As such, the evaluation of a system under test must go beyond discrete T&E blocks within the program lifecycle. It must be carried out across both capability and system lifecycles within the context of expected contributions service and Joint effectiveness. The JTC pilot provides a pathway toward implementation of service and Joint E2EL T&E both within the existing constraints and restraints and is flexible enough to adapt to shifting or emergent policies, authorities, requirements, and capabilities.

Recommendations
Evolution toward a JTC must come incrementally because changes, in the short-term, to existing T&E processes are unlikely. To achieve a JTC that can be implemented within the existing acquisition and T&E constraints and restraints but is flexible and scalable enough to adapt to the dynamic T&E environment, the focus for follow-on work must focus on developing JTC planning and implementation guidelines. Resulting artifacts would illustrate JTC implementation within the existing acquisition and T&E construct and include recommendations for changes that would enhance adoption.

Outcomes could be achieved through additional workshops and/or small COI working groups covering topics including the development of an:

- Overarching, dynamic, and realistic JTC immersion scenario
- E2EL T&E strategy execution framework that could be adopted across the entire Defense ecosystem; this framework will leverage existing government-owned planning and strategy models
- Immediate, mid, and long-range implementation pathways (including recommended policy and organizational changes) that benefit all stakeholders

Ultimately, this JTC team needs to help DOT&E create quick wins that will build momentum to drive the bigger change that is needed across existing policies and structures within the services. Capability portfolio management must transform and adapt to the way we will fight and win in a world that must be Joint and interoperable.

Pillar 2: Accelerate the Delivery of Weapons that Work

DATA SECURITY – T&E FOR DATA SECURITY

Research Objective
The AIRC research team's objective was to communicate industry best practices for securing data at rest and in-transit or computation by writing a survey paper and best practices brief. The literature review includes traditional cybersecurity/information assurance approaches that apply to machine learning (ML) development and operations pipelines and identifies new challenges and threats due to the deployment of ML.

Methods
The team's research methodology included:

1. Reading the DOT&E I-Plan to understand how data security fits into the strategic pillars. The I-Plan emphasizes "implementing industry best practices for secure authentication, access management, encryption, monitoring and protection of T&E data at rest, in transit, and in use," so these became the topics to address.
2. Forming an outline for the data security paper by identifying a list of topics pertinent to data security, specifically data at rest, in transit, and in computation. This list includes:

- Security properties (confidentiality, integrity, availability)
- Cryptography
- Authentication
- Authorization
- Access control
- Zero trust architectures
- Homomorphic encryption
- Cloud computing
- Privacy (k-anonymity, differential privacy)
- Privacy in machine learning
- Federated learning
- Attacks in machine learning (data poisoning, model inversion)

3. Identifying academic papers, government reports, and websites that comprehensively cover the list of topics and then reading and summarizing the material into a survey paper. Of the sources identified, the team cited 117 sources in the survey paper and reviewed additional sources that were not included due to space constraints.
4. Condensing the survey paper material into a concise best practice brief that included the basics on cybersecurity and new challenges for data security presented by ML.

Findings
This work resulted in a 29-page survey paper that summarizes the topics identified explicitly around protecting data at rest, in transit, or in computation. The team addressed newer methods in cybersecurity such as homomorphic encryption, attribute-based access control, multi-factor authentication, and zero trust architectures as well as the new challenge of ML attacks alongside established or older techniques such as digital signatures and k-anonymity; the review focused on data that already exists in an information system.

For ML, this occurs in the middle of the pipeline where models are trained or tested. The team did not address the additional security concerns that exist at the endpoints, such as when data is collected/created or how to securely wipe data from a model in "unlearning." Additionally, the team focused on data in more obvious information systems, such as databases, flowing across communication networks, or cloud platforms. Data also exists in transit and computation within cyber-physical systems and the ML use cases here are increasingly evident.

Recommendations
The field of cybersecurity has established data protection practices in information systems. No protection mechanism is perfect or can defend against every attacker, so best practices already include defense-in-depth, zero trust architectures, modern cryptography and access control mechanisms, user education, and data redundancy to achieve security properties of confidentiality, integrity, and availability. Increasingly, ML engineers and data scientists are employed for data management, and the training for these positions does not include understanding of cybersecurity practices for information assurance. Thus, ML engineers and data scientists should receive a baseline education to avoid creating vulnerabilities that an attacker could exploit. One example of this is not relying on the cloud service provider to encrypt the data owned by the organization but maintaining those keys locally and only sending encrypted data to the cloud. Homomorphic encryption is another option for confidentiality against the cloud service provider, but it may not be as easy to utilize.

Additionally, the subfield of adversarial ML is nascent and defending against ML attacks can require knowledge of ML itself. Cybersecurity professionals tasked with defending an ML-enabled system need to be aware of attacks vectors outside of the traditional information system. For example, even when training data is encrypted inside a database with appropriate access control mechanisms and the training code is trusted, training data may be extracted from a model via membership inference attacks. Similarly, even if a model's weights are encrypted, a surrogate model can be constructed via model inversion attacks.

DATA SECURITY – AI SYSTEM PERFORMANCE

Research Objective
The research objective was to evaluate alternative formats for communicating Artificial Intelligence (AI) model behavior to non-AI-expert decision-makers. The team was particularly interested in the decision-makers' ability to identify model errors and assess their importance in their specific application. This required identifying standard information formats, developing new ones, and designing a use-study to evaluate how the information formats impact decision-making.

Methods
As part of the research methodology, the team:

1. Conducted a literature review on AI explainability methods and then extracted and updated three candidate information formats for evaluation.

2. Designed a user experiment to evaluate them. This involved training a baseline ResNet50 image classifier pretrained on ImageNet-1K and imputing known errors to simulate reasonable model deficiencies. Next, the team created alternative information formats to communicate the behavior of each model and presented them to (a pilot set of) representative decision makers in a user experiment. The user experiment set the stage for users to prefer one type of error over another. This enabled a preliminary assessment of which information formats allowed users to identify and evaluate the model errors relevant to their use case.
3. Conducted a preliminary analysis of these results.
4. Is working on recruiting a large population for the full experiment.

Findings
The effort on the AI Systems Performance task began in June 2023, so the team's results are preliminary. To date, the team pilot tested two information formats: the t-Distributed Stochastic Neighbour Embedding (t-SNE) and the Mistake Structure Matrix (MSM). Despite a small sample, the MSM clearly outperformed the t-SNE. Results are below in Figure 3. This gives confidence that a full survey will provide insight into the difference between information formats in supporting decision-making.

As part of the pilot, the team also collected qualitative responses about the decision-making process. Participants responded that regardless of which visualization they were given, they used the numerical metrics provided with the visualization over the visual. What is interesting is that the numerical metrics were identical in both treatments, yet the participants chose the correct model more often when shown the MSM format. These results suggest that the MSM help non-AI experts understand model behavior and that the participants overestimated their numerical literacy.

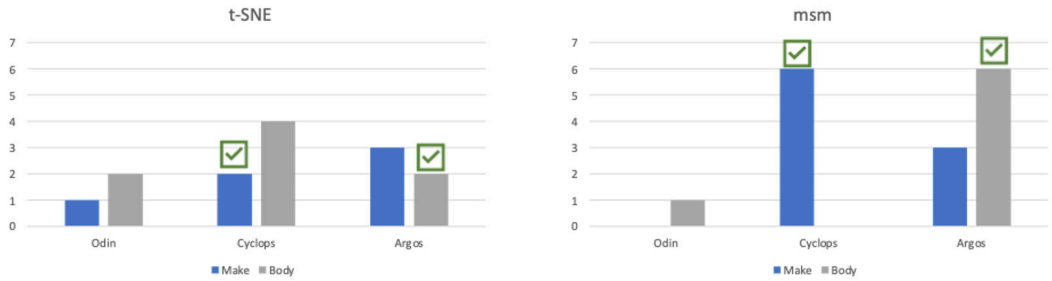


Figure 3: AI System Performance - Pilot Results (Correct model is the green check box)

Recommendations
Currently, numerical values including accuracy and precision recall are the standard information format presented to decision-makers. These measures hide key aspects of model behavior and rely on expert knowledge of underlying models to interpret the model's outputs. Results from further research will enable the team to recommend to the DoD what information formats are helpful and provide insight into how much AI literacy is required for personnel in the role of selecting AI-embedded systems.

INTEGRATED TESTING – BAYESIAN INFERENCE TECHNIQUES

Research Objective
The research team achieved their research objective by collaborating among researchers from Virginia Tech, the Air Force Institute of Technology (AFIT), and the Scientific Test and Analysis Techniques Center of Excellence (STAT COE). The primary objective has been to develop and illustrate Bayesian inference techniques for leveraging all available data, including potentially dissimilar data from earlier in the program life cycle, to better understand the characteristics of operational systems. Taking advantage of all the available data enabled the team to estimate system parameters more accurately and precisely and use operational testing resources more efficiently.

Methods
Bayesian inference is a statistical technique wherein the unknown is considered to be a random quantity. It involves three key quantities:

1. The *prior* is a probability distribution describing what is known about the parameter before data is collected. Priors can be developed from historical datasets, Subject Matter Expert (SME) input, or by considering natural ranges for parameter (e.g., some parameters are necessarily non-negative). The development of priors was the subject of Dr. Sieck of AFIT's Military Operations Research Society (MORS) Symposium 2022 presentation that won Honorable Mention for the Barchi Prize for best presentation. It is also the subject of the paper that Dr. Sieck and Dr. Krometis of Virginia Tech submitted to the Military Operations Research (MOR) Journal as part of this project.

2. The *likelihood* is a function characterizing how collected data relates to the unknown parameter being estimated. The likelihood is similar to traditional measures in that it will be high where parameters match the data well.
3. The *posterior* is a probability distribution describing the understanding of the parameter while incorporating both the data and the prior. This is the output of Bayesian inference. It will have high probability for parameter values that match both the prior and likelihood well and be lower where the parameter exhibits significant mismatch for one or the other. The posterior distribution typically cannot be computed analytically and instead must be numerically approximated, such as via Markov Chain Monte Carlo (MCMC) methods.

Bayesian inference is like human reasoning in many ways, where an initial understanding of the environment is developed and then updated systematically as data is collected. Except in very particular circumstances, a parameter estimate from Bayesian inference will converge to the traditional frequentist estimate as the number of observations grows large.

Findings

The Integrated Testing effort, thus far, has had two primary thrusts, each of which can be associated with a phase of testing. The first is *Post Test*, which has been the focus of a majority of the team's effort to date. For this thrust, the team considered cases where testing was completed and took all the available data into account to make the best performance estimates.

The second thrust is *During Test*, where the team looked at partial testing data and sought to determine whether the system had been tested enough to be confident in the outcome of the test. A key exemplar throughout has been the Stryker family of vehicles, for which data from both developmental testing (DT) and operational testing (OT) was publicly released as part of an analysis in 2015 (Dickinson, Freeman and Simpson).

The *Post Test* effort considered three different approaches for estimating operational behavior using both DT and OT data:

- ***Hierarchical models.*** This approach involves fitting a single model to all of the available data. This allows the modeler to incorporate all assumptions and to fix structure between testing phases – e.g., by enforcing that reliability estimates be lower in OT than in DT. This is a reasonable approach to take when considering all data after testing; however, this approach lacks the desirable property of accumulating knowledge in that if additional data were to be collected, then the whole model would need to be refit. Hierarchical models were the subject of a 2015 paper on the Stryker family of vehicles (Dickinson, Freeman and Simpson).
- ***Informative Priors with Fixed Down-weighting.*** This model involves conducting traditional Bayesian inference at each testing phase by using the initial prior distribution in the first phase and building priors from subsequent phases using the results from the inferences from the previous phase. This mimics, to some extent, how a human might interpret testing data as a system evolves. At some boundaries between test phases, the data from a previous phase might be "down-weighted" if the practitioner believes that because the system or environment has changed, the behavior might also be different.
- ***Normalized Power Priors (NPPs).*** NPPs are an informative prior technique where the down-weighting is treated as a random parameter and is estimated with the rest of the model parameters via Bayesian inference. In doing so, this technique involves fewer assumptions on behalf of the practitioner, but the results come at the expense of the posterior being significantly more computationally costly to estimate.

Dr. Krometis gave a talk at MORS Symposium 2023 comparing these three methods as applied to the Stryker dataset. In the coming months, the Integrated Testing team intends to submit this content to an academic journal and to develop a STAT COE Best Practices guide outlining the ideas and providing a detailed software implementation. The team completed assembling R notebooks to illustrate the application of each of these methods and is working on developing an R Shiny app to communicate the ideas via a more graphical interface. The paper on priors that Dr. Sieck and Dr. Krometis submitted to the MOR Journal used informative priors with fixed down-weighting applied to a single Stryker vehicle variant as a numerical example; Dr. Sieck also presented this content at MORS Symposium 2023 as part of her presentation's selection as a Barchi prize Honorable Mention.

The *During Test* analysis has thus far focused on identifying how much testing could be saved by conducting Bayesian inference during testing to determine whether the system has already reached or failed to reach the programmatic criteria with a sufficient confidence. This approach could help ensure that test resources are used only when more information can be learned about a system. Stryker has again provided an exemplar for this analysis with the outcome of indicating how many miles of testing could have been saved for each vehicle variant. Dr. Sieck presented results at the Institute of Mathematical Statistics (IMS)/ American Statistical Association (ASA) Spring Research Conference in May 2023 and intends to compile and submit them to an academic journal later this year. The research team is also starting to look at a design of experiments analyses to identify choices of tests that could yield the most information about system characteristics, which is another way of ensuring that limited test resources are used efficiently.

Finally, as an outreach and education effort, the STAT COE team led a full day short course titled "Applied Bayesian Methods for Test Planning and Evaluation" at the Defense and Aerospace Test and Analysis Workshop (DATAWorks) in April 2023. Course topics included the difference between Bayesian statistics and frequentist statistics, how to select priors and their effect on the posterior, the use of software to obtain a posterior in two different examples associated with diagnostics checks, and how to interpret several analyses in a Bayesian framework. Approximately 75 people attended both in person and remotely.

Recommendations
The papers, presentations, and training courses generated by this effort have increased awareness of Bayesian methods as applied to test and evaluation and have illustrated their potential for application to DoD problems. They have also helped establish a software base that can be leveraged in future efforts. Future project goals include:

- Continuing to develop methods for Bayesian approaches to integrated test and evaluation, especially for programmatic datasets that are more complex than Stryker
- Applying more During Testing approaches, like design of experiments, to illustrate how resources can be allocated efficiently
- Continuing to build community around the analytical approaches via outreach and training

INTEGRATED TESTING – DESIGN OF EXPERIMENTS (DOE) TECHNIQUES

Research Objective
Design of Experiments (DOE) is a systematic and efficient approach used in various science and engineering fields to optimize processes, improve product quality, and gather meaningful insights from experiments. The primary goal of DOE is to maximize the information gained from a limited number of experiments, thereby reducing cost, time, and resources required for experimentation. The research team investigated the operational testing of repairable systems and used DOE techniques to better design and implement testing plans such that the researchers could obtain the desired amount of information from the tests, such as the total testing time, number of systems, and constraints on testing equipment, even with limited resources.

Methods
DOE techniques involve carefully planning and controlling the experimental conditions to systematically explore how different factors or variables affect the outcome or response of interest. For a repairable system, these factors include system design variables and system operation variables, particularly the stress variables of the system's operating environment such as temperature, vibration, loads, and work cycles. Additional factors to consider include the number of testing systems, testing duration, failure mode, and repair effectiveness. Thus, designing test plans for repairable systems is much more complicated than designing plans for non-repairable systems.

Conventional DOE techniques consist of factorial designs and orthogonal designs where each factor can be set on several different levels and an experimental run is executed on a specific factor-level combination such that all main factor effects and some factor interaction effects can be distinctly identified and estimated by analyzing experimental outcomes without confounding. Furthermore, common DOE strategies also include:

- **Randomization:** Randomization is used to eliminate bias and ensure that the experimental results are not influenced by external factors. It involves random assignment of experimental units to different treatment conditions.
- **Replication:** Running experiments multiple times helps to account for variability and assess the consistency of results. Replication increases the reliability of the findings.
- **Blocking:** Blocking involves dividing the experimental units into homogeneous groups or blocks based on certain characteristics. This helps to control sources of variation that are not of primary interest.

After obtaining all experimental outcomes, the team can build a response model, which is typically a regression model that quantifies the relationship between experimental factors and system response. In reliability testing, experimental outcomes are typically the failure counts under different testing conditions, which are discrete variables, thus the regression model to be built may not follow the common Gaussian linear regression model. Instead, the team needs to look for Poisson regression models or other models that are suitable for discrete responses.

Findings
Classical DOE (e.g., full factorial designs, fractional factorial designs, Latin square designs, Box-Behnken designs, etc.) has a long history of success in a variety of industries. Recently, there has been a growing emphasis on optimal designs over classical designs within the design of experiments literature. Count data, such as the number of failures in testing repairable systems, often appear in industry and can be modeled as a Poisson distribution (instead of assuming the response variable is normally distributed). Applications of this model have been studied in medicine, social sciences, and toxicology. The research team explored the D-optimal design criteria as it applies to the Poisson Regression model on repairable systems, with a number of independent variables and under various modeling constraints such as the total time tested at a specific design point with fixed parameters.

For an ongoing paper, the team is investigating multiple regression models of the intensity parameter with operational stress factors. In this investigation, they look at the influence of testing time on designing optimal test plans and discuss various examples in determining the D-optimal design for the Poisson regression model when: 1) accounting for discrete design region in lieu of a continuous, 2) injecting a time factor, and 3) varying the design space. Finally, the team is investigating a Bayesian approach to optimal design to handle the model uncertainty in the regression coefficient.

Recommendations
The Poisson regression model the team introduced in this report is more suitable for modeling failure counts of a repairable system than modeling the classical Gaussian linear model. Based on the Poisson regression model, the team can account for the design space when developing the test plan of repairable systems. Through examples, the team has explored D-optimality in combining the design space, testing time, and the use of prior knowledge (i.e., the Bayesian approach). As they stepped through several homogeneous Poisson process (HPP) models while introducing the time factor at varying stress levels, the team investigated how testing time affects the optimal design. However, these HPP models ignore system degradation and repair effectiveness, thus the team recommends exploring optimal experimental designs for nonhomogeneous Poisson process (NHPP) models as the next step. The team plans to investigate Bayesian optimal designs for other optimality criteria, including A-, E-, and I-optimalities, and compare them with classical designs.

Pillar 4: Pioneer T&E of Weapon Systems Built to Change over Time

VERIFICATION VALIDATION UNCERTAINTY QUANTIFICATION FOR MODELING AND SIMULATION

Research Objective
The research team aimed to investigate agile processes in verification validation uncertainty quantification and accreditation (VVUQ&A) as they relate to modeling and simulation (M&S). They focused on increasing the use of reliable computational tools in operational Test and Evaluation (T&E). This goal aligns with the fourth pillar of the DOT&E Implementation Plan (I-Plan), "T&E of weapon systems built to change over time."

Working towards this aim, the team launched several specific research objectives during Year 1 to:

- Characterize the state-of-the-art (SoA) in physics-based M&S verification and validation (V&V), and uncertainty quantification (UQ)
- Identify technical barriers to the acceptance of M&S alternatives in T&E
- Identify regulatory and non-technical barriers to the acceptance of M&S alternatives in T&E
- Develop a roadmap that prioritizes technologies and standards needed to mitigate identified barriers
- Develop and conduct a M&S VVUQ case-study aimed at evaluating the risk that uncertainty poses to an assessment of operational performance

Methods
The team undertook three research tasks to begin addressing the research objectives listed above: (1) a systematic literature review to characterize the state-of-the-art in M&S V&V, and UQ, (2) an online survey, targeting M&S professionals, developed to better understand the technical and non-technical barriers to adoption, (3) design and resource an underwater acoustics case study aimed at demonstrating the use of UQ in a mission setting.

1. **Systematic Literature Review**

   The team undertook a comprehensive literature review to define the current state-of-the-art in M&S V&V, and UQ with the primary goal of offering recommendations to increase the use of M&S in operational test and evaluation (OT&E). Additionally, the review informed the creation of an online survey designed to gather insights from the T&E community about both technical and non-technical barriers to the use of published guidance on verification, validation, and accreditation (VV&A) of M&S used in operational test and live fire test and evaluation (LFT&E). The literature review drew from a wide variety of sources including peer-reviewed academic journals (Duque et al.; Roy and Oberkampf; Jatale et al.; White et al.; Roache; Richardson et al.; Oberkampf and Smith; Dienstfrey and Boisvert; Babuška et al.; Gregory and Salado), conference proceedings, textbooks (Oberkampf and Roy, Verification and Validation in Scientific Computing), technical reports (Wojton, Avery and Freeman; Cortes, Wong and Cortes-Morales), and DoD policy and guidelines to provide a well-rounded analysis that led to the development of practical recommendations and a targeted survey.

2. **Survey on Verification, Validation, and Uncertainty Quantification Practices in Modeling & Simulation**

   To evaluate the current application of VVUQ within the United States Department of Defense (DoD), the team designed a survey to gather insights from developers, decision-makers, and users. The survey assesses the implementation of verification, validation, uncertainty quantification, and accreditation (VVUQ&A) practices within the DoD's M&S enterprise and identifies obstacles that hinder their full adoption across various organizational M&S operations. Structured to provide a detailed analysis, the survey includes inquiries about the role and trust level of M&S, specific details about the V&V process and its scalability, and potential obstacles for utilizing V&V and UQ. Participants are also asked to rank their top three barriers to adhering to VVUQ&A guidelines. The survey's findings will enable the refinement of the case study and may pave the way for a process that can mitigate barriers within the developers' control, promoting the broader application of these practices within the DoD.

3. **Underwater Acoustics Case Study Design and Resourcing**

   The team will conduct an underwater acoustics case study in year 2, and they will refine the study's methodology and objectives for the case study based on the findings from the literature review and survey conducted in year 1. The case study's objective is to show how uncertainty, particularly epistemic uncertainty, affects an assessment of operational performance in underwater acoustic systems. The case study will focus on the Sediment Volume Search Sonar (SVSS) developed by the Applied Research Laboratory at Pennsylvania State University (ARL-PSU) (Brown, Johnson and Brownstead), which produces three-dimensional synthetic aperture sonar (SAS) images of sediment layers in near-shore environments. The data collected by SVSS is processed for detection, classification, and localization (DCL) of unexploded ordnance (UXO) for remediation efforts by the US DoD Strategic Environmental Research and Development Program (SERDP). Underwater acoustics presents unique challenges for VVUQ due to difficulties in characterizing the environment, misunderstandings about modeling limitations, expensive data collection, and the complexity of model inputs. The SVSS offers an ideal subject for this case study as the data and results can be widely shared, and there is access to software, tools, and validation data (Brown, Johnson and Brownstead) for simulating and processing mission-level data enabling a robust evaluation of the uncertainties involved. An example of the data collected by SVSS is shown in Figure 4 below.

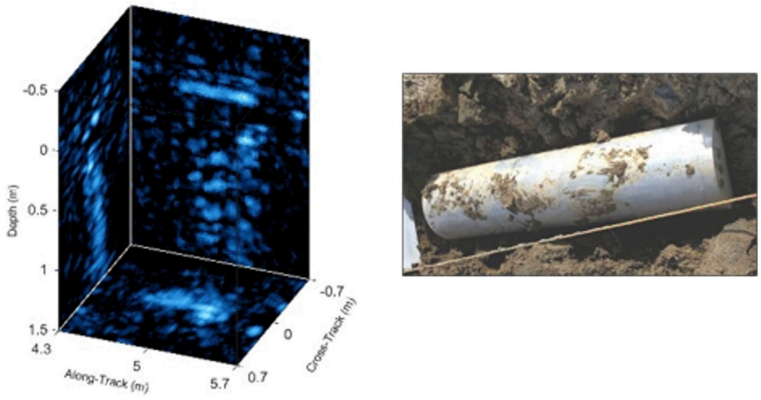


Figure 4: VVUQ - Aluminum cylinder prior to burial 3 cm below water-sediment interface, and maximum intensity projection of the volumetric image produced by SVSS. (Williams and Brown)

Findings

1. **Systematic Literature Review**

   Several key findings emerged from the literature review on VVUQ for M&S. The verification and validation processes, which include standardized methodologies and procedures that demonstrate how a model is implemented correctly and how well it represents real-world phenomena (Oberkampf and Roy, Verification and Validation in Scientific Computing), are relatively well-understood. Model maturity assessment methods, which focus on the gradual development and confidence in a model's reliability through various stages (Cortes, Wong and Cortes-Morales), also form a vital part of the literature. Basic uncertainty quantification techniques are also well-established, particularly when probability distributions are known (Wojton, Avery and Freeman). However, the literature reveals ongoing challenges in areas such as advanced uncertainty quantification, especially in the presence of epistemic uncertainty, and the multifaceted and often subjective process of accreditation. Complexities arise in distinguishing between model errors and uncertainties in model inputs and integrating VVUQ in high-dimensional and intricate models. Additionally, non-technical barriers related to organizational culture and competing interests add further complexities to effective VVUQ. These findings indicate a landscape where fundamental aspects of VVUQ are well-grounded, while more complex and nuanced areas continue to require active research and development.

2. **Survey on Verification, Validation, and Uncertainty Quantification Practices in Modeling & Simulation**

   Before distributing the survey, it was determined that an Institutional Review Board (IRB) exception would be required since the survey's focus is on gathering information about organizations rather than individuals. Virginia Tech successfully obtained this IRB exception and forwarded it to ARL-PSU for their processing, which also successfully received IRB exception. Regrettably, the time taken to apply for the exception and await decisions from both universities led to delays in sending out the survey. As a result, the findings from the survey are now expected to be available in early October.

3. **Underwater Acoustics Case Study Design and Resourcing**

   To prepare for case study implementation, the team has arranged with DOT&E for an allocation on DoD High Performance Computing Centers (HPCC) (DoD High Performance Computing Modernization Program), which will provide adequate resources to perform UQ activities. We have also been coordinating with SVSS project personnel to gain access to the simulation and processing software.

Recommendations
The team recommends the following based on the literature review findings:

- Encourage decision-maker alignment by implementing specific training programs
- Enhance the modeling process by promoting continuous integration of VVUQ
- Create a central repository to house and organize all VVUQ tools for easier access and collaboration
- Streamline the progress of projects by facilitating the sharing of data and benchmarks across different programs
- Offer essential resources and guidance to assist independent verification and validation (V&V) assessments

T&E FOR MULTI-FIDELITY AI MODELS

Research Objective
The research objectives included:

1. Establishing scientific, literature-based best practices for requirements development and agile testing of AI/ML-based systems
2. Creating a framework for multi-fidelity T&E of AI/ML-based systems that:
   - Supports holistic systems evaluation using data & models at multiple levels of development, and different levels of representativeness of the operating environment
   - Establishes how emerging ML-testing techniques can be used throughout the Joint Capabilities Integration and Development System (JCIDS) process, specifically, how to optimally combine white-box, data-box, and black box testing of AI systems
   - Guides the design of future tests
   - Helps quantify the cost and value of model-related intellectual property (e.g., training data, model parameters) in acquisition of AI-based systems.

Methods
The team's research methodology included three key components:

1. **Literature Review**

   The literature review focused on:
   - Recent testing methods for ML techniques such as DeepXplore, DeepTest, and the use of synthetic data
   - T&E frameworks developed specifically for autonomous vehicles
   - Regulations and standards such as UL 4600, SAE G-34, and ISO 26262
2. **Framework development**

   The team created a conceptual framework that integrated JCIDS with multi-fidelity T&E.
3. **Testbed development**

   The team developed an initial software/hardware testbed with the use cases of identifying and tracking people and vehicles. Scenarios with different amounts of training data, different training conditions (soccer vs. combat), different versions of the deep neural network, different hardware quality (e.g., different camera resolution), and different test conditions (e.g., simulation vs. video footage) considered (see Figure 5). The details are documented in (Sonanis).

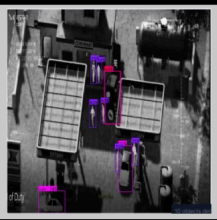
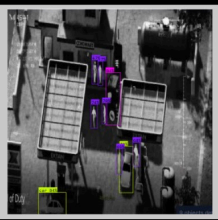
Image Source	YOLOv8	Footage Training	Simulation Training	Combined Training (67% Real, 33% Sim)
Simulation				
Live Footage				

Figure 5: Multi-Fidelity AI Models – Examples of Multi-Fidelities Considered

Findings

The literature review resulted in an understanding of how AI/ML-based systems are fundamentally different from traditional systems in terms of their requirements, development processes, system architecture, and evolution/maintenance. The main implications on the T&E process include:

- Reproducibility challenges arising from random data, random observation order, random initialization for weights, random batches fed to the network, random optimizations in different versions of frameworks and libraries
- The oracle problem
- Changing behavior under test
- Lack of test adequacy criteria

The range of objectives for *Test* include correctness, relevance, robustness, security, data privacy, efficiency, fairness, and interoperability. The literature review also resulted in a classification of the techniques for testing ML models.

The preliminary implementation of the framework for the use-case highlighted that testing measures, such as correct/incorrect detections and F1 score, can be used to create "meta-models" that quantify how the error varies with the features of the system and the environment (e.g., hardware quality, distance from the boundary, weather conditions, etc.). Such meta-models (see Figure 6) can be used to determine future test conditions.

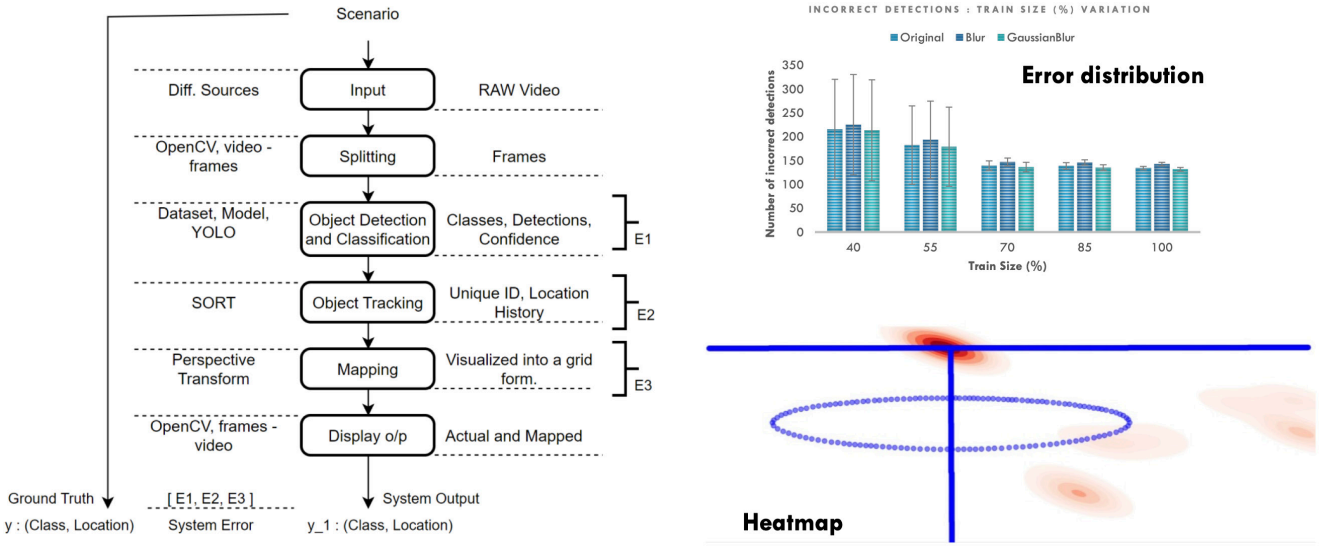


Figure 6: Multi-Fidelity AI Models - Error Propagation in the ML Model Chain (Left) Meta-models of Error to Drive Further Testing (Right)

Recommendations
Recent ML testing methods are valuable for individual algorithmic-level development and testing, but they need to be integrated from a systems perspective in the overall JCIDS process. Meta-models from ML model testing can be used to communicate between contractors and the T&E community. The proposed framework should be extended using decision analytics techniques to assess cost-value tradeoffs.

The research team's recommendations include:

- Further developing the framework that integrates JCIDS with multi-fidelity T&E to demonstrate it for a simulated acquisition scenario
- Using code sharing platforms (e.g., GitHub) for sharing test method codes with the research community
- Using online model platforms (e.g., NanoHub) for training the T&E user community

PENETRATION TESTING

Research Objective
The research team's objectives included:

- Surveying the SoA automated penetration testing
- Exploring integration of ML into Penn State's Automated Attack Framework for Test and Evaluation (AAFT)

Methods
The team's research methods included:

- The VTNSI team simulating network attacks using reinforcement learning
- The VTNSI and ARL-PSU teams collaborating on attack tree pruning

Findings
Relating to the future direction for automated penetration testing, the team identified whole campaign emulation as an organizing principle for benchmarking AI/ML-based penetration testing systems and for directing future research efforts. The team wrote an extensive review paper of the technology for automating penetration testing with reinforcement learning and attack graphs and published an article in IEEE Instrumentation and Measurement Magazine.

Relating to the RL-based attack tree search, the team implemented a method for searching attack trees using reinforcement learning. Using this method, the team was able to identify weighted shortest paths after training. The team identified that for reinforcement learning to offer a favorable advantage, there would need to be more uncertainty or stochasticity in the attack paths, otherwise deterministic search outcompetes the AI/ML at searching attack trees.

Recommendations

- AI/ML offers significant multiplier effects for cyber operators and current state of the art should be considered as a supplemental technology. Future development of autonomous cyber test systems should begin development from the perspective that AI/ML will be central.
- Current systems that treat deterministic search and logic programming as central are often not set up to tolerate data-heavy sub-routines that AI/ML rely on. Side-by-side comparisons of deterministic and logic programming systems with AI/ML systems should consider the network scale, actions and behavioral scope, and covertness of comparison tasks.

CONCLUSIONS

In support of DOT&E's I-Plan and thus the National Defense Strategy, the AIRC research team has completed foundational work advancing in focus topics on a Joint Test Concept to test the way we fight, maturing the use of Bayesian and Design of Experiments, in addition to maturing Data Security practices to accelerate delivery of weapons that work. The team has also matured automated Penetration Testing, furthered the application Digital Engineering for Test and Evaluation, developed methods for Test and Evaluation of AI enabled systems, and pursued the application of further Verification, Validation, and Uncertainty Quantification via analytical models to help pioneer T&E of weapon systems built to change over time. Through this first year of research, industry best practices have been captured and shared via public engagements and publications, and the AIRC research team has collaborated across the T&E community forming great partnerships for future research.

Though the first-year accomplishments are many, there is still much work to be done to realize the ultimate objectives of DOT&E. Recommendations captured in this report are already being discussed with DOT&E in addition to addressing other areas of focus such as the development of an Agile and enduring T&E enterprise workforce. It is the hope and recommendation of the AIRC team that DOT&E leverage the breadth of academic expertise in the AIRC UARC in the coming year to realize their I-Plan objectives.

APPENDIX A. EVENTS COORDINATED/ATTENDED

Pillar 1: Test the Way We Fight

JOINT TEST CONCEPTS

Joint Test Concept Workshop #1

- **Date:** 3/9/2023
- **Sponsor:** DOT&E
- **Purpose/theme of the event:** Determine additional competencies required for the JTC COI, refine initial JTC framework and identify T&E best practices.
- **Attendee's participation:** DOT&E and VT-ARC participated in the workshop
- **Insights/takeaways:**
  - The current study team does not span the entire T&E community, and there are additional participants needed for future workshops in order to fill all areas of competence.
  - JTC layer descriptions were updated from their initial framework.

Joint Test Concept Workshop #2

- **Date:** 5/17/2023
- **Sponsor:** DOT&E
- **Purpose/theme of the event:** Answering the question "How do we prepare for the future fight without losing our ability to execute the fight tonight?"
- **Attendee's participation:** VT-ARC coordinated the event and led the workshop; they ran a designed thinking workshop, and coordinated all the participation between the groups, including the prompts, groupings, and themes.
- **Insights/takeaways:**
  - Identified pathways for effective joint T&E integration into the complete system lifecycle from concepts through sustainment.
  - Identified JTC data considerations that support T&E best practices within program office timelines.
  - Explored solutions to address the problem of poor data quality and strategies for changes to existing data practices to help new policy adoption.

Joint Test Concept Workshop #3

- **Date:** 7/13/2023
- **Sponsor:** DOT&E
- **Purpose/theme of the event:** Focusing on strategy design for systems that support the Joint Concept for Command and Control and/or Joint All-Domain Command and Control.
- **Attendee's participation:** VT-ARC coordinated the workshop, designing the tabletop exercises and coordinating participation amongst the different groups.
- **Insights/takeaways:**
  - Identified the team composition that would be required for a joint test team to function well.
  - Identified key metrics for success criteria for a joint test team's implementation.

Digital Engineering & T&E Connect the Dots Workshop

- **Date:** 6/27/2023 - 6/29/2023
- **Sponsor:** MITRE, via DTE&A
- **Purpose/theme of the event:** Exploring the continuity of T&E across system and addressing alignment of T&E with model-based engineering.
- **Attendee's participation:** Ms. Christina Houfek attended in-person

- **Insights/takeaways:**
  - » Several tooling vendors have developed products to aid in the integration of the various lifecycle engineering efforts for model-based test and evaluation.
  - » There are several programs actively integrating systems engineering models with test plans and test results to include physics-based test results.
  - » Identified the team composition that would be required for a joint test team to function well.
  - » Identified key metrics for success criteria for a joint test team's implementation.

4th T&E Renaissance Workshop

- **Date:** 9/13/2023 - 9/14/2023
- **Sponsor:** Naval Surface Warfare Center – Port Hueneme Division
- **Purpose/theme of the event:** For the Naval T&E community to meet with the goal of re-invigorating the science and art of T&E through collaboration and knowledge sharing.
- **Attendee's participation:** Ms. Christina Houfek, participating and presenting in-person

Pillar 2: Accelerate the Delivery of Weapons that Work

INTEGRATED TESTING

DATAWorks Workshop

- **Date:** 4/25/2023 - 4/27/2023
- **Sponsor:** DOT&E, NASA, Institute for Defense Analyses (IDA), Statistics in Defense and National Security (SDNS)
- **Purpose/theme of the event:** To showcase a combination of applied problems, unique methodological approaches, and tutorials from leading academics, and to facilitate collaboration among all involved, including other government agencies.
- **Attendee's participation:** Dr. Victoria Sieck, Dr. Corey Thrush, and Dr. Cory Natoli, STAT COE delivered a day-long short course on "Applied Bayesian Methods for Test Planning and Evaluation"
- **Insights/takeaways:**
  - » Approximately 75 people attended both in person and remotely, learning the purpose of and how to implement Bayesian methods.
  - » The Integrated Testing team came away with a better understanding of additional training needs and new points of contact in the test community.

Military Operations Research Society (MORS) 91st Symposium

- **Date:** 6/12/2023 - 6/15/2023
- **Sponsor:** US Military Academy
- **Purpose/theme of the event:** For the national security community to exchange information, examine research and discuss critical national security topics
- **Attendee's participation:**
  - » Dr. Justin Krometis, VTNSI participated and presented in-person
  - » Dr. Victoria Sieck, STAT COE participated and presented in-person
- **Insights/takeaways:**
  - » Overall interest level in the methods seemed very high.
  - » The team came away with new points of contact in both government and industry and generated interest in the applications of these methods to T&E.

APPENDIX B. RESULTING SUPPORTING PRODUCTS

Pillar 1: Test the Way We Fight

JOINT TEST CONCEPT

Presentation: Joint Test Concept Pilot

- **Presenter:** Ms. Christina Houfek
- **Presentation Date/Location:** 9/13/2023 - 9/14/2023, Port Hueneme, CA
- **Audience:** Navy T&E community participating in the 4th T&E Renaissance Workshop

Pillar 2: Accelerate the Delivery of Weapons that Work

DATA SECURITY

Report: A Survey of Data Security: Practices from Cybersecurity and Challenges of Machine Learning

- **Authors:** Mr. Padmaksha Roy, Dr. Jagan Chandrasekaran, Dr. Erin Lanus, Dr. Laura Freeman, Dr. Jeremy Werner
- **Date:** 8/10/2023

Sponsor Brief: Data Security Best Practices Brief

- **Presenters:** Dr. Erin Lanus, Dr. Jagan Chandrasekaran
- **Presentation Date:** 8/10/2023
- **Audience:** DOT&E AI Sponsors Dr. Jeremy Werner, Chris Colclough

INTEGRATED TESTING

Short Training Course: "Tutorial on Applied Bayesian Methods for Test Planning and Evaluation" at DATAWorks 2023

- **Presenters:** Dr. Victoria Sieck, Dr. Cory Natoli, and Corey Thrush
- **Presentation Date/Location:** 4/25/2023, Alexandria, VA
- **Audience:** Members of the defense and aerospace test and analysis community

Educational Material: Tutorial on Applied Bayesian Methods for Test Planning and Evaluation Training Materials

- **Developers:** Dr. Victoria Sieck, Dr. Cory Natoli, and Corey Thrush
- **Used for:** DATAWorks workshop 4/25/2023, Alexandria, VA
- **Audience:** Members of the defense and aerospace test and analysis community

Presentation: "A Comparison of Bayesian Methods for Integrated Information from Developmental and Operational Test and Evaluation" at Military Operations Research Society (MORS) Symposium 2023

- **Presenter:** Dr. Justin Krometis
- **Presentation Date/Location:** 6/13/2023, West Point, NY
- **Audience:** National security community including military, government, industry, and academic professionals

Presentation: "A Framework for Using Priors in a Continuum of Testing" at Military Operations Research Society (MORS) Symposium 2023

- **Presenter:** Dr. Victoria Sieck
- **Presentation Date/Location:** 6/13/2023, West Point, NY
- **Audience:** National security community including military, government, industry, and academic professionals

Application: R-shiny app

- **Developers:** Dr. Justin Krometis, Mr. Kyle Risher
- **Delivery Date to Sponsor:** July 2023

Programming Tool: DOT&E Plotting Style Package

- **Developers:** Dr. Justin Krometis, Mr. Kyle Risher
- **Delivery Date to Sponsor:** July 2023

Pillar 4: Pioneer T&E of Weapon Systems Built to Change over Time

VERIFICATION VALIDATION UNCERTAINTY QUANTIFICATION FOR MODELING AND SIMULATION

Sponsor Brief: VVUQ SERC Updates

- **Presenter:** Dr. Sheri Martinelli
- **Presentation Date:** 1/30/2023
- **Audience:** Dr. Jeremy Werner (DOT&E), Capt. Kenneth Cooke (DOT&E / Naval Warfare) and Mr. Jose Arteiro (Naval Warfare)

Sponsor Brief: VVUQ for Modeling and Simulation

- **Presenter:** Dr. John Gilbert
- **Presentation Date:** 5/24/2023
- **Audience:** Dr. Jeremy Werner (DOT&E)

Sponsor Brief: Practical Guidance on M&S VVUQ Based on Current SoA

- **Presenters:** Dr. John Gilbert, Dr. Justin Kauffman, Dr. Sheri Martinelli
- **Presentation Date:** 7/10/2023
- **Audience:** DOT&E Sponsors: Dr. Sandra Hobson, Dr. Jeremy Werner, Dr. Kristen Alexander, Dr. Tyler Englestad, AIRC Research Team

Survey: T&E Best Practices in M&S DOT&E

- **Authors:** Dr. John Gilbert, Dr. Justin Krometis, Dr. Sheri Martinelli
- **Release Date:** TBD
  In-progress; survey questions submitted with report; survey will be distributed by the end of the period of performance
- **Target Audience:** Individuals in government, industry, and academia that utilize or influence M&S use across a range of organization roles

T&E FOR MULTI-FIDELITY MODELS

Sponsor Brief: T&E of AI/ML-based Systems: Literature Review and Overview of Framework

- **Presenter:** Dr. Jitesh Panchal
- **Presentation Date:** 4/13/2023
- **Audience:** Dr. Kristen Alexander (DOT&E AI Sponsor)

Sponsor Brief: T&E of AI/ML-based Systems: Framework, Testbed, and Initial Results

- **Presenter:** Dr. Jitesh Panchal
- **Presentation Date:** September 2023 (delayed)
- **Audience:** Dr. Kristen Alexander (DOT&E AI Sponsor)

PEN TESTING

Sponsor Brief: State of the Art Brief on Automated Penetration Testing

- **Presenter:** Dr. Tyler Cody
- **Presentation Date:** 2/28/2023
- **Audience:** Dr. Kristen Alexander (DOT&E AI Sponsor)

APPENDIX C. RESULTING PUBLICATIONS

Pillar 1: Test the Way We Fight

JOINT TEST CONCEPTS

Nix, Maegen, Timothy Crone, and Christina Houfek. "Joint Test Concept: Setting the Foundation Workshop Report." *Stevens Institute of Technology, Virginia Tech, and Virginia Tech Applied Research Corporation*, Apr. 2023.

Houfek, Christina. "Joint Test Concept Workshop II Report: Building the Structure." *Stevens Institute of Technology, Virginia Tech, and Virginia Tech Applied Research Corporation*, June 2023.

Houfek, Christina. "Joint Test Concept Workshop III Report: JTC Pilot Final Report." *Stevens Institute of Technology, Virginia Tech, and Virginia Tech Applied Research Corporation*, August 2023.

Pillar 2: Accelerate the Delivery of Weapons that Work

INTEGRATED TESTING

Sieck, Victoria R.C., Justin Krometis, and Steven Thorsen. "A Framework for Using Priors in a Continuum of Testing." Military Operations Research Journal, (Submitted, May 2023).

Pillar 4: Pioneer T&E of Weapon Systems Built to Change over Time

T&E FOR MULTI-FIDELITY MODELS

Sonanis, Atharva M. A multi-fidelity Modeling and Experimental Testbed for Test and Evaluation of Learning-based Systems. West Lafayette: Purdue University, 2023. MS Thesis.

PEN TESTING

Cody, Tyler, et al. "Whole Campaign Emulation with Reinforcement Learning for Cyber Test." IEEE Instrumentation & Measurement Magazine 26.5 August 2023.

REFERENCES

Amarantou, Vasiliki, et al. "Resistance to change: an empirical investigation of its antecedents." Journal of Organizational Change Management (2018).

Babuška, I., F. Nobile and R. Tempone. "A Systematic Approach to Model Validation Based on Bayesian Updates and Prediction Related Rejection Criteria." Computer Methods in Applied Mechanics and Engineering (2008): 2517 - 2539. <https://doi.org/10.1016/J.CMA.2007.08.031>.

Bonvillian, William and Charles Weiss. Technological Innovation in legacy sectors. New York: Oxford University Press, 2015.

Brown, Daniel C., Shawn F. Johnson and Cale F. Brownstead. "Sediment Volume Search Sonar Development - Executive Summary." MR-2545. 2021.

Brown, Daniel, et al. "Acoustic modeling for Volumetric Sonar Systems." OCEANS 2022 Hampton Roads. Hampton Roads: Institute of Electrical and Electronics Engineers Inc., 2022. 1-5.

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

Deputy Assistant Secretary of Defense Systems Engineering. Digital Engineering Strategy. Digital Engineering Strategy. Washington, DC: U.S. Department of Defense, 2018. PDF. <https://sercuarc.org/wp-content/uploads/2018/06/Digital-Engineering-Strategy_Approved.pdf>.

Dickinson, Rebecca M, et al. "Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study." Journal of Quality Technology 47.4 (2015): 400–415.

Dienstfrey, Andrew and Ronald Boisvert. "Uncertainty Quantification in Scientific Computing." IFIP Advances in Information and Communication Technology (2012). <https://doi.org/10.1007/978-3-642-32677-6>.

DoD High Performance Computing Modernization Program. centers.hpc.mil/about/index.html. 21 February 2022. 25 July 2023.

Duque, Earl, et al. "Summary of the CFD 2030 Integration Committee Invited Panel on Physics Based Model Improvement and Uncertainty Quantification for the Digital Engineering Transformation." AIAA SCITECH 2023 Forum (2023). <https://doi.org/10.2514/6.2023-1200>.

Gregory, Joe and Alejandro Salado. "Model-Based Verification Strategies Using SysML and Bayesian Networks." (n.d.).

Guertin, Nickolas. DOT&E Strategy Update 2022. Washington, DC: U.S. Department of Defense, 2022. PDF. <https://www.dote.osd.mil/Portals/97/pub/reports/FINAL%20DOTE%202022%20Strategy%20Update%2020220613.pdf?ver=KFakGPPKqYiEBEq3UqY9lA%3D%3D>.

Jatale, Anchal, et al. "Multiscale Vvalidation and Uncertainty Quantification for Problems with Sparse Data." Journal of Verification, Validation and Uncertainty Quantification 2.1 (2017): 011001. <https://doi.org/10.1115/1.4035864>.

Kotter, John. Leading change. Boston: Harvard Business Review Press, 2012.

Lipmanowicz, Henri and Keith McCandless. The Surprising Power of Liberating Structures: Simple Rules to Unleash a Culture of Innovation. Liberating Structures Press, 2016.

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.

Nix, Maegen, et al. "Training in Innovation and Emerging Technology Adoption." Washington, DC: Acquisition Innovation Research Center, 2023.

Oberkampf, William and Barton Smith. "Assessment Criteria for Computational Fluid Dynamics Model Validation Experiments." Journal of Verification, Validation and Uncertainty Quantification (2017): 031002. <https://doi.org/10.1115/1.4037887>.

Oberkampf, William and Chris Roy. Verification and Validation in Scientific Computing. Cambridge University Press., 2010.

Office of the Director, Operational Test and Evaluation. DoD Instruction 5000.89 Test and Evaluation. Washingon, DC: U.S. Department of Defense, 2020. PDF. <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500089p.PDF>.

Richardson, Robin, et al. "EasyVVUQ: A Library for Verification, Validation and Uncertainty Quantification in High Performance Computing." Journal of Open Research Software 8.1 (2020). <https://doi.org/10.5334/JORS.303>.

Roache, Patrick. "Interpretation of Validation Results Following ASME V&V20-2009." Journal of Verification, Validation and Uncertainty Quantification 2.2 (2017): 024501. <https://doi.org/10.1115/1.4037706>.

Rogers, Everett. Diffusion of Innovations. 5th ed. New York: The Free Press, 2003. Book.

Roy, Christopher and William Oberkampf. "A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing." Computer Methods in Applied Mechanics and Engineering 200.25-28 (2011): 2131-2144. <https://doi.org/10.1016/J.CMA.2011.03.016>.

Sonanis, Atharva M. A multi-fidelity Modeling and Experimental Testbed for Test and Evaluation of Learning-based Systems. West Lafayette: Purdue University, 2023. MS Thesis.

Tacy, Adam. Innovation resistance the Forgotten Cause Innovation Failure. 23 May 2021. Print. 2023. <https://solvinnov.com/resistance-of-innovation/>.

White, Andrew, et al. "Multi-Metric Validation Under Uncertainty for Multivariate Model Outputs and Limited Measurements." Journal of Verification, Validation and Uncertainty Quantification 7.4 (2022). <https://doi.org/10.1115/1.4056548>.

Williams, David P. and Daniel C. Brown. "Three-dimensional convolutional neural networks for target classification with volumetric sonar data." Proc. Mtgs. Acoust. Virtual Conference: 6th Underwater Acoustics Conference & Exhibition, 2021. 070005.

Wojton, Heather, et al. Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation. Technical Report NS D-10455. Alexandria: Institute for Defense Analyses, 2019.

DISCLAIMER

The Acquisition Innovation Research Center 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.