ACQUISITION INNOVATION RESEARCH CENTER

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

EXECUTIVE SUMMARY AND REPORT

SEPTEMBER 2023

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

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

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

RESEARCH TEAM

Name	Organization	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
John Gilbert	VTNSI	Research Assistant Professor
Justin Kauffman	VTNSI	Research Assistant Professor
Kelli Esser	VTNSI	Associate Director, Intelligent Systems Division
Nicola McCarthy	VTNSI	Research Assistant Professor
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
Himanshu Upadhyay	Florida International University (FIU)	Associate Professor - ECE
Raymond Warner	Georgia Tech Research Institute (GTRI)	Senior Research Engineer
Alton (AJ) Schultheis	GTRI	Research Engineer I
Brandon Stringfield	GTRI	Research Scientist II
Chase Harrison	GTRI	Research Engineer II
Jennifer Sharpe	GTRI	Research Engineer I
Zoe Szajnfarber	George Washington University (GWU)	Professor
Robert Pless	GWU	Professor
Christine Krueger	GWU	Graduate Student
Justine Manning	GWU	Graduate Student
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

Alejandro Salado	University of Arizona	Associate Professor of Systems Engineering
Emily Ayala	University of Arizona	Undergraduate Research Assistant
Joe Gregory	University of Arizona	Post doctoral Associate Researcher
Sam Goldsmith	University of Arizona	Undergraduate Research Assistant
Samuel Cornejo	University of Arizona	Graduate Research Assistant

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
AI	Artificial Intelligence
AIRC	Acquisition Innovation Research Center
ARL	Applied Research Laboratory
ARL-PSU	Applied Research Laboratory - Pennsylvania State University
ASA	American Statistical Association
ASU	Arizona State University
CCAAW	Cognitive Communications for Aerospace Applications Workshop
COI	Community of Interest
CSER	Conference on Systems Engineering Research
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
FIU	Florida International University
GTRI	Georgia Tech Research Institute
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
IoT	Internet of Things
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
JHU-APL	Johns Hopkins University Applied Physics Laboratory
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
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
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.

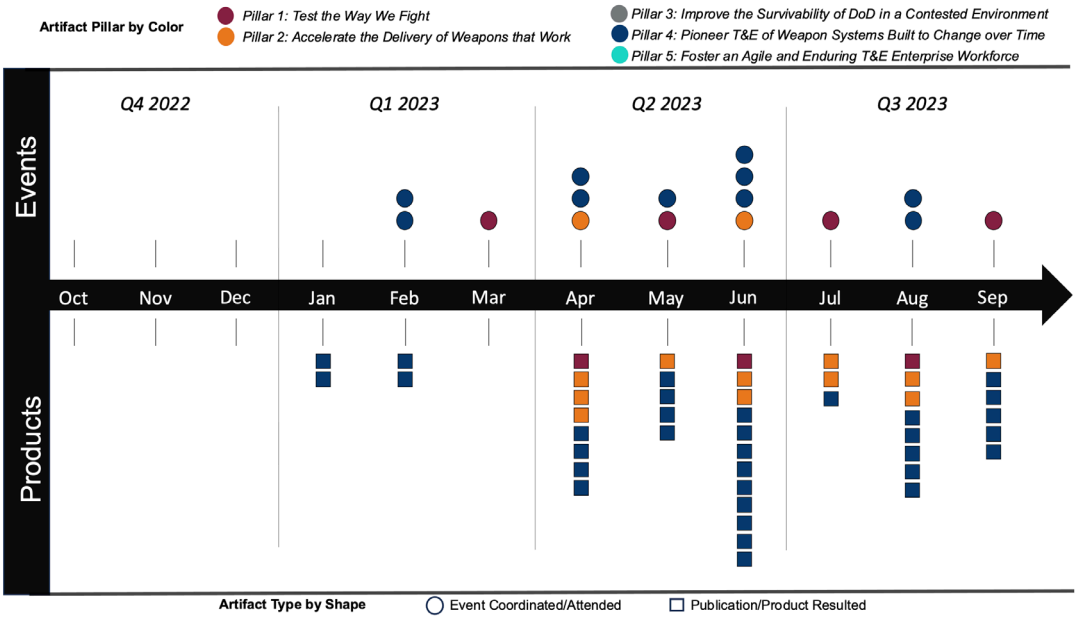

Figure 1: AIRC Year 1 Resulting Products and Event Participation

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 3 workshops (1 under 1071), 22 presentations/ sponsor briefs (12 under 1071, 3 under both), 1 webinar (under 1071), 17 reports/publications (10 under 1071), 2 apps/programming tools.

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

DIGITAL ENGINEERING LEVERAGING MODEL-BASED SYSTEMS ENGINEERING TO IMPROVE T&E – *DIGITAL ENGINEERING FOR T&E*

Research Objective

The objective for the DE application to T&E line of effort was to explore current industry practice in applying digital engineering methods to test and evaluation of DoD acquired systems and provide recommendations on how to promote further adoption in the T&E community. The research team intends to improve on current practices and address difficulties experienced by various practitioners. As part of this effort the team seeks to investigate the use of Digital Twins for T&E.

Methods

Since the release of the DoD Digital Engineering Strategy (Deputy Assistant Secretary of Defense Systems Engineering), DoD T&E leadership has embraced the pursuit in DoDI 5000.89 (Office of the Director, Operational Test and Evaluation) and is seeking how to gain the best value for achieving T&E transformation in concert with the larger DE transformation. The pathway toward implementation is not well defined. The DOT&E Strategy Update (Guertin) publication offers a framework, but further deliberation and planning by the T&E community is needed to deliver actionable recommendations for implementation over a prescribed timeline. To support the DOT&E team, the AIRC research team:

- Conducted a deep literature search of current practices and visions of future adaptation of DE methods to DoD acquisition/ development programs. The team searched 67 publications and summarized findings and best practices.

- Held discussions with eleven industry and government Subject Matter Experts (SME) over seven sessions on current DE for T&E practices focusing on best practices and leading challenges. Seven discussions were conducted to gain a broad perspective of the experiences of T&E professionals. The team has summarized the findings from these discussions and has shared this information with the DOT&E sponsor.
- Embedded itself with the T&E practitioners forming strong networks and partnerships by attending and presenting at five separate workshops and summits focused on application of DE implementation in the T&E professional practitioner. In addition, the AIRC team partnered with the DOT&E sponsor and conducted a two-day workshop on Model-Based (MB) Test and Evaluation Master Plan (TEMP) and Integrated Decision Support Key (IDSK) practices, hosting academic, government, and industry participants.
- Developed an initial ontology to support modeling TEMPs. The ontology has been used to capture an actual, document-based TEMP from the DoD; a sanitized version of the TEMP of a radar system. This has proven feasibility of a digital TEMP, as well as identified benefits of adopting digital TEMPs.
- Started to explore the novel capabilities that a digital TEMP enables over document-based TEMPs. These explorations have mainly consisted of queries to assess different aspects of the TEMP, and of connections between a systems modeling language (SysML) model and a Bayesian model to feed information from planning to evaluation and vice versa.
- Developed a testbench to validate research in digital engineering for T&E. The testbench consists of a physical hardware/software (HW/SW) rover that enables variations in design and datasets that reflect the performance of some components and some configurations.
- Shared their findings through:
  - 5 publications/papers
  - 2 webinars/workshops
  - 7 presentations/panels/briefs

Findings

Through literature searches, discussions and networking with practitioners, the research team determined that the design and development part of Model-Based Systems Engineering has advanced DE practice beyond that of the T&E portion. To realize the improvement in development timelines, complex system of system capability verification, and operational suitability that the DoD desires, the T&E practices need to be brought to the same level of maturity and integration in the development lifecycle as the other domains of digital engineering. This maturation includes improvements and standardization in ontology, integration of tooling, updates to modeling standards in SysML and other modeling language, improvements in data architecture, and connectivity with existing Modeling and Simulation and Test tracking tooling. The research team has identified three focus areas, as shown in Figure 2, to meet the challenges in the T&E community.

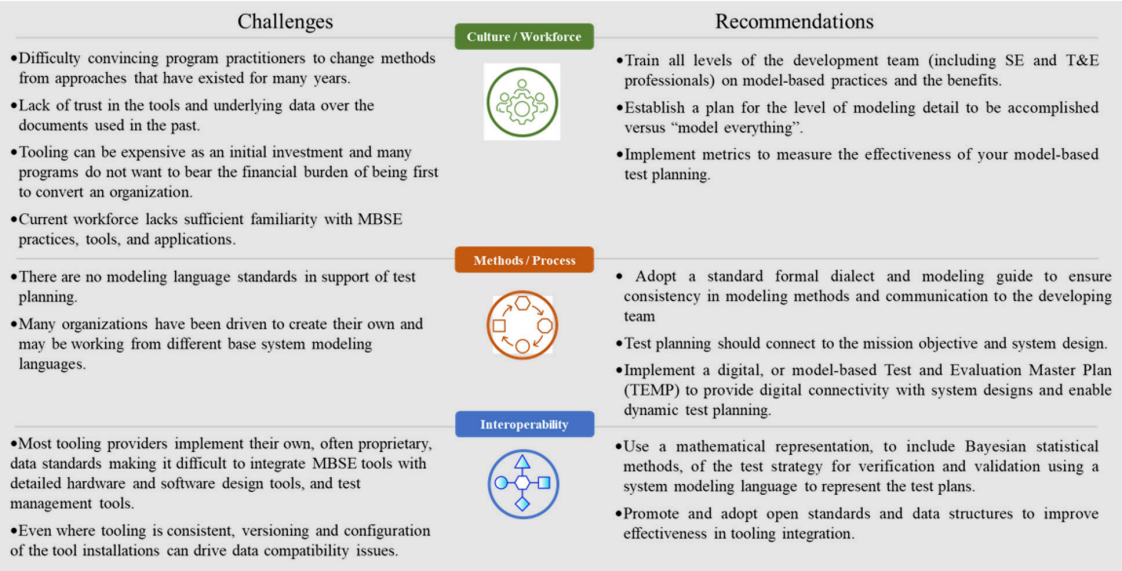

Figure 2: DE for T&E - Model Based T&E Challenges and Recommendations

As noted in Figure 2, the research team also identified cultural issues that are contributing to the slow advancement and adaptation of digital engineering practice in the DoD government and industry community. Because there are not widely documented standards as there are in other domains, many T&E professionals have found digital engineering integration cumbersome and resorted back to established processes and tooling. There has been hesitancy in believing in the potential benefits of advancing their methods. Programs understand that there is an investment to be made in process and tooling change and remain unconvinced of the return on that investment. Finally, commercial tooling interoperability force programs to develop unique application program interfaces to allow data sharing between applications. These constructs are not consistent and can become fragile as commercial tools continually mature.

By establishing community of practice networks in partnership with the DOT&E sponsors, the research team has discovered that there are numerous organizations/teams working to better the practice and further collaboration. Both DOT&E and Office of the Director for Development Test, Evaluation, and Assessments (DTE&A) have sponsored and facilitated workshops to rapidly advance tooling, process, ontology, and awareness. This continued focus on collaborative development is important to realize DOT&E I-Plan goals of key enabler 4.1.1 focused on Digital Engineering advancements. More specifically DOT&E has expressed an interest in leveraging models and simulations to reduce the cost and time related to T&E aspects of development. Digital Twins of DoD systems may be used to evaluate the operational suitability of a system class, though a single twin cannot do this, twin information may shape an operational assessment and reduce testing efforts. The team intends to explore the use of Digital Twins in the continued research efforts.

The AIRC research team has begun shaping a Model-Based T&E Roadmap from the research performed on this project. This roadmap was described in an International Test and Evaluation Association (ITEA) journal article and illustrated below in Figure 3. The team will continue to mature this roadmap in follow-on research.

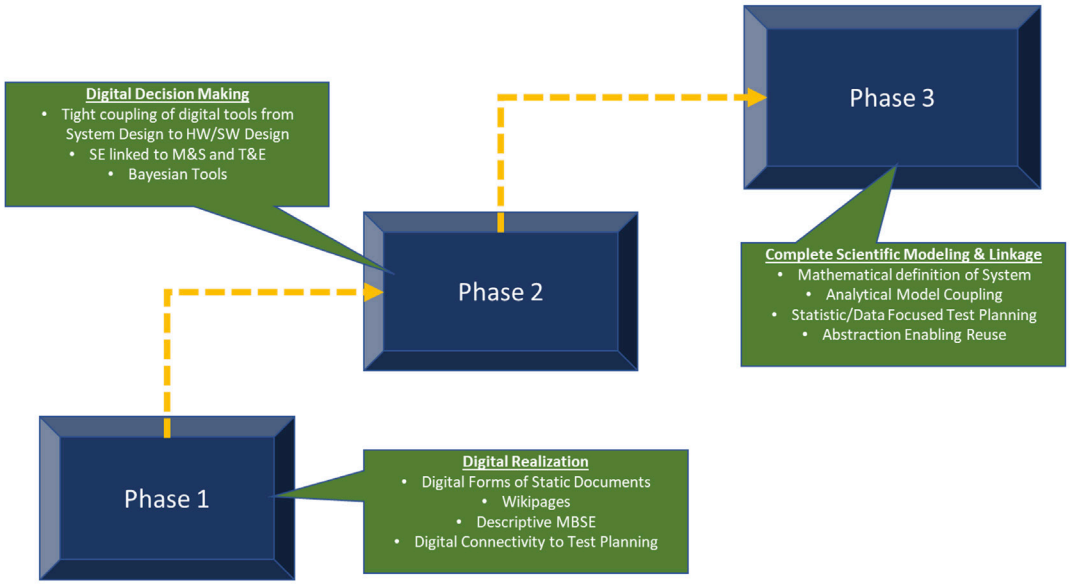

Figure 3: DE for T&E - Model Based T&E Roadmap

The AIRC research team has started progressing in this roadmap. It has demonstrated the feasibility of establishing a digital engineering environment to support T&E, as well as provided indications of benefits that a digital environment can provide with respect to the current document-based approach. Digital modeling enables agile bidirectional data sharing from planning through execution with the information classically captured in document TEMPs being dynamically captured in the model. With an ontology shared with the developers, gaps and conflicts can be rapidly identified and addressed.

Recommendations

The DE research performed by the AIRC team over the last year has yielded several recommendations for continued focus by DOT&E and the AIRC research team, including:

- A case study be pursued in the option year to validate and mature methods and tooling for T&E planning integration with Mission Engineering/System Design. This focused effort will also help further mature the model-based T&E roadmap and illustrate to the T&E community the benefits to be gained by adapting DE principles into T&E process.

- Continued work on maturing an ontology focused specifically on DE practices in T&E.
- A dedicated focus and workshop with DOT&E stakeholders to further explore the use of Digital Twins for T&E. This recommendation is in direct support of Key Enabler 4.1.1 of the I-Plan.
- DOT&E consider partnering with AIRC to bring digital engineering short courses to the government T&E offices. AIRC affiliated universities have already developed supporting material on DE that could be leveraged in support of Pillar 4 objectives, as well as Pillar 5 objectives to foster an agile and enduring T&E workforce.
- Pursuit of further definition of phase 3. The focus of the current year has largely been on phase 1 and phase 2. To reach the "complete scientific modeling and linkage" of phase 3, additional research is needed to characterize existing frameworks, capture gaps, chart roadmaps, and create products for workforce development and scalability to practice.

DIGITAL ENGINEERING LEVERAGING MODEL-BASED SYSTEMS ENGINEERING TO IMPROVE T&E – *ADVERSARIAL MACHINE LEARNING RESEARCH*

Research Objective

The adversarial machine learning study's main objective was to investigate and analyze the phenomenon of adversarial attacks in machine learning (ML) systems for test and evaluation. The research aimed to provide a thorough understanding of the fundamentals, taxonomy, attack techniques, defense mechanisms, and implications of adversarial attacks across various domains and applications. By conducting an extensive review of relevant research papers and literature, the study sought to identify the key challenges and limitations in the field and explore potential future directions for research.

Methods

To achieve the research objective, a systematic review approach was adopted. Extensive searches were conducted in academic databases, including IEEE Xplore, ACM Digital Library, Google Scholar, and arXiv, to collect research papers and articles related to adversarial attacks in the T&E domain. The search focused on papers published within the last five years to ensure the inclusion of the most recent advancements and developments in the field. Relevant keywords such as "adversarial attacks," "defense mechanisms," "robustness metrics," and "transferability" were used to retrieve relevant literature.

The collected papers were carefully reviewed and analyzed to extract essential information related to adversarial attacks, such as attack methodologies, evaluation metrics, and defense strategies. Special attention was given to studies involving attack techniques like gradient-based attacks, transfer-based attacks, decision-based attacks, physical attacks, and generative models. The evaluation of defense mechanisms, including adversarial training, defensive distillation, certified defenses, gradient masking, and detection-based approaches, was closely examined.

Findings

The analysis of the literature revealed a significant body of research focused on adversarial attacks, indicating the growing concern and interest in this area. It was observed that adversarial attacks exploit the vulnerabilities of machine learning models, leading to misclassification and erroneous predictions. Gradient-based attacks, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), emerged as potent techniques for crafting adversarial examples with minimal perturbations.

Various defense mechanisms were explored to improve model robustness against adversarial attacks. Adversarial training demonstrated promising results in enhancing model resilience. The review highlighted the importance of standardized evaluation metrics and benchmarks to comprehensively assess the effectiveness of different defenses, and the need for rigorous evaluation of defense mechanisms, including adversarial training, defensive distillation, and certified defenses, in real-world scenarios.

The implications of adversarial attacks were identified across diverse applications, including computer vision, natural language processing, autonomous vehicles, facial recognition systems, speech recognition, cybersecurity, and Internet of Things (IoT) devices. Adversarial attacks pose significant risks to safety, privacy and critical infrastructures, warranting immediate attention from researchers, practitioners, and policymakers. Furthermore, the review emphasizes the importance of T&E in the assessment of AI systems' security and robustness against adversarial attacks.

Recommendations

The research team recommends the application of generative AI in adversarial attacks within the test and evaluation domain. This method not only enhances the depth and scope of testing, exposing potential vulnerabilities in machine learning models, but also contributes significantly to bolstering these models' resilience against real-world adversarial threats. Thus, generative AI has the potential to be a game-changer, delivering more robust and reliable systems that can withstand sophisticated adversarial attacks.

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 T&E. This goal aligns with the fourth pillar of the DOT&E 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), 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 DoD, the team designed a survey to gather insights from developers, decision-makers, and users. The survey assesses the implementation of 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, Sediment Volume Search Sonar Development - Executive Summary), 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.

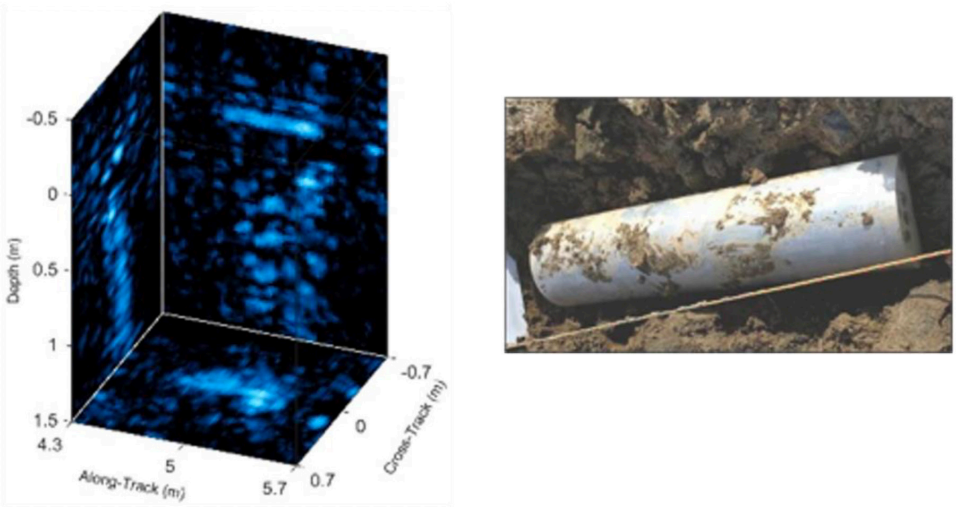

Figure 4: VVUQ - Aluminum cylinder prior to burial 3 cm below water-sediment interface, and maximum intensity projection of the volumetric image produced by SVSS. (Williams and Brown)

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), 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 V&V assessments

T&E FOR AI/ML

Research Objective

The objective was to assess and address gaps in operational T&E of AI/ML. Accordingly, the research team pursued the following sub-objectives:

- Survey the literature for best practices in T&E of AI/ML, develop an ontology for T&E of AI/ML that covers the life cycle of AI/ML development, deployment, sustainment, and retirement, and develop a risk-based framework for T&E of AI/ML
- Engage with the community, participate in workshops, and provide subjective matter expertise in the form of input and feedback to frameworks, policy, and curriculum development
- Conduct basic research into T&E of AI/ML including experimental design, combinatorial coverage, model-based test, and reference architectures

Methods

The multi-university research team was made up of researchers from VTNSI, Georgia Tech Research Institute (GTRI), and University of Arizona. The team's methodology included contributing to:

- Combinatorial coverage metrics to test deep learning for spectrogram classification
- Metric learning method that learns meaningful latent space for combinatorial coverage analysis of machine learning performance
- Systems modeling methodology for learning systems applied to explainable deep reinforcement learning based unmanned aerial vehicle flight controller
- A testbed for intelligent systems research based on computer vision
- A SysML model of the explainable deep reinforcement learning based flight controller and conceptualization of requirements-based test generation for AI/ML
- A study of design of experiments and experimental design perspectives on test set generation for AI/ML

Findings

- **Risk-based framework:** A document that describes risk across the life cycles of AI/ML systems and the programs that develop them. The framework (Figure 5) considers the training pipeline as well as the support mechanisms required for successfully operating over the life cycle. The framework emphasizes how risk can be assessed and how testing AI/ML and their support mechanisms can address risk.

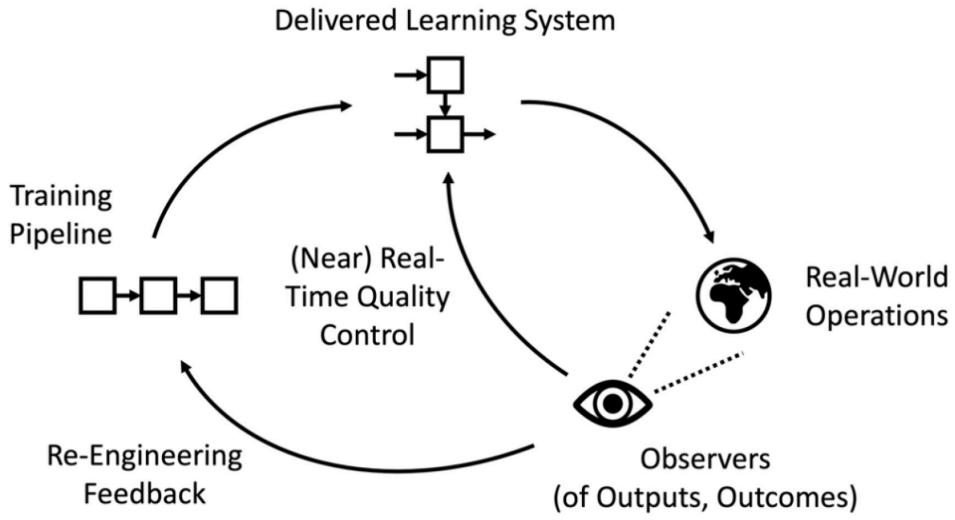

Figure 5: The Life Cycle of a Learning System from the Training Pipeline, to the Field, and Back

- **Ontology:** A document that captures the stages of AI/ML T&E concerns starting from initial scoping all the way through sustainment and retirement. The ontology (Figure 6) unfolds into concerns and questions for each stage. Prior to model T&E, the focus is on asking the right questions so that tests can be planned and resourced in advance. The stages after model T&E are not widely considered in the academic literature; however, these stages are where the execution of tests occurs. The stages generally occur from left to right as a program progresses; however, this progression is not made explicit, as stages may be revisited over time. As a program progresses, later stages of testing, like system T&E, may be revisited without first performing extensive model T&E as the engineering "V" flattens and engineering velocity for AI/ML model development accelerates.

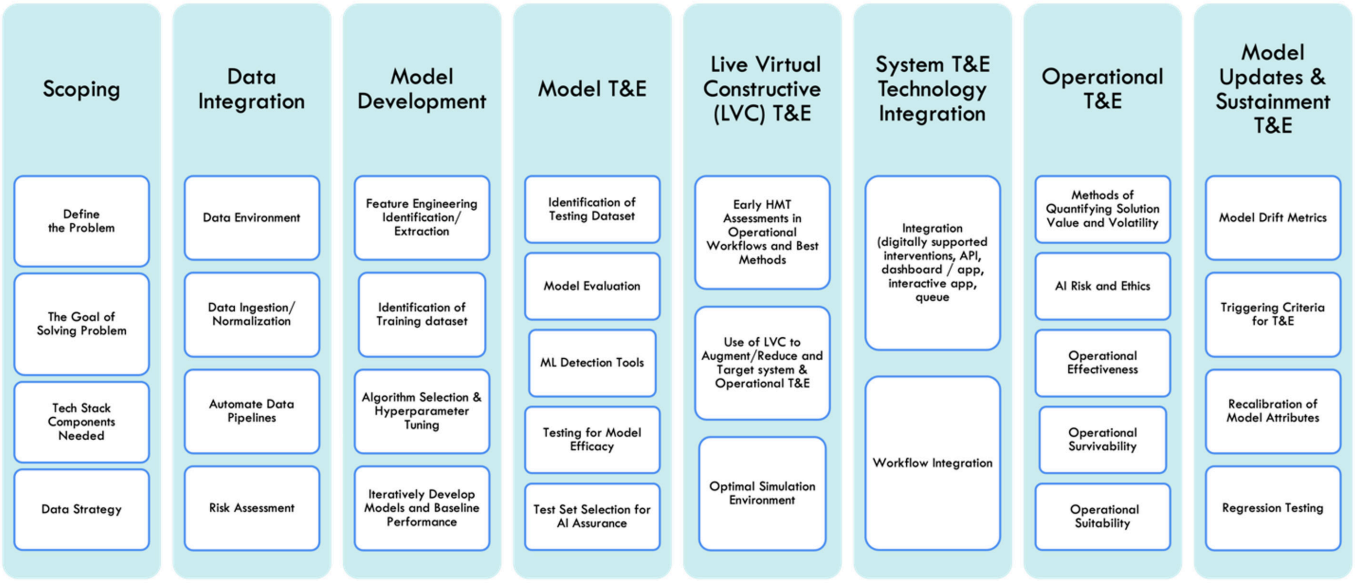

Figure 6: An Ontology for AI/ML T&E

- **Best practices:** A document that surveys the academic literature for best practices and the state-of-the-art AI/ML test methods including a cyclical view of T&E best practices for AI/ML (Figure 7). The survey focuses on the "when" and "why" of using test methods as opposed to detailing their technical underpinnings. The survey considered over 200 papers, most of which made it to the final version of the survey. Notably, academic literature focuses on what can be considered developmental testing. This underscores the importance of the top-down construction of an ontology; the academic literature as surveyed does not span the full scope of AI/ML T&E concerns.

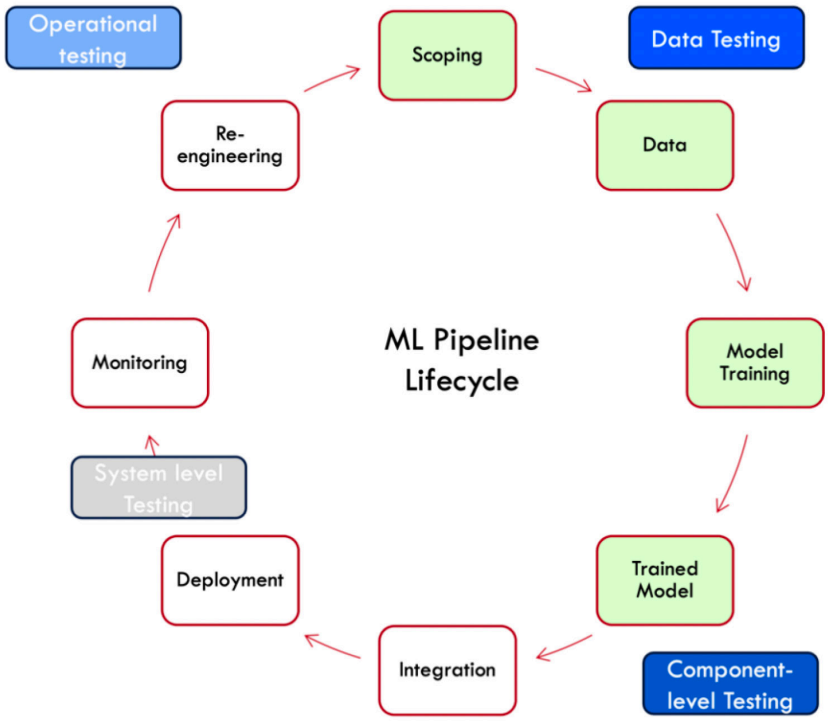

Figure 7: A Cyclical View of T&E Best Practices for AI/ML

Recommendations

The research team recommends:

- The risk-based framework, ontology, and best practices should be used as input to the next stage of policy development
- Fundamental research aligned with OT&E concerns should be funded to address the gaps in the literature and best practices
- A testbed should be created to implement, test, and iterate on the framework, ontology, and methods. This testbed should be tied to workforce development, education, and training material generation

CONCLUSIONS

In support of DOT&E's I-Plan and thus the National Defense Strategy, the AIRC research team has concluded 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 4: Pioneer T&E of Weapon Systems Built to Change over Time

DIGITAL ENGINEERING LEVERAGING MODEL-BASED SYSTEMS ENGINEERING TO IMPROVE T&E

Systems Engineering Research Center (SERC) Talk, Myths and Facts About Digital Engineering in DoD Acquisition

- **Date:** 2/15/2023
- **Sponsor:** SERC
- **Purpose/theme of the event:** Panel/webinar of Subject Matter Experts on the true practices of DE and some inaccurate beliefs currently in the industry.
- **Attendee's participation:** Mr. Geoffrey Kerr submitted the abstract and moderated the panel discussion.
- **Insights/takeaways:**
  - The panel confirmed the difficulties of tooling integration in furthering digital engineering practices across DoD programs.
  - The panel confirmed a need to focus on contracting for digital engineering as well as the digital engineering practice itself.
  - The panel spoke to numerous cultural issues still impeding some of the desired progress in implementing digital engineering practices.

JHU-APL TEMP Workshop #2

- **Date:** 2/21/2023 - 2/23/2023
- **Sponsor:** DOT&E
- **Purpose/theme of event:** To provide an opportunity for T&E community discussion on issues relating to the development of a MBTEMP.
- **Attendee's participation:** Dr. Kelli Esser, Dr. Paul Wach, and Dr. Alejandro Salado attended in-person.
- **Insights/takeaways:**
  - Improved understanding of where the acquisition & T&E communities are on introducing DE and MB practices to T&E, including current best practices and common roadblocks regarding the implementation of MBSE for T&E.
  - VTNSI and AIRC partners were motivated by this event to publish an ITEA paper to discuss key issues and proposed path forward regarding T&E adoption of DE and MB practices. The resulting paper, "Positioning Test and Evaluation for the Digital Paradigm", was published in the ITEA Journal of Test and Evaluation, vol. 44, no. 2, June 2023.

Naval Postgraduate School (NPS) 20th Annual Acquisition Research Symposium

- **Date:** 5/10/2023 - 5/11/2023
- **Sponsor:** Naval Postgraduate School Acquisition Research Program
- **Purpose/theme of the event:** Share research and exchange ideas with the defense community; with a focus on senior policymakers, practitioners, and distinguished scholars.
- **Attendee's participation:**
  - Dr. Paul Wach and Dr. Laura Freeman attended in-person and participated in the symposium panel and presented on their research.
- **Insights/takeaways:**
  - The event was attended by a wide range of DoD officials, Service members, industry professionals, tooling vendors, and academic institutions.
  - The event bolstered the desire and emphasis on transformative digital engineering practices.
  - Some of the research shared complements the research being conducted by the AIRC. However, there continues to be mounting evidence toward the need of the more fundamental progress being made by the AIRC research team.

DE for Defense Summit

- **Date:** 6/14/2023 - 6/15/2023
- **Sponsor:** Defense Strategies Institute
- **Purpose/theme of the event:** To bring together representatives from government, industry, and academia to discuss the current state of practice of DE for the DoD and determine priorities for further implementation.
- **Attendee's participation:** VTNSI, Mr. Geoffrey Kerr participated in the event and briefed general VTNSI information as well as summaries of current AIRC projects at VTNSI.
- **Insights/takeaways:**
  - The event confirmed the desire of the Office of Under Secretary of Defense (OUSD) imperative to implement DE practices on DoD programs, new and legacy.
  - The briefings reiterated many of the findings of the AIRC research team on the current state of practice of DE in and for the DoD.
  - The Summit also identified DE tooling vendors and IT practices to further Digital Engineering integration.

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:**
  - Dr. Kelli Esser and Dr. Peter Beling attended in-person
  - Mr. Geoffrey Kerr attended virtually
- **Insights/takeaways:**
  - Several tooling vendors have developed products to aid in the integration of the various lifecycle engineering efforts for model-based T&E.
  - 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.

AIRC Model-Based TEMP Strategy & IDSK Workshop

- **Date:** 8/2/2023 - 8/3/2023
- **Sponsor:** DOT&E, VTNSI, JHU APL
- **Purpose/theme of the event:** To gather the T&E community across the DoD, industry, and academia to develop a common vision for digitally transforming T&E practices.
- **Attendee's participation:**
  - Dr. Laura Freeman attended in-person
  - Mr. Geoffrey Kerr led the organization and coordination of the event and attended virtually
  - Dr. Kelli Esser attended and presented virtually
- **Insights/takeaways:**
  - The T&E community is in full support of transitioning to more integrated model-based approaches to test planning and execution.
  - There are numerous organizations that are actively implementing model-based TEMPs and IDSKs. Most of this development is not consistent across the practitioners.
  - The workshops are very beneficial to sharing best practices and driving toward consistent processes and integrated tooling. More workshops are planned.

TEST AND EVALUATION FOR AI/ML

DoD AI T&E Systems Engineering Process to Test Artificial Intelligence Right (SEPTAR) Workshop

- **Date:** 4/10/2023 – 4/11/2023
- **Sponsor:** MITRE
- **Purpose/theme of the event:** To gain consensus on the SEPTAR document, to develop an action plan to resolve any items that may be blocking consensus, and to build collaboration and rapport amongst participants.
- **Attendee's participation:** Dr. Nicola McCarthy, VTNSI, attended in-person
- **Insights/takeaways:**
  - » Use an Agile methodology (e.g., user defined requirements, data required, iterative development, etc.) for AI development once terminology is standardized across different government branches.
  - » More scenarios in each SEPTAR use cases required (library of use cases is required).
  - » Requirements should guide performance testing and performance testing is essential during development.

12th Annual International Workshop on Combinatorial Testing

- **Date:** 4/16/2023
- **Sponsor:** Institute of Electrical and Electronic Engineers (IEEE), NASA Glenn Research Center
- **Purpose/theme of event:** To provide a common forum for researchers, scientists, engineers and practitioners throughout the world to present their latest research findings, ideas, developments and applications in Software Testing, Verification and Validation.
- **Attendee's participation:** Dr. Tyler Cody attended and presented virtually
- **Insights/takeaways:**
  - » Even if "good" factors are not available for using combinatorial coverage methods with AI/ML, we can use AI/ML to learn "good" factors from the AI/ML inputs. This result positions combinatorial coverage methods as a general-purpose tool in testing AI/ML as it relaxes the need for discrete factors by learning discrete factors that are meaningful for coverage analysis.

2023 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)

- **Date:** 6/20/2023 - 6/22/2023
- **Sponsor:** IEEE Cleveland Section, NASA Glenn Research Center
- **Purpose/theme of event:** To discuss and present recent advancements and future interest in the field of AI and ML with applications toward the development of cognitive telecommunication systems for aerospace.
- **Attendee's participation:** Dr. Tyler Cody attended and presented virtually
- **Insights/takeaways:**
  - » The resulting conference presentation and papers implemented the general, domain-agnostic ideas developed to support the high-level deliverables for specific systems.
  - » The cognitive communications community is building AI/ML into their systems (satellite networks, UAVs) for critical functions but there is a gap in related systems and test engineering methods.

T&E for Artificial Intelligence and Autonomous Systems Workshop

- **Date:** 8/24/2023
- **Sponsor:** DOT&E
- **Purpose/theme of event:** To gather experts and stakeholders from across the DoD to discuss and advance frameworks for T&E of AI and autonomous systems.
- **Attendee's participation:** Dr. Tyler Cody and Dr. Laura Freeman attended and presented in-person

APPENDIX B. RESULTING SUPPORTING PRODUCTS

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

DIGITAL ENGINEERING LEVERAGING MODEL-BASED SYSTEMS ENGINEERING TO IMPROVE T&E

Webinar Moderation: *SERC Talk, Myths and Facts about Digital Engineering in DoD Acquisition*

- **Moderator:** Mr. Geoff Kerr
- **Date:** 2/15/2023
- **Event:** SERC Talk, Myths and Facts about Digital Engineering in DoD Acquisition webinar

Presentation: *Digital Engineering Enhanced T&E of Learning-Based Systems*

- **Presenter:** Dr. Laura Freeman (presentation research co-authored by Dr. Paul Wach, Dr. Jitesh Panchal, and Dr. Peter Beling)
- **Date:** 5/10/2023 - 5/11/2023
- **Event:** NPS 20th Annual Research Symposium

Presentation: *Exploring Program Archetypes to Simplify Digital Transformation*

- **Presenter:** Dr. Paul Wach
- **Date:** 5/10/2023 - 5/11/2023
- **Event:** NPS 20th Annual Research Symposium

Symposium Panel: *Leveraging Digital Engineering in Acquisition*

- **Panelists:** Dr. Laura Freeman, Dr. Paul Wach
- **Date:** 5/10/2023 - 5/11/2023
- **Event:** NPS 20th Annual Research Symposium

Report: *DE for T&E Best Practices*

- **Author:** Mr. Geoff Kerr
- **Date:** 6/2023

Presentation: *Practical Research in Digital Engineering*

- **Presenter:** Mr. Geoff Kerr
- **Date:** 6/14/2023 - 6/15/2023
- **Event:** DE for Defense Summit

Presentation: *Joint Test Concept*

- **Presenter:** Ms. Christina Houfek
- **Date:** 6/27/2023 - 6/29/2023
- **Event:** Digital Engineering & T&E Connect the Dots Workshop

Presentation: *Approach for Developing Front-end IDSK with Flexibility for Mission Engineering (ME) Alignment*

- **Presenter:** Dr. Kelli Esser
- **Date:** 6/27/2023 - 6/29/2023
- **Event:** Digital Engineering & T&E Connect the Dots Workshop

Workshop Summary: Model-Based TEMP Strategy & Integrated Decision Support Key Workshop Summary

- **Event Organizer/ Project Manager/ Author:** Mr. Geoff Kerr
- **Moderator/ Author:** Dr. Laura Freeman
- **Date:** 8/2/2023 - 8/3/2023
- **Event:** AIRC Model-Based TEMP Strategy & IDSK Workshop

  Workshop summary currently under review with the Defense Office of Prepublication and Security Review (DOPSR)

Presentation: *Approach for Developing Front-end IDSK with Flexibility for Mission Engineering (ME) Alignment (Updated)*

- **Presenter:** Dr. Kelli Esser
- **Date:** 8/2/2023 - 8/3/2023
- **Event:** AIRC Model-Based TEMP Strategy & IDSK Workshop

Report/Publication: *Development and Application of a Digital Test and Evaluation Master Plan*

- **Authors:** Dr. Joe Gregory, Dr. Alejandro Salado
- **Date:** September 2023

  In-progress- draft version will be delivered by the end of the PoP

Report/Publication: *An Ontology-based Digital Test and Evaluation Master Plan (dTEMP) Compliant with DoD Policy*

- **Authors:** Dr. Joe Gregory, Dr. Alejandro Salado
- **Date:** September 2023

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:** DOT&E Sponsor Dr. Jeremy Werner

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 PoP
- **Target Audience:** Individuals in government, industry, and academia that utilize or influence M&S use across a range of organization roles

TEST AND EVALUATION FOR AI/ML

Sponsor Brief: *SysML for Learning Systems*

- **Presenter:** Ms. Jennifer Sharpe
- **Presentation Date:** 4/13/2023
- **Audience:** DOT&E AI Sponsor Dr. Kristen Alexander

Presentation: *Anticipating Spectrogram Classification Error with Combinatorial Coverage Metrics*

- **Presenters:** Dr. Tyler Cody, Dr. Laura Freeman
- **Presentation Date:** 4/16/2023 - 4/20/2023, Virtual
- **Event:** IEEE International Conference on Software Testing, Verification, and Validation (ICST)
- **Audience:** Attendees of the IEEE International Conference on Software Testing, Verification, and Validation Workshops (ICSTW)

Presentation: *Metric Learning Improves the Ability of Combinatorial Coverage Metrics to Anticipate Classification Error*

- **Presenters:** Dr. Tyler Cody, Dr. Laura Freeman
- **Presentation Date:** 4/16/2023, Virtual
- **Event:** IEEE International Conference on Software Testing, Verification, and Validation (ICST)
- **Audience:** Attendees of the IEEE International Conference on Software Testing, Verification, and Validation Workshops (ICST IWCT)

Presentation: *Applying Learning Systems Theory to Model Cognitive Unmanned Aerial Vehicles*

- **Presenter:** Dr. Tyler Cody
- **Presentation Date:** 6/20/2023 - 6/22/2023, Virtual
- **Event:** 12th Annual International Workshop on Combinatorial Testing
- **Audience:** Attendees of the IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)

Presentation: *Anticipating Spectrogram Classification Error with Combinatorial Coverage Metric*

- **Presenter:** Dr. Tyler Cody
- **Presentation Date:** 6/20/2023 - 6/22/2023, Virtual
- **Audience:** Attendees of the IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)

Presentation: *Frameworks for OT&E of AI/ML*

- **Presenters:** Dr. Tyler Cody, Dr. Laura Freeman
- **Presentation Date:** 8/24/2023, In-person
- **Audience:** Attendees of the T&E for Artificial Intelligence and Autonomous Systems workshop

Report: *AI/ML Best Practices Report*

- **Authors:** Dr. Jaganmohan Chandrasekaran, Dr. Tyler Cody, Dr. Laura Freeman, Dr. Erin Lanus, Dr. Nicola McCarthy
- **Date:** September 2023

  In-progress- draft version will be delivered by the end of the PoP

APPENDIX C. RESULTING PUBLICATIONS

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

DIGITAL ENGINEERING LEVERAGING MODEL-BASED SYSTEMS ENGINEERING TO IMPROVE T&E

Gregory, Joe, and Alejandro Salado. "Model-Based Verification Strategies Using SysML and Bayesian Networks." Conference on Systems Engineering Research (CSER), January 2023. *Received Best Paper Award.

Freeman, Laura, Peter Beling, Kelli Esser, Paul Wach, Geoffrey Kerr, Alejandro Salado, Jeremy Werner, and Sandra Hobson. "Positioning Test and Evaluation for the Digital Paradigm", *The ITEA Journal of Test and Evaluation*, vol. 44, no. 2, June 2023.

TEST AND EVALUATION FOR AI/ML

T. Cody and L. Freeman, "Metric Learning Improves the Ability of Combinatorial Coverage Metrics to Anticipate Classification Error," 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Dublin, Ireland, 2023, pp. 206-213, doi:10.1109/ICSTW58534.2023.00045.

Cody, Tyler and Peter Beling. Applying Learning Systems Theory to Model Cognitive Unmanned Aerial Vehicles presentation, Institute of Electrical and Electronic Engineers (IEEE) Cognitive Communications for Aerospace Applications Workshop (CCAAW), June 2023.

Cody, Tyler, and Laura Freeman. Anticipating Spectrogram Classification Error with Combinatorial Coverage Metrics presentation, Institute of Electrical and Electronic Engineers (IEEE) Cognitive Communications for Aerospace Applications Workshop (CCAAW), June 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

Copyright © 2023 Stevens Institute of Technology and Virginia Tech National Security Institute (VTNSI). 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#0510.

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

No Warranty.

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

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