RESEARCH
AND ENGINEERING

OFFICE OF THE UNDER SECRETARY OF DEFENSE
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

MEMORANDUM FOR TEST AND EVALUATION PROFESSIONALS

SUBJECT: Report on Advancements in Test and Evaluation of Autonomous Systems

Autonomous systems (AS) continue to garner pronounced interest in the Department of Defense and military Services with a focus on developing and maintaining U.S. advantages in a field in which near-peer adversaries are quick to capitalize on rapidly changing technological capabilities. As part of its modernization priorities, the DoD targets the seamless integration of diverse unmanned/mixed team capabilities that provide flexible options for the Joint Force. The challenges associated with proper test and evaluation (T&E) of these systems to ensure appropriate developmental and operational mission assurance are significantly more complex than typical weapons systems.

As part of my office's focus in this area, the Scientific Test and Analysis Center of Excellence (STAT COE) recently concluded its first study on advancements in test and evaluation of autonomous systems (ATEAS). This effort owes a debt of gratitude to the diversity of thought gleaned from the participation and contribution of government, FFRDC, and industry partners. Both this report and the 2019 ATEAS Workshop report publication contain essential insight to share with the T&E professional community.

The report documents suggest refinements in current challenges and gaps in DoD methods, processes, and test ranges to rigorously test and evaluate autonomous systems derived from the 2019 ATEAS workshop, review of policy and literature (DoD and Academia), and data calls (DoD and Industry). It also identifies several methods, processes, and lessons learned to enable rigorous T&E of autonomous systems, including formal methods, combinatorial methods, and T&E technical planning tools (Range Adversarial Planning Tool (RAPT), Boundary Explorer).

Developmental testing is a key component in obtaining decision quality information for informing acquisition decisions and the systems engineering process. Current complex cyber-physical systems continue to exhibit challenges to the testing, planning, and effective data analysis, upon which these processes and decisions are contingent. These challenges are even more acute for increasingly complex and adaptive autonomous systems. This report does much to capture recent advancements in test planning and assessing autonomous systems and further refines the challenges present in the T&E of autonomous systems which must be addressed to close developmental mission assurance gaps and improve management of program risks.

COLLINS.CHRISTOPHER.CLAY.1050967561

Digitally signed by COLLINS.CHRISTOPHER.CLAY.1050967561
Date: 2020.10.19 10:36:47 -04'00'

Christopher C. Collins
Director, Developmental Test, Evaluation, and Assessments

cc:
DOT&E
JITC

2019 Advancements in Test and Evaluation of Autonomous Systems (ATEAS) Workshop Report

August 31, 2020

Dr. Darryl Ahner, Director
Dr. Steven Thorsen, Associate Director
Kaity Jones, Ctr
Emily Divis, Ctr
Dr. Troy Welker, Ctr
Dr. Steven Oimoen, Ctr
Dr. Lenny Truett, Ctr
Dr. Bill Rowell, Ctr

Office of the Secretary of Defense (OSD)
Scientific Test & Analysis Techniques (STAT) Center of Excellence (COE)
Department of Operational Sciences
Air Force Institute of Technology
2950 Hobson Way
Wright-Patterson AFB, OH 45433-7765

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. COE products are available at
https://www.afit.edu/stat

Executive Summary

Autonomous systems continue to garner great interest in the Department of Defense. Autonomous defense systems will offer new capabilities, may display emergent behavior, and will need to operate in dynamically changing environments. The challenges associated with properly testing and evaluating these systems to ensure appropriate developmental and operational mission assurance still need to be studied and addressed. This need was noted as early as 2011 when the Secretary of Defense noted autonomy, and specifically called out Test & Evaluation and Verification and Validation, as one of the seven priority science and technology investment areas in the FY 13-17 Program Objective Memorandum [1]. The Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)), Deputy Director for Developmental Test, Evaluation and Prototyping (DD(DTEP)), sponsored a study focused on Advancements in Test and Evaluation of Autonomous Systems (ATEAS). DD(DTEP) selected the OSD Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE) to lead this study.

In 2015, D(DT&E) directed the STAT COE to conduct a workshop to address "[h]ow to conduct test and evaluation of autonomous systems and what specific testing methodologies and capabilities need to be addressed." Workshop participants identified several challenges categorized into eight challenge areas specifically for test and evaluation of autonomous systems. The 2019 Workshop summarized in this report confirmed the eight challenge areas as still relevant to the T&E of autonomous systems and revealed two additional challenge areas discovered in the interim period of the two workshops. These ten challenge areas are:

- Requirements and Measures*
- Test Infrastructure and Personnel*
- Design for Test*
- Test Adequacy & Integration*
- Testing Continuum*
- Safety/Cybersecurity for Autonomous Systems*
- Testing of Human Systems Teaming*
- Post Acceptance Testing*
- Simulation for Developmental Mission Assurance
- Data

* *Indicates challenge identified in 2015 workshop report*

The 2019 ATEAS Study has several objectives to meet, and the Workshop provided a great opportunity to engage the larger enterprise in answering them. The two primary objectives of the ATEAS Workshop were:

- Refine current challenges/gaps in DoD methods, processes, and test ranges to rigorously test and evaluate autonomous systems
- Identify and develop methods and processes, and identify lessons learned needed to enable rigorous T&E of autonomous systems

The ATEAS Workshop (referred to as "the Workshop" throughout the rest of this report) took place over three days of presentations and panel-led discussions. Each day had a specific theme. The first day focused on identifying new and verifying previously understood gaps and challenges with respect to T&E of autonomous systems. The second day focused on best practices employed to test and evaluate autonomous systems. The third day featured the way forward and the resources required for rigorous T&E of autonomous systems.

The Workshop generated a more complete overview of the current state of T&E of autonomous systems to include processes, methods, and necessary test ranges. The scope included participation from a wide swath of DoD, including some select industry and academia. The Workshop also initiated a data call to obtain additional information from both participants and their relative contacts, to gather insight with respect to T&E of autonomous systems and the approach to move forward. Further, the Workshop generated terrific opportunities to collaborate throughout the DoD, academia, and industry to further identify current best practices, methods, and approaches to assist the development of rigorous, effective, and efficient test and evaluation of autonomous systems, and to understand and manage the risks associated with such systems.

The Workshop afforded attendees the opportunity to discuss future engagement with the STAT COE for PhD-level consultation on test planning and analysis support with DoD programs that would serve as "pilot programs" for continued development of best practices and lessons learned. Potential pilot programs identified during the Workshop were the Army's Autonomous Systems Test Capability, the Air Force's Skyborg small unmanned aerospace system, and the Navy's unmanned maritime systems portfolio. These future engagements should specifically document best practices and case studies that put forth solutions that close the gaps identified by the challenge areas. Other follow-on activities include:

- Analysis of the data call responses and results
- Development of an autonomy T&E taxonomy with associated lexicon and ontology
- Future funding for test planning and analysis efforts within programs
- Plans for further T&E Autonomy workshops/seminars

The STAT COE acknowledges the great amount of participation in the Workshop and is very grateful to all who attended and imparted their wisdom. The ideas captured will be acted upon well beyond the reach of this report.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

Autonomous systems have continued to increase in importance throughout the Department of Defense (DoD). Autonomy enables a particular action of a system to be automatic or, within programmed boundaries, self-governing. For the purposes of this report, autonomy is defined as a system *having a set of intelligence-like capabilities (i.e., learned behaviors) that allows it to respond to situations that were not pre-programmed or anticipated (i.e. learning-based responses) prior to system deployment*. Autonomous systems (AS) have a degree of self-governance and self-directed behavior, possibly with a human's proxy for decisions. The DoD defines mission assurance (MA) as: "A process to protect or ensure the continued function and resilience of capabilities and assets, including personnel, equipment, facilities, networks, information systems, infrastructure, and supply chains, critical to the execution of DoD mission-essential functions in an operating environment or condition" [2]. MA applies throughout each individual phase of the defense acquisition lifecycle: design, test, evaluation, production, deployment and sustainment. MA identifies and manages risks. The specific purpose of developmental test and evaluation (DT&E) during the "test" phase is to "manage and mitigate risks during development, to verify that products are compliant with contractual and operational requirements, and to inform decision makers throughout the program life cycle" [3]. Ahner synthesizes MA and DT&E, stating "developmental mission assurance addresses activities as outlined above, during engineering development and prototyping that ensure the system performs according to the desired end state" [4]. Implementing "shift-left" initiatives and applying STAT throughout system and component level development, from conceptual designs to physical prototypes and eventually the first fully-developed system, supports the conduct of efficient and effective tests during development. Addressing the gaps and challenges in test and evaluation of autonomous systems discussed in this report and in [4] will directly improve the rigor of progressive sequential developmental testing and overall mission assurance. Because of these intelligence-based capabilities and their coexistence with human actors, autonomous systems pose new challenges in achieving developmental mission assurance through adequate performance, safety, ethics, and cybersecurity outcomes. The ends of the autonomous system spectrum start with simple logic that can repeat simple tasks based upon a finite number of possible events. On the extreme other end, however, is the limitless application of artificial intelligence with trans-human data gathering and analyzing capabilities acting within the confines of a defining set of tasks and truly learning as the system ages. What hasn't kept up with the development of AS technology are the processes, methods, and techniques by which to gain developmental mission assurance. This need was noted as early as 2011 when the Secretary of Defense noted autonomy by specifically calling out its Test & Evaluation, and Verification and Validation, as one of the seven priority science and technology investment areas in the FY 13-17 Program Objective Memorandum [1]. Continuing this emphasis, in June 2019, D(DT&E) sponsored a study through the OSD Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE) on Advancements in Test and Evaluation of Autonomous Systems (ATEAS) to support three lines of effort:

1. Assist acquisition programs and rapid fielding efforts in the use of innovative and efficient DT&E strategies to ensure production readiness and fielded weapon systems meet warfighter needs

2. Improve the Defense Acquisition T&E workforce "practice of the profession"
3. Advance test and evaluation (T&E) policy and guidance

1.2. Objectives

Specific objectives associated with these lines of effort, not in priority order, are to:

- Collaborate with D(DT&E), Service components, OSD R&E Autonomy Community of Interest, and other DoD activities to develop methodologies and processes needed to conduct rigorous testing of autonomous systems
- Identify current best practices, lessons learned and new approaches (DoD and Industry) for testing and evaluation of autonomous systems
- Identify gaps and challenges in current DoD processes or methodologies to evaluate autonomous systems
- Identify current gaps in test ranges to evaluate autonomous systems

The STAT COE commenced the study by addressing the objectives listed above with planning of an ATEAS Workshop to include participants from across DoD, industry, and academia.

2. ATEAS Workshop Basis

The 2019 STAT COE ATEAS Workshop (the "Workshop") built upon the 2015 STAT COE T&E of Autonomous Systems Workshop by refining current challenges and gaps issues within the DoD. The 2015 Workshop Report identified multiple challenges and gaps for T&E of autonomous systems, which were categorized into eight major challenges for T&E of autonomous systems:

- Requirements and Measures
- Test Infrastructure and Personnel
- Design for Test
- Test Adequacy & Integration
- Testing Continuum
- Safety/Cybersecurity for Autonomous Systems
- Testing of Human Systems Teaming
- Post Acceptance Testing

Furthermore, the 2015 report did not prioritize the challenges, noting "no one challenge was identified as being more important than the other as they span the acquisition life on any autonomous system acquisition program".

During the Workshop, the STAT COE attempted to discern the current state of autonomous systems regarding T&E of autonomous systems processes, methods, and necessary test range used within the DoD, industry, and academia, with a focus on the test, evaluation, verification, and validation (TEVV) of those systems. This Workshop was a preliminary attempt to address the ATEAS objectives and provide data in its current state from DoD, industry, and academia about the current state of T&E of autonomous systems and direction for the future.

2.1. Workshop Objectives

The 2019 STAT COE ATEAS Workshop focused on two primary objectives across three days of autonomous systems presentations:

- Refine current challenges/gaps in DoD methods, processes, and test ranges to rigorously test and evaluate autonomous systems
- Identify and develop methods and processes, and identify lessons learned needed to enable rigorous T&E of autonomous systems

The first day featured presentations and panel discussions related to current gaps and challenges. The second day focused on current T&E efforts and practices. The third day addressed how to move forward and discussed required resources. The workshop also captured suggestions for topics, agenda and speakers for future T&E of autonomy workshops and other collaborative engagements.

2.2. Workshop Participation

The following DoD, government, academia, and industry organizations participated:

- Air Force, Army, and Navy Test Centers
- DoD Autonomy Community of Interest (COI)
- Galois, Inc.
- Institute for Defense Analyses
- Johns Hopkins University Applied Physics Lab
- Joint Artificial Intelligence Center
- Department of Homeland Security Science & Technology Directorate
- National Institute of Standards and Technology
- Naval Surface Warfare Center
- OSD Developmental Test, Evaluation, and Prototyping
- STAT COE
- Test Resource Management Center
- University of Memphis/FedEx Institute of Technology
- Virginia Polytechnic Institute and State University

2.3. Workshop Presentations

Briefings and panel discussions created a collaborative environment, allowing the workshop participants to share their unique set of expertise and experiences. Each day's presenters participated in the panel discussions. The three tables below list the panel discussion topics, presenters, organizations, and presentation titles for each day of the workshop. The briefings denoted with an asterisk are limited distribution documents. Please contact the authors or their organizations for more information on obtaining copies of the presentations.

Table 2.1. Workshop Presentations – Day 1

Day 1		
Panel Discussion Topic: Gaps and Challenges in Test and Evaluation of Autonomous Systems		
Presenter	**Organization**	**Presentation**
Dr. Todd Stewart, Maj Gen, USAF (Ret.)	Director and Chancellor, Air Force Institute of Technology	Opening remarks
Dr. Greg Zacharias	Office of the Secretary of Defense	Autonomous Systems: Test and Evaluation Implications
Dr. David Tate	Institute for Defense Analyses	What Counts as Progress in Test and Evaluation of Autonomy?
Dr. Laura Freeman	Virginia Polytechnic Institute and State University	Artificial Intelligence as a Change Agent for Test and Evaluation
Mr. Rick Kuhn	National Institute of Standards and Technology	Explainable Artificial Intelligence and Autonomous Systems
Dr. Darryl Ahner	Scientific Test and Analysis Techniques Center of Excellence	Developmental Mission Assurance Challenges
Mr. Jean-Charles Ledé	Air Force Office of Scientific Research	DoD Autonomy Community of Interest Overview
Dr. Jane Pinelis	Johns Hopkins University Applied Physics Laboratory	Challenges in Test and Evaluation of AI: DoD's Project Maven*

Table 2.2. Workshop Presentations – Day 2

Day 2		
Panel Discussion Topic: Current Actions and Methods, and Gathering Best Practices within DoD		
Presenter	**Organization**	**Presentation**
Dr. Signe Redfield	Naval Research Laboratory/Joint Artificial Intelligence Center	Verification of Autonomous Systems: Challenges and Best Practices
Mr. Matt Clark	Galois, Inc.	Learning Enabled Continuous Assurance*
Dr. Mark Gillenson	University of Memphis	Test and Evaluation for Unmanned Aerial Systems
Dr. Jim Wisnowski	STAT COE/Adsurgo LLC	How to Effectively Inject STAT/DOE into Autonomy Test and Evaluation
Mr. Reid McAllister	Naval Surface Warfare Center	PMS406/Warfare Center Workshop Report
Dr. Craig Lennon	Army Research Laboratory	Draft Autonomy Community of Interest Test, Evaluation, Verification, Validation (TEVV) Roadmap
Dr. Laura Humphrey	Air Force Research Laboratory	Formal Methods for Verification, Validation, and Certification

Table 2.3. Workshop Presentations – Day 3

Day 3		
Panel Discussion Topic: What are the Current and Future Needs for the DoD to Effectively Test and Evaluate Autonomous Systems		
Presenter	**Organization**	**Presentation**
Mr. Brian Nowotny	Test Resource Management Center	Current Capabilities and Gaps at Ranges for Test and Evaluation of Autonomous Systems
Mr. Paul Kwashnak	Army Test and Evaluation Command	Enabling Effective Test and Evaluation of Autonomous Systems using Modeling and Simulation*
Lt Col David Aparicio	Air Force Test Center	Emerging Technologies Combined Test Force Mission Brief
Mr. Andrew Pollner	STAT COE/ALP International Corporation	Testing Certifications: Current and Emerging
Dr. Jim Simpson	STAT COE/JK Analytics LLC	Sequential Testing and Modeling and Simulation Validations of Autonomous Systems
Dr. Lance Fiondella	University of Massachusetts Dartmouth	Software and System Reliability Engineering for Autonomous Systems Incorporating Machine Learning
Mr. Paul Stankiewicz	Johns Hopkins University Applied Physics Laboratory	Range Adversarial Planning Tool
Dr. Chad Bieber	Institute for Defense Analyses	Operational Testing of Systems with Autonomy*

2.4. Additional Workshop Objectives

Secondary objectives related to the study goals leveraged the diverse expertise assembled for the workshop. They were to:

- Identify autonomy pilot programs from each branch of service to partner with the STAT COE with respect to T&E of autonomous systems
- Identify organizational points of contact to complete the ATEAS Data Call.
- Identify key autonomy T&E personnel throughout the DoD, academia, and industry
- Identify the future direction for rigorous T&E of autonomous systems

3. Workshop Content

3.1. Day 1: Gaps and Challenges in Test and Evaluation of Autonomous Systems

While the first day focused on refining current challenges/gaps in DoD methods, processes, and test ranges to rigorously test and evaluate autonomous systems, this objective was treated during all three days. This section summarizes the first day and subsequent thought and discussions across the widespread autonomous systems community regarding these gaps as well.

Approximately half of the briefings identified specific "challenges" in a number of areas. The challenges come from design and construction difficulties to software and hardware testing and integration to general T&E challenges that have plagued autonomous systems since their introduction (e.g. test range availability and capability, human acceptance, certification, licensure, and fielding).

The overall scope of the challenges was consistent across the presenters and demonstrated agreement with the initial eight challenges presented in the 2015 Workshop Report. The workshop concluded with two additional challenges added to the original eight as a result of the discussions. These additional challenges are referred to as 1) Simulation for Developmental Mission Assurance, and 2) Data. Each following section discusses specific aspects of a challenge identified during presentations.

3.1.1. Requirements and Measures

T&E of an autonomous system needs to assess its ability to successfully perform the required tasks and functions it was created to fulfill and evaluate its decision-making capability. To do this, it is necessary to have a thorough understanding of the system's environmental model and decision functions. This understanding informs the generation and development of system requirements, objectives, and metrics, which are the means used to measure the success of the system in performing its task, from a high level viewpoint of task completion to the low level requirements related to the manner in which the system behaves. One important point of note is the use of the two terms "measure" and "metric." While "measure" is used in the original title of this major challenge, it should be noted that throughout the body of this document, it is used interchangeably with the term "metric" to mean either a standard of measurement or a basis of comparison for expectations.

3.1.1.1. Generate precise requirements

The increasing complexity of software, physical, and cyber-physical systems for weapons and other military applications emphasizes the need for clear, precise requirements when designing, testing, and evaluating these systems. It is imperative that those involved in these efforts have great knowledge of the systems under test and their associated requirements upfront [4]. This knowledge of requirements is necessary to understand and manage the mission and its associated risks and contingencies, assess situational awareness and mission effectiveness, and assess the capabilities and overall safety of an autonomous system.

3.1.1.2. Develop metrics

T&E suffers from a lack of overall metrics to apply to autonomous systems. While a few metrics do exist, these are largely specific to a given application or system and do not translate well into generalized T&E of autonomous systems [4]. Within the T&E team, those who operate the autonomous system, referred to as Operators, have a need to expressly conduct a specific mission. We rely on them for both operationally relevant technical metrics and those metrics that define the performance of the mission.

At the system level, the quality of measurement and evaluation techniques is poor. Operators do not have metrics for how system behavior should be evaluated. This lack of metrics is notable at the component level as well. Measures of Performance (MoPs) and Measures of Effectiveness (MoEs) are common forms of evaluation for systems and their components. According to Defense Acquisition University (DAU), MoPs are "System-particular performance parameters," or "distinctly quantifiable performance features. Several MOPs may be related to achieving a particular Measure of Effectiveness (MOE)." [5] They allow users to evaluate systems before a mission begins to determine which system is best suited for use, based on tactical assumptions. MoEs are "The data used to measure the military effect (mission accomplishment) that comes from using the system in its expected environment. That environment includes the system under test and all interrelated systems, that is, the planned or expected environment in terms of weapons, sensors, command and control, and platforms, as appropriate, needed to accomplish an end-to-end mission in combat." [5] MoEs allow users to evaluate the success of the system after the mission is completed. Replacing utility and cost functions which the system itself uses during operation to choose the actions which should best carry its mission to completion with MoPs and MoEs may allow for more comprehensive evaluation of the system based on the goals which operators seek to achieve [6]. Not only should operators have metrics by which they evaluate autonomous systems, developers should also have the means by which they evaluate the system's ability to coordinate. The way in which a system chooses its actions and sub-tasks to coordinate behavior and improve performance can provide developers with much information about the system as a whole and measurements by which to assess the system's decision-making process and overall mission success.

3.1.1.3. Develop CONOPS

The Office of the Secretary of Defense (OSD) has specifically identified changes in how future concept of operations (CONOPS) will be developed based on the use and sophistication of autonomous systems. Historically, science and technology have been the driving force behind CONOPS development. Autonomous systems are projected to take over from science and technology and become the new driver for CONOPS as autonomy capabilities improve and grow [5]. However, before this occurs, improvements must be made in the realm of requirements and metrics.

3.1.1.4. Generate behavior requirements

The ability to specify precise, accurate system requirements is yet another aspect of this challenge, and the need for greater requirements at the operational and behavioral levels, as well as their traceability over the testing lifecycle, has been specifically emphasized by the OSD [7] [6].

Currently, there is no clear delineation between defining system requirements and designing the system. This is dangerous because it has the potential to devolve into a situation where designers have designed their own solution to one specific problem, but nothing else, thus creating a mismatch between evaluation criteria and original requirements. Autonomous system behavior requirements must then be precise and accurate enough to ensure that the system does what is needed and can handle operation along performance boundaries, without being so restrictive as to prevent the system from having real autonomy.

3.1.1.5. Define levels of abstraction

Due to the growing layers of complexity of cyber-physical and autonomous systems, determining the appropriate level of abstraction to use for system design or to capture the overall problem can be difficult. However, it is important to define a way to do so, because the level of abstraction greatly impacts the ability to generate accurate system requirements for design, test, and evaluation [4].

3.1.2. Test Infrastructure and Personnel

The OSD has emphasized the current lack of testbeds, ranges, and personnel to carry out T&E of autonomous systems as a challenge area [5]. Rigorous T&E of autonomous systems requires test methods, processes, strategies, physical ranges that may not currently exist, and the pool of personnel available to develop those tools and capabilities. Personnel from traditional engineering, mathematics, and computer backgrounds, as well as those with skill sets in the realm of psychology, philosophy, linguistics, and the like, will need to be recruited and trained. Additionally, special curriculums focusing on effective T&E must be designed and implemented within the workforce, and perhaps even in technical or collegiate programs.

3.1.2.1. Expand and educate workforce

A major piece of comprehensive T&E of autonomous systems is an appropriately skilled workforce. Accomplishing this will require updated policies which emphasize workforce development and executive education with a focus on both developmental and operational T&E of autonomous systems [4] [8]. It will also require the inclusion of those with understanding of psychology and language to develop the insight into system behavior necessary to design appropriate tests, into a sphere currently dominated by personnel who specialize in mathematics, computer science, and engineering. Efforts which yield the most success will need to be recorded as best practices in T&E of autonomous systems for the workforce of the future.

3.1.2.2. Develop tools

We have some tools and processes which are application and system specific, and a few which are for general use. However, we need far more and varied tools and processes than we currently have or need to determine how our current abilities can be better applied across a wide range of projects and systems [4].

3.1.2.3. Construct test ranges

Test ranges which simulate the field environment in which autonomous cyber-physical systems will operate are another concern for the T&E community. Robotic system navigation requires that systems be able to estimate their current state, the space around them, plan a route through it, and execute those maneuvers. The ranges necessary to test those capabilities must ultimately be able to simulate the sort of environment which would be faced by systems in the field, including a variety of weather conditions, terrains, foreign objects and obstacles, and other agents [7]. Constructing such ranges is not only difficult from a physical standpoint, but from that of Modeling and Simulation (M&S) as well, which must recreate weather, terrain, elevation, and local traffic accurately and in an easily controllable fashion. Additionally, these ranges must be able to support testing nationwide, for all Services. Testers must have assurance that these ranges allow them to test safely and securely, launch and recover their systems, and monitor system instrumentation, data, and movements [8].

3.1.3. Design for Test

Design for Test is yet another area of major interest to the OSD [7]. T&E of autonomous systems is made difficult by the system's internal decision-making process, which essentially operates like a "black box." To the extent possible, systems need to be designed in such a way that adequate T&E can occur even without full knowledge of the inner workings of the system, and without the need for continuous, real-time monitoring by users.

3.1.3.1. Test early, often, and efficiently

In designing systems with a focus on testing early and often, specific initiatives toward testing efficiency need to be outlined. One such initiative is simply the early testing of components and the system as a whole that follows the "shift left" paradigm which aims to find flaws in design, construction, and execution much earlier in the product or acquisition life cycle, when they are less costly to remedy [4] [10] [11].

3.1.3.2. Set user expectations of systems

One issue in designing systems so that testing can be done regularly and well is that the user perspective of the system is often incorrect. The user does not design the system in such a way that it will always function as originally intended, or designs the system as they intended and realizes that his or her foundational assumptions were flawed [6]. Additionally, the funding for such projects is not always guaranteed in terms of both dollar amount and timing. This instability makes it difficult to design a system with testing in mind if there is a chance that the funds to build and test the system will not be available in the long term [6]. Those not actively engaged in the design and construction of systems will likely have unrealistic expectations of what can be accomplished with a given budget and timeline, even when the technology and capability to create the desired system actually exists. This problem is not unique to any one discipline of engineering and is exacerbated by the paucity of verification tools available for autonomy and autonomous systems.

3.1.3.3. Improve general reproducibility

Another issue in the realm of designing systems for testing is that of reproducibility. Systems which learn and change with each successive run do not always allow for replication and reproduction of test results. One way in which testers may be able to obtain test outcomes which are more easily reproduced is the generation of consistent data which is used in system navigation. Simultaneous Localization and Mapping (SLAM) data sets are one such option, as they provide consistent, useful information to a system about its surroundings [4]. Another option is environmental interaction, which also informs the system about its location and general surroundings, including boundaries and obstacles [6].

The combination of SLAM data sets and environmental interaction may serve to improve loop closure and lead to improvements in system perception. However, these things alone do not account for the non-deterministic nature of autonomous systems, and thus do not ensure reproducibility of results.

3.1.3.4. Define system goals

Cognitive instrumentation or implementation of machine learning algorithms may lead to transparency in a system's lower level decision making [10]. To achieve this transparency, it will be necessary to enforce a structure for decision making, including clear specifications for decision making, and an overall goal that the system ought to achieve. These specifications may lead to unwanted interactions and cross-pollination between low-level decision-making and high-level goals unless goals are clearly represented and defined in the system's cognitive instrumentation. However, this is not currently common practice.

While modular verification can be employed in some instances, the lower level goals represented in such components are unlikely to either translate to system level goals or generalize to any autonomous system [4]. Those systems which are designed for a single purpose or mission are likely to be hard-coded and require less extensive TEVV than more generalized systems, thereby preventing testers from linking autonomous capabilities (or lack thereof) to system performance and raising the question of how to design systems and tests such that passing a given test is indicative of a general capability, rather than one which is application or system specific.

3.1.4. Test Adequacy, Integration, & Methods

T&E of autonomous systems is inherently difficult due to a lack of information that would allow testers to adequately define and quantify risk and performance. The OSD has stated that this challenge is further compounded by the dynamic nature of autonomous systems, which existing T&E methods are not well-suited to handle [7]. Autonomous systems are likely to interact not only with their environment but also with other systems, leading to emergent, unanticipated behaviors that are difficult for humans to track. When coupled with the classic issue of state space explosion that calls into question when sufficient testing has occurred, this issue will be likely be one of the most persistent.

3.1.4.1. Improve test efficiency

The overall adequacy of T&E of autonomous systems, knowing when the point of diminishing returns has been reached, is a long-standing question. When a system is constantly changing and learning, its state space is continually expanding, leading to the question of where testers determine that they have enough evidence to declare that an autonomous system's behavior has been verified. Additionally, the black box nature of most autonomous systems makes it difficult to determine any model components which are not originally included in the system architecture but may become pertinent during testing [4]. When these obstacles are coupled with finite resources in both budget and scheduling, the conclusion that systems cannot be exhaustively tested is obvious [11]. However, because testing must be done to the extent that we have confidence and trust in the system and its behavior even in conditions that have not been tested, an initiative for efficiency in testing must be championed, and reasonable limits determined [4].

3.1.4.2. Refine sequential testing methods

Sequential testing is a powerful tool used by the Scientific Test and Analysis Techniques Center of Excellence (STAT COE) when testing systems over time. It does however have its shortcomings, such as difficulty designing a test with limited understanding of the system under test, how to handle covariates and a changing state space, how to accommodate changes in factors, levels, variables, and responses, and how to create new measures for coverage of state space and power. Finding solutions to these short comings would go a long way toward improving T&E of autonomous systems [14].

3.1.4.3. Utilize multivariate techniques

Because T&E will need to continue over the lifecycle of an autonomous system, testers must be prepared to work with temporal responses and know how to prepare and clean data for response generation, fit model curves via splines or Fourier basis, use Principal Component Analysis, model response surfaces with as many as six factors, and use generalized regression methods [15]. These multivariate techniques will be key to working with models which cover such large state spaces.

3.1.5. Testing Continuum

T&E of autonomous systems presents the challenge of not having a predetermined test phase but requiring testing throughout the system life cycle. It is necessary to decide when, how, and what aspects of a system to test under initial assumptions; how to retest after learning has occurred; and how to use that information to make decisions. Constructing a continuum for T&E of autonomous systems will be a major factor in encouraging human trust in these systems.

3.1.5.1. Create a testing continuum framework

In order to produce relevant information for users, T&E of an autonomous system must occur often and throughout the lifecycle of the system. Creating a continuum of testing is perhaps the best chance that the T&E community has of being able to assure end users that a system that deliberately alters itself over the course of operation fulfills the requirements and meets the metrics that it ought to at that time.

In the case of systems which intermittently change or update themselves, current efforts amount mainly to alerts and warnings that the system may be performing poorly as it learns and overfitting its decision making models based on the data it has received [6].

In the case of systems which are designed to update and change themselves continuously, understanding and potential solutions are limited. Run-time verification is one potential solution which is often mentioned, though in some cases a system may then require constant supervision [6]. Safety caging and user construction and implementation of performance boundaries may also provide solution in certain circumstances. Finally, licensures for autonomous systems which are updated periodically may prove effective verification tools, provided a reasonable timeline for such license renewals is established ahead of time and the licensing process is comprehensive [6].

3.1.6. Provide Safety/Cybersecurity for Autonomous Systems

T&E of autonomous systems will need to provide testers (and ultimately users) with assurance that the system will not perform actions that are deemed unsafe or undesired. It will also need to provide assurance that the system is reasonably resistant to physical and cyber-attacks, with cyber being of greater concern. The OSD has noted that both system software and data are vulnerabilities, subject to fault through the introduction of subtle, incorrect data or algorithms which can cause poor performance, system errors, and undesired behaviors [7].

3.1.6.1. Verify system safety

One of the primary concerns to the T&E community, users, and society at large is assuring the safety of autonomous systems prior to fielding. The ability of an autonomous system to learn and alter itself continually elevates safety concerns at all levels, but for end users in particular. It also creates a unique challenge for those involved in T&E to ensure that cyber and cyber-physical autonomous systems are robust to cyber-attacks and provide evidence that vulnerabilities are well-monitored and protected. In the case of both cyber and cyber-physical systems, verification must be provided that any emergent behaviors which would be deemed unsafe or undesirable can be prevented or corrected [12].

3.1.6.2. Deliver security capability

Test ranges will also be impacted by safety and cyber security, as networked autonomy requires information sharing and protection against cyber-attacks; the ability to exchange tactics, techniques, and procedures (TTPs), within the network; mission-system-task allocation; and relevant data and information management [9]. Delivering these capabilities will be key to ensuring the safety of humans working alongside autonomous systems and overall mission success.

3.1.7. Testing of Human System Teaming

T&E of autonomous systems will need to address the ability of any combination of humans and machines to perform as partners and to determine how to measure the effectiveness of that team. Currently, societal feeling toward autonomous systems leans toward distrust. Overcoming this bias within the DoD and cultivating an appropriate level of trust between humans and autonomous

systems will be critical to system acceptance and use. This challenge was broken out into two broad categories, discussed below.

3.1.7.1. Facilitate cultural change and build trust

A major hurdle in the acceptance of autonomous systems by the ultimate users and society as a whole is a willingness on the part of humans to learn to use the system, place their trust in it, and communicate openly with the system and with testers. Specifically in the case of building trust between human and autonomous cyber-physical systems, there is much to learn about how humans exhibit trust in the system, how to measure the level of trust exhibited, and how much trust in a system is warranted [6].

Calibration of trust is a major issue. There may be an assumption on the part of human teammates too eager to trust the system that because a machine has done one thing as a human would, it will generally do so. Due to the dynamic nature of autonomous systems, this assumption may not necessarily be true. At the opposite end of the spectrum is the possibility that, because the human teammate deeply distrusts autonomy, the system is never given a chance to earn the human's trust by doing its share of the mission. Striking a balance between over and under-reliance on autonomous systems and inclusion of CONOPS and training in system and test design will contribute significantly toward human acceptance of autonomous teammates [12].

3.1.7.2. Develop human-system teaming

Human-System teaming elements that will need to be developed for use in test ranges will include platform interfaces which allow the human reliable communication with the system, the ability to affect system propulsion and vehicle controls, and the ability to monitor weapons, payloads, and the health of the system and possibly other human teammates [9].

Challenges will also be found in the realm of language and ethics as humans collaborate with machines in the field to correctly interpret and act on Commander's Intent, adhere to Rules of Engagement, prioritize mission objectives, and maintain awareness of the state of the system in mission [9].

3.1.8. Post Acceptance Testing

T&E of autonomous systems requires testing throughout the life of the system (see Testing Continuum challenge). The OSD has acknowledged that the blending of autonomy development, CONOPS, and continual T&E with rapid system prototyping and dynamic CONOPS will increase the need for post-fielding testing to occur [5]. Consequently, methods will need to be developed to test, evaluate, use, and update systems already in the field. This testing and assurance would not only contribute to the need for a designed testing continuum, but to continued human trust in autonomous systems.

3.1.8.1. Continue testing fielded systems

Because autonomous systems are designed to continually learn and alter themselves, it is imperative that there is a paradigm of continuing assessment in place that tests systems even after they have been in operation, to ensure that original assumptions made about the environment of

operation, the system and how it has changed, and whether it is still used for its intended purpose remain valid [12]. Violation of these assumptions could lead to failure in continued acceptance and use of the system. Thus, tests and licensure or certification methods will need to be continually updated to accommodate the new system after learning and use has occurred [6].

3.1.9. Simulation for Developmental Mission Assurance

Due to the interplay of sensors, software, and human agents, autonomous systems represent a complex development environment. Hence, modeling and simulation are critical components for mission assurance. In some instances, these techniques are the only ones available to test and evaluate a system, either due to budget or test range constraints. Consequently, it is of pivotal importance to develop high-quality models and simulations to test system behavior.

3.1.9.1 Develop models for simulation

As cyber and cyber-physical autonomous systems advance, the T&E community is experiencing a need for high-fidelity models and simulations to allow for verification of these systems. However, doing so requires the development of models not only for verification, but common models to standardize evaluation as well [4]. This assumes that testers are able to accurately model the behavior and decision-making process of black box autonomy, which has been an ongoing effort. This also assumes that the models which are used to train AI and cyber-physical systems are comparable to reality and that systems are robust to things such as environmental differences and timing differences between simulated and real events [6].

While simulations may ultimately be the most helpful tool developed for the TEVV of autonomous systems, they are still in their infancy.

3.1.10. Data

Data was highlighted as a vital resource in the T&E of autonomy systems. Successfully modeling an autonomous system's decision-making process depends on the quality of the empirical data which informs the model. The difficulty of acquisition, ownership, storing, sharing, handling, labeling, and quality control of data are therefore major obstacles to the autonomy T&E community. Collectively, they form the tenth major challenge to the advancements in testing and evaluation of autonomous systems.

3.1.10.1. Standardize, own, and distribute data

Central to the ability to comprehensively test and evaluate autonomous systems is the data used to do so. Data is central to model design, verification, and validation, and autonomous systems and artificial intelligences generate staggering amounts of data. However, the OSD has noted a paucity of T&E processes, accepted data formats, and common frameworks and architectures for autonomous systems [7]. Additionally, outside of industry and possibly academia it seems that very few of those involved in T&E of autonomous systems own their data. A great deal of DoD training data is either public or borrowed from external sources. Ownership of data has implications for both data quality and an organization's ability to share data within the DoD [11]. Assessing the quality of data is much harder to do when an organization is not the original

generator or owner of the data. Low-quality data can lead to the creation of low-quality models and poorly trained AI algorithms, thereby complicating the TEVV process [12]. Additionally, sharing data within the DoD is difficult even if DoD owns their own data due to self-imposed file and information sharing restrictions placed on both real and synthetic data. While these measures are necessary, they do actively prevent the open sharing of data across the DoD in a timely manner.

3.1.10.2. Process large amounts of data

Additionally, with large AI systems that generated terabytes of data on a daily basis, the ability to parse the data down to a manageable amount is a great concern. As systems learn and generate output, this output leads to classification and labeling of different data types that then creates more (meta)data that must be stored. Handing these ever-increasing amounts of data in a secure way is quickly becoming a major concern for testers [16].

3.1.10.3. Leverage synthetic data

Finally, the use of synthetic data in modeling and simulation development was discussed. The current practice is to use synthetic to augment real data when the amount of it is insufficient because the use of only synthetic data can lead to inaccurate or overfit models when the data are not representative of real data [6].

3.2. Day 2: Current Actions and Methods, and Gathering Best Practices within DoD

The focus of the second day was the ATEAS Study Objective: Identify and develop methods and processes, and identify lessons learned needed to enable rigorous test and evaluation of autonomous systems.

While identification of current efforts within academia, industry, and DoD was the theme of the presentations and discussion on Day 2, relevant information pertaining to the given objective was presented on each day of the workshop. Summarized below are presentations that concentrated on covering methods, processes, and lessons learned regardless of the actual day they were presented.

3.2.1. Autonomous System Verification Tools

Dr. Signe Redfield of the Joint Artificial Intelligence Center (JAIC) presented several tools currently in use and being developed for use specifically in the verification of autonomous systems.

Figure 3.1 depicts a sample Capability Analysis Table, which is a general purpose tool (independent of design approach) that can be used to analyze the behavior implementation and integration of a system [6]. This table lays the foundation for implementing explainable AI, improved documentation and debugging.

Adaptive Feature Detection	Position on Map		Time		Feature Status		Task		Notes
	Cond	Beh	Cond	Beh	Cond	Beh	Previous State	New State	
Vehicle Position Control	Not yet at seach area	Transit					Start	Transit	Start in Transit behavior
	At search area	Spiral					Transit	Spiral	Start Spiral from Transit
			No feature after time (T)	Spiral			Turn	Spiral	Start Spiral from Turn
					Found feature	Follow	Spiral	Follow	Start Follow from Spiral
					Found feature	Follow	Turn	Follow	Start Follow from Turn
					Lost feature	Turn	Follow	Turn	Start Turn from Follow

Figure 3.1. Capability Analysis Table [6]

The contents of the entire table are determined by the task that testers assign to a system. The top row and side panel specify platform elements specific to the system and task. Conditions, behaviors, and state transitions are specified on the second row. The contents of the inner table are behavior or sub-task specifications. This configuration connects platform and behavior design elements, and drives specification of sub-tasks and behaviors.

A tool currently in development is a set of ontology standards (Figure 3.2) which could allow for improved requirement and model specifications, helping to reduce cost early in the TEVV process by improving system knowledge and understanding that testers need in order to verify their models. Task ontology will specify the details of the system structure and what information testers must have to able to properly define a task in a manner that has meaning to a system [6].

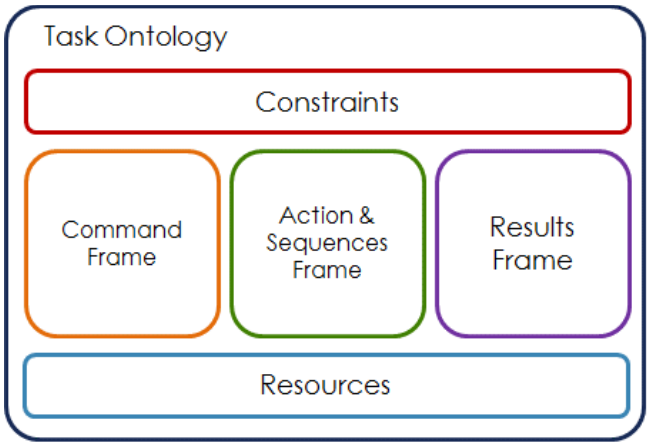

Figure 3.2. Task Ontology Architecture [6]

Solutions focused on system capabilities provide other tools testers can use to verify their systems. Once such tool is a capability representation map. This map or tree allows users to select goals and evaluation metrics based on a chosen task by following "branches" to designer-selectable options for a how a task should be specified, what it ought to achieve, and how a system's success in achieving that task ought to be measured. Related to capability representation maps are capability-centered relationships that link autonomous robot and task ontologies by providing precise definitions for the relationships between tasks and the behaviors of a system that allow it to complete those tasks (Figure 3.3) [6]. Such maps and models provide ways to more easily determine the specifics of a given task or problem, and what requirements will be imposed upon the system in order to achieve the desired outcome.

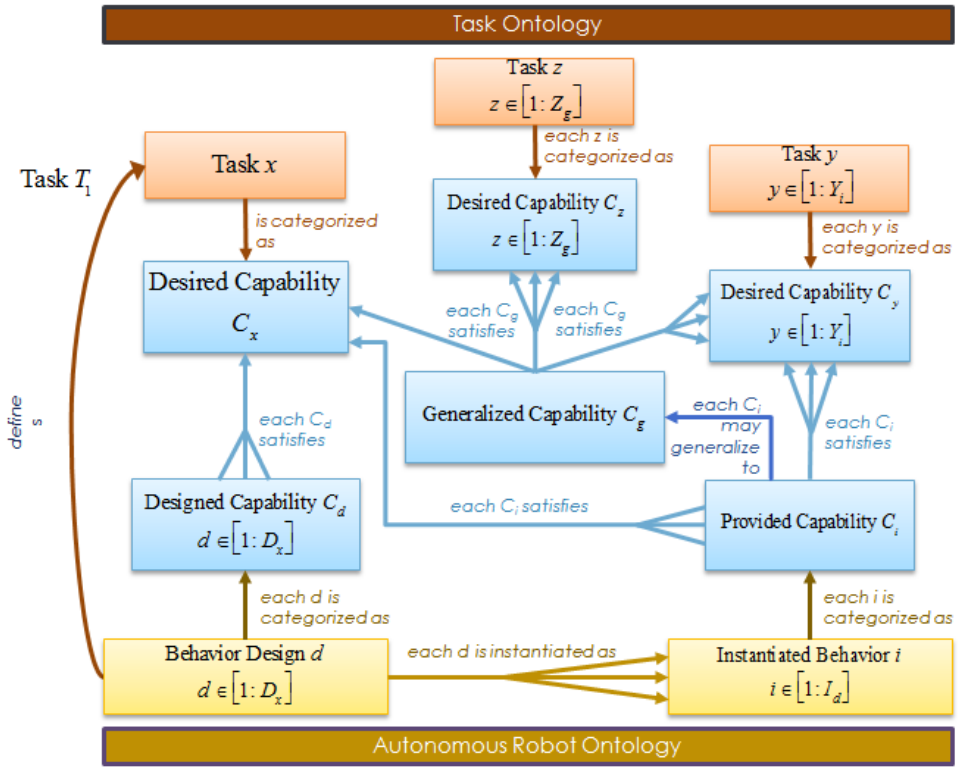

Figure 3.3. Capabilities Relationship Table [6]

Such maps and models provide ways to more easily determine the specifics of a given task or problem and what requirements will be imposed upon the system in order to achieve the desired outcome.

One possible solution to finding and solving problems within the system during verification is the use of formal methods, such as linear temporal logic and model checking. These methods are useful when problems are well-defined and the assumptions that have been made about the system and its operating environment are correct, though they provide no guarantees on system behavior when this is not the case. It is possible that formal methods could be applied to emerging types of autonomy and improve both system specifications and verification tools. However, formal methods are currently only for certain types of control and systems where much of the information about operational conditions is known in advance. Procuring that information when it is not already available can be costly. More in-depth information about the successful use of formal methods is detailed later in this section, courtesy of Dr. Laura Humphrey of the Air Force Research Laboratory (AFRL) [10].

Tools which are currently needed for the generalized verification of autonomous systems can be used to determine performance surfaces and boundaries for AI and cyber-physical systems. Such tools would ideally be able to accurately characterize and evaluate system performance over the operational space. This sort of performance profile would allow testers to generalize a system's behavior in ways which cannot presently be done because knowledge of system performance and behavior over the entire operating space is not known. The metaphor of "Swiss cheese" is often used in this context, where the holes in the cheese correspond to regions of poor performance in

the performance surface [6]. However, testers do not currently have a way to determine where those regions of unacceptable performance will occur, and the proliferation of such regions renders the use of standard machine learning techniques such as gradient descent unhelpful. Furthermore, as systems learn, the performance surface changes, eliminating a tester's ability to generalize performance between test scenarios.

The Range Adversarial Planning Tool (RAPT), which would make heavy use of modeling and simulation to determine the boundaries which shape the performance surface, is one possible solution. RAPT is described in detail by Paul Stankiewicz of the Johns Hopkins University Applied Physics Laboratory (JHU APL) later in this section [17].

Another possible solution is the use of mixed reality testing, which allows for the use of both virtual and real-world data in modeling and simulation. This technology is currently in development at the 412th Test Wing Emerging Technologies Combined Test Force (ET CTF) at Edwards Air Force Base [18]. The Testing of Autonomy in Complex Environments (TACE) system is being developed to function as a go-between for autonomy and artificial intelligence and aircraft autopilot systems. TACE is able to command aircraft without direct intervention by humans, by sending commands from an autonomous system to the autopilot and sending state information about the aircraft back to the autonomous system. This allows the system to independently guide the aircraft autopilot.

TACE also has the ability to recall the autonomous system to a safe point, should any safety specifications (e.g. communications, proximity) be violated during flight, and allows for the use of Live-Virtual-Constructive Testing (LVC) [18]. This type of testing allows autonomous systems to interact with live aircraft, even to the extent of being a virtual wingman, in areas where GPS-denial is simulated. LVC testing ultimately improves the safety of live flight tests that rely on the teaming of humans and cyber-physical systems.

Physical or physics-based simulation is another mixed-reality tool currently being refined which can be utilized to verify the behavior of autonomous cyber-physical systems in areas where communications between the system and its user are limited [6]. These simulations allow testers to see first-hand how their systems move within and interact with their environments, especially when such environments are dynamic in nature. Testbeds which mimic those environments provide testers with information about system performance and behavior that is more accurate than full simulation.

Finally, there are tools which monitor the changeability of autonomous systems. Such tools include run-time verification, safety caging (purposely restricting performance boundaries), licensing, and system-generated warnings [6]. These tools, while useful to some degree in verification, are still nascent. As autonomous systems learn and change, the original verification metrics may no longer be accurate or even applicable. Thus, a process of continuous verification over the lifecycle of the system will be necessary in order to maintain user trust. Most of the techniques discussed require significant human interaction and monitoring of the system.

While none of these efforts alone is likely to completely solve the problem of autonomous system verification, each may serve to move us closer to the solution.

3.2.2. PMS406/Warfare Center Workshop Report (McAllister)

Mr. Reid McAllister summarized the Naval Surface Warfare Center's PMS 406 portfolio, which encompasses several unmanned maritime systems for use in surface, expeditionary, and undersea warfare. The details presented in the workshop report are not approved for public release. A Distribution D version of the report is available by contacting COE@afit.edu.

3.2.3. Range Adversarial Planning Tool (Stankiewicz)

Mr. Paul Stankiewicz focused his presentation on the Range Adversarial Planning Tool (RAPT), which is currently in development at JHU APL [17]. RAPT makes use of adaptive sampling to find boundary conditions and near-boundary conditions for autonomous system performance. This ability will allow testers to verify their systems in the most comprehensive way available within schedule and cost constraints. The details of the presentation are not cleared for public release. A Distribution D version of the report is available by contacting COE@afit.edu.

3.2.4. Autonomous System Test Capability (ASTC) M&S Effort

The Army Test and Evaluation Command (ATEC) is responsible for overseeing developmental and operational testing for the Army. In recent years, the Undersecretary of Defense Research and Engineering (USD R&E) has deemed autonomy and AI critical to battlefield success. Thus, ATEC is attempting to build trust in autonomous systems early in their life cycle through a three-phased approach consisting of lab-based assessments (M&S), system level testing via hardware-in-the-loop (HWIL), and safe live testing. Early development of trust in autonomous systems will increase the trust in systems and the T&E procedures used, reduce risk of both failure and emergent behavior, and shorten test schedules. Ultimately, the Army can expect to acquire and field autonomous systems faster and cheaper due to streamlined live developmental testing and reduced test scenarios.

The Autonomous System Test Capability (ASTC) is meant to enable the building of trust in robots and autonomous systems using digital modeling and simulation (M&S), HWIL, and open-air range testing. ASTC's three phases will yield two deliverables: DRIVE (Digital RAS (robotic and autonomous system) Integrated Virtual Environment) and SEER (Safety, Environment, Engagement and Response) [19]. The details of the presentation are not cleared for public release. A Distribution D version of the report is available by contacting COE@afit.edu.

3.2.5. Combinatorial Coverage Testing Tools

Testing of autonomous systems poses unique issues because safety- and life-critical software and systems evaluation and verification techniques used for manned systems are not applicable to autonomous software and systems. One of the most challenging issues in testing is the state space expansion of autonomous systems; as the number of test factors increases, the number of combinations of factors that must be tested increases. At some point, it becomes first cost prohibitive, and then physically impossible to exhaustively test a system. Understanding what factor combinations are part of the test design is critical to designing tests that provide enough information to evaluate and verify autonomous systems. The combinatorial coverage table developed by NIST provides a visual representation of which factor combinations are covered and

a measure of the extent to which certain 2-way or higher-order factor combinations are covered [20].

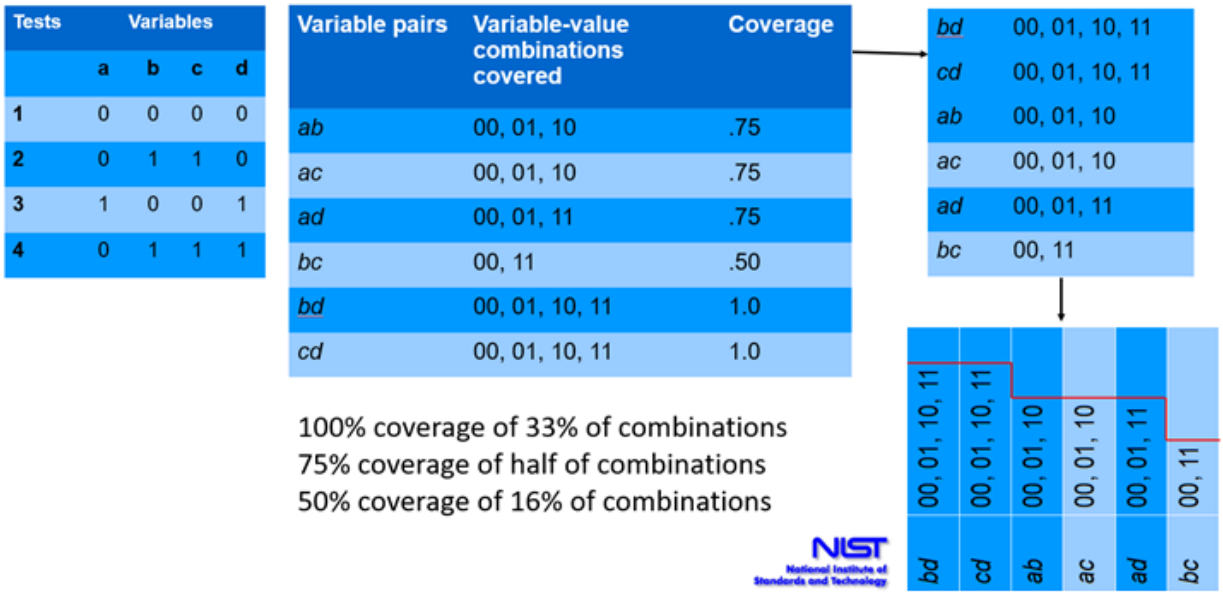

Figure 3.4. NIST Combinatorial Coverage Table [20]

Figure 3.4 provides an example of the combination table developed by NIST which allows testers to specify factors and levels [20]. Combinations of these factors and levels covered in each test can then be determined. The coverage of combinations [0, 1.0] can then be determined by the percentage of combinations covered by the percentage of total tests. The factor combinations are then ordered from greatest coverage to least and rotated counterclockwise. This rotation provides the coverage measurement chart displayed below (Figure 3.5) [20]. The red line delineates the maximum coverage for the given combinations, with covered combinations below and untested combinations above. This shows definitively which areas of the operational space were not tested and, thus, where users may expect to encounter problems if they operate.

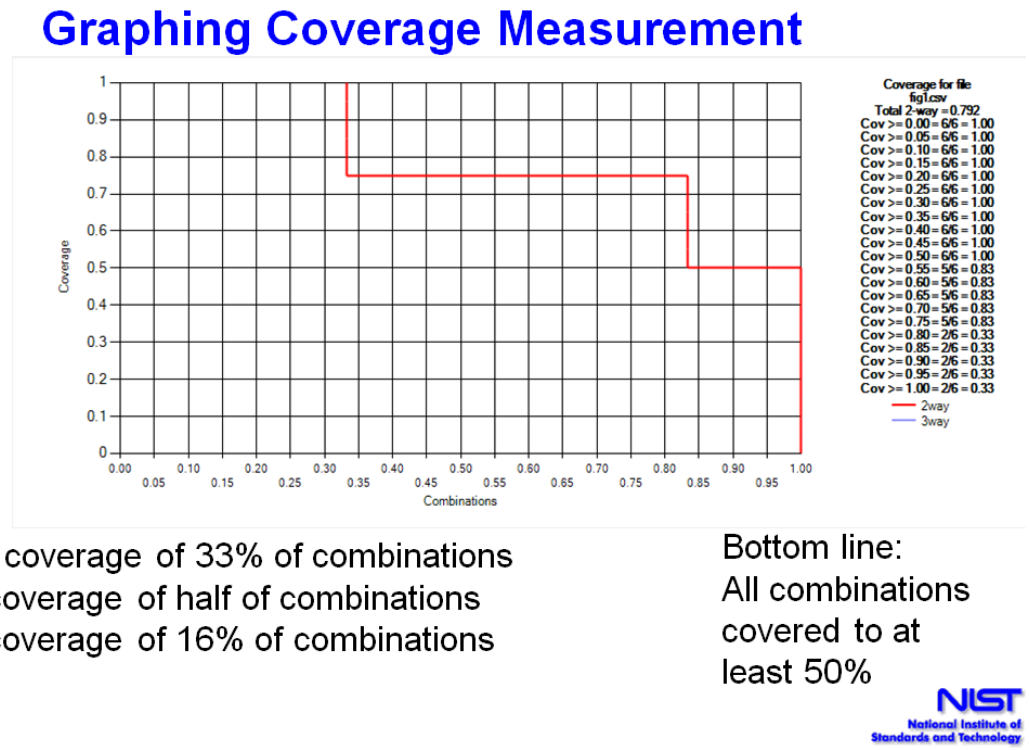

Figure 3.5. Combinatorial Coverage Measurement Chart [20]

This same tool can be applied to AI in an effort to improve "explainability" of autonomous software decision-making. The idea behind this new application is to determine what combination(s) of factors lead AI software to a certain decision. In some cases, the combinations the tester wishes to isolate may be those which cause the system to fail or arrive at the wrong conclusion, to determine faults in the system. In other instances, it may be more useful to know what combinations of factors lead AI software to the right conclusion via a process of elimination of incorrect factor combinations from tests. In either event, this tool could prove useful not only for AI explainability, but for debugging and model checking as well.

3.2.6. Formal Methods

There is an overwhelming need in the DoD for TEVV methods for autonomous systems that are both fast and rigorous. Formal methods based on mathematical techniques for specification, design, and TEVV of hardware and software are one possibility. Formal methods rely on formal logic, discrete mathematics, and computer-readable languages to support T&E. Formal methods are often employed for modeling requirements throughout the system, source and object code, and system architecture, as well as for analyzing compliance, traceability, robustness, completeness, and consistency of a system [10]. Because it is based in mathematical techniques, the logic used in formal analysis must be sound. Uses of formal methods tend to fall into the categories of proving theorems with axioms and inference, model checking by searching for counterexamples of model properties, and abstract interpretation to construct and analyze software representations.

The use of formal methods is often complicated due to the efforts that formal methods make to reconcile the discrete logic and discontinuous models of computer science with the physical laws and well-conditioned models of engineering. Additionally, the equations needed to apply formal methods typically employ millions of variables and often require iterative experimentation and T&E. However, formal methods are uniquely useful for improving the analysis of automated software and systems by removing much of the ambiguity that surrounds autonomous decision-making and behavior.

One tool actively in use is the SPARK programming language, used for functional verification of code by implementing low-level requirements and pre- and post-conditions of functions to ensure that scripts run as intended without generating overflow or math errors. SPARK can also perform analysis of data dependency contracts and verify that all data is traceable to system architecture and low-level requirements [10].

Formal methods can be of considerable help in the T&E of autonomous systems but will require better integration of source code and requirements into existing workflows, as well as better education and availability of case studies to testers. AFRL has worked to develop tools for increasingly complex architectures, as well as high- and low-level requirements of autonomous systems [10]. They have also developed error-finding and design automation tools.

It is possible to either supplement or replace formal methods with hybrid TEVV systems which make use of both continuous and discrete logic, assurance cases which are useful on a case-by-case basis, and runtime assurance for systems that are simply too complex to be verified with standard methods, formal or otherwise [10].

AFRL currently uses formal methods to verify UAVs, air and space collision avoidance methods, and algorithms implemented in autonomous systems [10]. In industry, Airbus, Amazon Web Services, Microsoft, and others have made use of formal methods to verify their autonomous system software [21] [22] [23] [24] [25].

3.3. Day 3: Current and Future Needs for the DoD to Effectively Test and Evaluate Autonomous Systems

Day 3 presentation generally supported Objectives 1 and 2 and are included in the discussion above. A focus of the panel discussion was current and future needs for the DoD to effectively test and evaluate autonomous systems. The highlights from all 3 panel discussions are detailed in the next section.

4. Panel Highlights

4.1. Introduction

Each workshop day ended with a Panel Discussion. The speakers of that day participated as panel members with active input from the balance of workshop attendees. Panel highlights are grouped based on recurring themes that presented during discussion.

4.2. Purpose and Definition of Autonomy

An autonomous system can have any of the following purposes: to replace a human at performing a task that is too "dull, dirty, or dangerous" for the human, or when the precision and/or response time demands are too high for a human to meet; or to replace an unmanned system in a mission where there is no guarantee of continuous communication with a human operator, thus necessitating the system's ability to make decisions. In the case of the autonomous system performing a task that would have otherwise been assigned to a human, the working group generated the following questions:

- Would the autonomous system need to be tested to prove that it performs at least as well as its human counterpart?
- How would such testing be done when human performance is generally not tested rigorously?
- Is this comparison necessary?
- Is there enough value in the autonomous system simply freeing up the human to perform other tasks that an autonomous system would not be capable of?

Workshop participants could not agree on a definition of autonomy. Several definitions were proposed, but no one of them mustered a consensus. There was some agreement about which specific examples discussed were autonomous systems, but attempts to offer a clear distinction between autonomy, automation, and artificial intelligence were strained. Properly addressing challenges of establishing frameworks for testing and evaluation of autonomous systems requires the autonomy T&E community to collectively establish a taxonomy that is useful to all.

4.3. Data

Data was highlighted as a vital resource in autonomy T&E. Data must be available to model an autonomous system's decision-making process, and the quality of any model built from empirical data is limited by the quality of that data. The difficulty of acquisition, ownership, storing, sharing, handling, and quality control of data are therefore major obstacles to the autonomy T&E community.

The practice of labeling data may improve the organization and allow quick reference. However, in order to be useful, the labels would most likely be mission-specific and therefore not easily reused. Furthermore, the labeling process highly time-consuming and labor intensive, perhaps even necessitating automation.

Collecting empirical data for testing and evaluating autonomous systems is expensive. Synthetic data may prove useful for augmenting empirical data thereby drastically reducing this cost. However, synthetic data will only be as good as the model that generates it. This raises the question of how to quantify the "goodness" of models and synthetic data.

To facilitate acquisition, availability, and ownership of data; participants recommended one central hub be created for data collection and storage, accessible broadly by the autonomy T&E community.

4.4. Modeling and Simulation

For validation and risk reduction of autonomous systems, claiming simulation is a useful tool is an understatement. Indeed, simulation is unavoidable and necessary to this end. However, simulation runs the risk of inaccuracy when failing to account for all relevant factors. A simulation is only as good as the model that informs it. It may be possible to augment simulations with virtual reality, or to test in a partially real and partially simulated environment. Formal methods may help to inform model construction as well.

4.5. Test and Evaluation of Emergent Behavior

More challenges identified and discussed during the panel sessions were concerned with testing and evaluation in general. Applying DOE to systems as complex as autonomous systems with potential emergent behaviors may be difficult; the more factors there are to examine, the design space expands, requiring additional test runs to characterize the system. Using continuous main effects and continuous interaction factors can reduce the number of runs to cover a test space. However, continuity of factor interactions cannot be assumed in the context of autonomy. Even more complexity is introduced when the system in question is a black box system.

The working group recommends a twofold approach: to "shift left," or test at every stage of development; and to focus on creating AI with explainable rationale, that is, to allow autonomous systems to be white box systems to reduce the difficulty of testing.

4.6. Safety, Assurance and Trust

Assurance can be shown through empirical data, but there is a distinction between assurance and trust. If the users of an autonomous system were to place an inappropriate level of trust or mistrust in an autonomous system, they would misuse, refuse to use, or abuse the system, preventing it from performing its intended task regardless of its capabilities. There is no known established metric for evaluating the level of trust human users have in autonomous systems. How to establish such a metric, how to instill the appropriate level of trust in autonomous systems by their users, and how to maintain this level of trust once it is reached are all potentially useful areas of study.

The working group suggested including users in testing to improve trust and provide the users more familiarity with the autonomous system to decrease the likelihood of inappropriate level of trust and consequently inappropriate use. Explainable AI would also benefit testing of autonomous systems.

4.7. Programs and Prototypes

The working group proposed a potential near-term approach to identify already existing pilot programs and prototypes for autonomous systems. These programs and prototypes would provide the benefit of being studied to answer the question, "What do right and wrong look like?" In return, the autonomy T&E community can assist with development, test planning, and analysis for these programs. Central funding for such an effort would be greatly helpful.

4.8. Centralized Effort

Perhaps the most important "next step" for the DoD autonomy T&E community is centralized coordination of effort. A central hub where data is stored, managed, and shared; funding, investments, acquisition, and program progress are tracked; a unifying framework for establishing requirements is defined; and experts from multiple disciplines can come together to exchange knowledge and ideas would grow the autonomy T&E community. Current lines of effort are quite stove piped within Military Services, perhaps born out of classification requirements and rapid acquisition emphases. It is clear there is no single agency guiding the development of these systems and there is no single vision on where the DoD is going.

5. Autonomy Test and Evaluation Collaboration

A critical component to forward the goals of this effort is an establishment of partnerships and collaboration efforts throughout the autonomy community. Policy and collaboration efforts have limited utility without their application to delivering combat capability to the warfighter. The STAT COE provides PhD-level T&E expertise for development of practical T&E guidance with specific actions to support programs. The STAT COE is actively seeking opportunities to work with DoD autonomy programs. Any DoD autonomous program is encouraged to avail itself of this consultation expertise.

5.1. Army Test and Evaluation Command (ATEC)

The STAT COE is currently establishing a test support role with the Army's Autonomous System Test Capability (ASTC). Initial conversations focused on STAT COE engagement with ASTC and emerging use cases. ASTC has integrated the Range Adversarial Planning Tool (RAPT), developed by Johns Hopkins University as discussed above, into their integrated virtual environments. While ASTC's capabilities have Joint applications, the initial engagement focused on two Army projects. The first is Leader/Follower, through Product Manager Appliqué and Large Unmanned Ground Systems, an emergent program linking unmanned Follower Palletized Load Systems (PLSs) to a soldier-operated Leader PLS vehicle for increased throughput and Soldier protection both on the road and off road [26]. The second is Project Quarterback, a nascent artificial intelligence effort through the Army Futures Command Next Generation Combat Vehicle Cross Functional Team. STAT COE participated in a workshop for ASTC at Aberdeen Proving Ground during the week of 9 December 2019 to refine the scope of involvement with these two projects. STAT COE members are currently in continuing discussions with ASTC to begin structuring test cases, scoring methodologies, sprint plans, and requirements prioritization for autonomous systems in the land domain.

5.2. Air Force Test Center (AFTC)

Following workshop networking, the STAT COE engaged with Emerging Technologies Combined Test Force (ET-CTF) at Edwards Air Force Base to discuss upcoming testing with the Air Force Research Laboratory Aerospace Systems Directorate. The STAT COE hosted a mission brief with ET-CTF and AFRL during the week of 25 November 2019 to identify potential test programs and opportunities to support. One outcome involved the Skyborg project. Skyborg is an autonomy and artificial intelligence "AFRL Vanguard" project that will integrate "attritable" unmanned combat air vehicles into operations in highly contested battlespace. The STAT COE is now consulting with Skyborg's Test Planning Working Group as the program prepares for upcoming flight test events.

5.3. Air Force Research Laboratory (AFRL)

STAT COE met with a group from AFRL/RW to discuss possible support for their upcoming test of a collaborative system, Golden Horde. The group consisted of Mr. Eddie McAllister, Gray Wolf test manager; Lt Col Olivia Elliott, AFRL/RW Test Lead; and Mr. Matt Alsleben, Golden Horde Test Manager. The work would include contributions to test planning for three upcoming separate

multiple weapon release tests, with the first test scheduled August 2020. The program was especially seeking assistance with data analysis and reporting, which could also involve work with GTRI.

5.4. Naval Sea Systems Command (NAVSEA)

Prior to the workshop, STAT COE participated in a teleconference led by the NAVSEA Deputy for Test and Evaluation (T&E) to discuss preparations for the NAVSEA Autonomy T&E Summit scheduled for early 2020. Topics included objectives, scope, format, and other elements necessary for successful collaboration and synergy within the NAVSEA Autonomous T&E Community.

STAT COE identified an opportunity to collaborate with the NAVSEA T&E community to inform the establishment of a standardized framework and best practices across the wider Department of Defense (DoD) autonomy community, particularly following the conclusion of our workshop. STAT COE observed that significant alignment exists between the goals and objectives of the ATEAS study and the goals that NAVSEA communicated for its upcoming autonomy summit, and that STAT COE may be able to contribute to the advancement of autonomy T&E through participation with NAVSEA. Expanding on that alignment yields the following areas as potential collaboration opportunities.

NAVSEA Autonomy T&E Summit has identified the following goals:

- Provide insight into the NAVSEA Enterprise efforts addressing autonomous systems T&E
- Establish framework to develop, standardize, and promulgate T&E policy, guidance, and best practices for NAVSEA program managers (PMs) executing autonomous systems acquisitions
- Ensure alignment of proposed NAVSEA autonomous systems T&E efforts with ongoing DoD and Department of the Navy (DoN) autonomy efforts.

NAVSEA also identified a working "problem statement" to guide the discussions: "NAVSEA autonomous systems T&E efforts (policy, procedures, workforce, [and] training) are disparate, uncoordinated, and not optimally organized to support NAVSEA autonomous systems acquisition efforts." STAT COE reviewed and provided feedback on their problem statement, objectives, taxonomy, and working definitions concerning autonomy T&E within the NAVSEA enterprise.

5.5. Unmanned Maritime Systems (UMS) Program Office (PMS 406)

STAT COE made initial contact with the Navy's PMS-406 Test and Evaluation Department and gave a STAT COE mission brief for their September 2019 Autonomy Workshop. STAT COE has had initial discussions with PMS-406 to assist with rigorous T&E of specific projects, such as Snakehead, Orca, and Razorback.

5.6. ATEAS Data Call to DoD, Academia, and Industry

STAT COE developed a data call and distributed it to workshop participants to identify organizational points of contact and solicit responses for answering the ATEAS study goals and

objectives. Attendees also sent out the data call to the wider test and evaluation community within their organizations. The data call is included as an appendix to this report.

The University of Memphis—FedEx Institute of Technology is developing a questionnaire to send out to industry. One of the key emphases of this effort was to reach outside of the DoD. This survey will allow the STAT COE to investigate current challenges and best practices in use across industry for test and evaluation of autonomous systems. The findings from both the STAT COE data call and University of Memphis questionnaire will be incorporated into the ATEAS final report.

6. Conclusions and Recommendations

The STAT COE conducted the ATEAS Workshop in October 2019 to 1) define and refine current challenges and gaps existing in DoD methods, processes, and test ranges capabilities to rigorously test and evaluate autonomous systems, and to 2) identify and develop methods, processes, and identify lessons learned needed to enable rigorous test and evaluation of autonomous systems. During the three-day workshop, selected experts from across the Department of Defense, academia, and industry provided insights into addressing the workshop goals and objectives. This was accomplished through presentations, questioning, and panel discussions with workshop attendees. That information exchange during the workshop advanced the understanding of challenges facing the autonomous systems test and evaluation community and provided an opportunity for several distinct follow-on engagements.

The workshop participants identified ten distinct challenge areas of test and evaluation of autonomous systems:

- Requirements and Measures
- Test Infrastructure and Personnel
- Design for Test
- Test Adequacy & Integration
- Testing Continuum
- Safety/Cybersecurity for Autonomous Systems
- Testing of Human Systems Teaming
- Post Acceptance Testing
- Simulation for Developmental Mission Assurance
- Data

The Workshop attendees' presentations, discussions, and panels demonstrated the original eight challenges identified in 2015 are still relevant in 2019, while also determining two additional challenge areas merited inclusion for a total of ten challenges. The follow-on data call may provide further information on these and identify other challenge areas as well. These ten challenge areas will continue to help frame future discussions and efforts in autonomy test and evaluation.

In June 2019, D(DT&E) sponsored a study to focus on Advancements in Test and Evaluation of Autonomous Systems (ATEAS) to support three lines of effort:

1. Assist acquisition programs and rapid fielding efforts in the use of innovative and efficient DT&E strategies to ensure production readiness and fielded weapon systems meet warfighter needs
2. Improve the Defense Acquisition T&E workforce "practice of the profession"
3. Advance test and evaluation (T&E) policy and guidance

Specific objectives of these lines of effort, not in priority order, are to:

- Collaborate with D(DT&E), Service components, OSD R&E Autonomy Community of Interest, and other DoD activities to develop methodologies and processes needed to conduct rigorous testing of autonomous systems
- Identify current best practices, lessons learned and new approaches (DoD and Industry) for testing and evaluation of autonomous systems
- Identify gaps and challenges in current DoD processes or methodologies to evaluate autonomous systems
- Identify current gaps in test ranges to evaluate autonomous systems
- Determine current demand for simulation and live testing, followed by identification of simulation and live testing demand for autonomous systems

To address the three lines of effort and satisfy the objectives above, the STAT COE conducted the Workshop and as follow-on actions intends to:

- Continue engagements to offer test planning and analysis to pilot programs from each Service. These engagements should specifically document best practices and case studies that put forth solutions that close the gaps identified by the challenge areas,
- Analyze of the data call responses and results,
- Develop an autonomy T&E taxonomy with associated lexicon and ontology to help guide the framework of future DoD autonomy T&E efforts,
- Create a funding plan for future STAT efforts within autonomy programs. To maintain an independent capability to provide test planning and analysis support, OSD and STAT COE should collaborate to scope and budget resources for pilot program engagement and assistance, and
- Assess the utility of future T&E Autonomy workshops/seminars.

The future of developmental mission assurance may require that autonomous systems require more "testing" than classical systems to define response surfaces from which the autonomy will make decisions. Therefore, use of efficient design of experiments test strategies and designs will be critical to this effort and this highlights the need to widely socialize continuing improvements recently made across the acquisition community. This includes the need for methods to improve automated software testing (AST), to increase effectiveness and repeatability of testing needed for autonomous systems. AST is an area of weakness in the T&E community. Finally, human-systems interface testing needs to be considered early for effective deployment and teaming as well as to be correctly considered into test planning and designs. Developing best practices, socializing lessons learned, and creating rigorous, efficient, and effective new methods for T&E are key to closing the gaps identified in the ten challenge areas and reduce the additional risks inherent to autonomous systems.

7. References

[1] R. Gates, *Science and Technology (S&T) Priorities for Fiscal Years 2013-17 Planning,* Washington, DC: Secretary of Defense, 2011.

[2] "DoD Directive 3020.40," Office of the Under Secretary of Defense for Policy, 29 November 2016. [Online]. Available: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/302040p.pdf?ver=2018-09-11-131221-983. [Accessed 24 January 2020].

[3] "DoD Instruction 5000.02T," Department of Defense, 26 November 2013. [Online]. Available: https://www.acq.osd.mil/fo/docs/DSD%205000.02_Memo+Doc.pdf. [Accessed 24 January 2019].

[4] D. Ahner, "Test and Evaluation of Autonomous Systems," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[5] Defense Acquisition University (DAU), "DAU Glossary," Defense Acquisition University (DAU), [Online]. Available: https://www.dau.edu/tools/t/DAU-Glossary. [Accessed 18 December 2019].

[6] S. Redfield, "Verification of Autonomous Systems: Challenges and Best Practices," in *Advancements in Test and Evaluation of Autonomous Systems*, Dayton, 2019.

[7] G. Zacharias, "Autonomous Systems: Test and Evaluation Implications," in *Advancements in Test and Evaluation of Autonomous Systems*, Dayton, 2019.

[8] R. McAllister, "Results from PMS 406/Warfare Center UUV/USV Workshop Conducted 22 Mar 2019," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[9] B. Nowotny, "Current Capabilities and Gaps at Ranges for T&E of Autonomous Systems," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[10] L. Humphrey, "Formal Methods for V&V and Certification," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[11] L. Freeman, "Artificial Intelligence as a Change Agent for Test and Evaluation," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[12] D. Tate, "What Counts as Progress in the T&E of Autonomy," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[13] C. Bieber, H. Wojton, D. Porter and M. McAnally, "Preparing for the Future: A Test Framework for AI and Autonomy," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[14] J. Simpson, "Sequential Testing and Simulation Validation for Autonomous Systems," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[15] J. Wisnowski, "How to Effectively Inject STAT/DOE into Autonomy T&E," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[16] J. Pinelis, "Challenges in T&E of AI: DoD's Project Maven," in *Advancements in Test and Evaluation of Autonomous Systems*, Dayton, 2019.

[17] P. Stankiewicz and J. Horris, "Range Adversarial Planning Tool (RAPT)," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[18] K. Thuloweit, "AFIT alum discusses work on autonomous flight testing," Air Force Institute of Technology, 07 March 2019. [Online]. Available: https://www.afit.edu/news.cfm?article=876. [Accessed 2 November 2019].

[19] P. Kwashnak, "Enabling Effective T&E of Autonomous Systems Using Digital Modeling and Simulation," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[20] R. Kacker and R. Kuhn, "Explainable AI, Combinatorial Methods, and Autonomous Systems Assurance," in *Advancements in Test and Evaluation of Autonomous Systems Workshop*, Dayton, 2019.

[21] J. Backes, P. Bolignano, B. Cook, C. Dodge, A. Gacek, K. Luckow, R. Neha, O. Tkachuk and C. Varming, "Semantic-based Automated Reasoning for AWS Access Policies using SMT," in *Formal Methods in Computer Aided Design (FMCAD)*, Austin, 2018.

[22] T. Ball, B. Cook, V. Levin and S. K. Rajamani, "SLAM and Static Driver Verifier: Technology Transfer of Formal Methods inside Microsoft," *Integrated Formal Methods,* vol. 2999, pp. 1-20, 2004.

[23] A. Chudnov, N. Collins, B. Cook, J. Dodds, B. Huffman, C. MacCarthaigh, S. Magil, E. Mertens, E. Mullen, S. Tasiran, A. Tomb and E. Westbrook, "Continuous Formal Verification of Amazon s2n," in *Computer Aided Visions*, 2019.

[24] C. Dodge and S. Quigg, "A simpler way to assess the network exposure of EC2 instances," in *AWS Security Blog post*, 2019.

[25] J. Souyris, V. Wiels, D. Delmas and H. Delseny, "Formal Verification of Avionics Software Products," in *FM 2009: Formal Methods*, Dayton, 2019.

[26] U.S.Army, "Product Manager Applique and Large Unmanned Ground Systems," Program Executive Office Combat Support and Combat Service Support, [Online]. Available: https://www.peocscss.army.mil/pdmalugs.html. [Accessed 6 Dec 2019].

Appendix 1: ATEAS Data Call

Advancements in Test and Evaluation of Autonomous Systems Data Call

A 2015 workshop held at the Air Force Institute of Technology (AFIT) assessed the current state of the art in the test and evaluation (T&E) of autonomous systems. The workshop identified nine major challenges to the T&E of autonomous systems. To aid in the accurate completion of this data call each challenge is listed below with a short description.

- **Requirements and Measures – T&E of autonomous systems needs to assess the system's ability to successfully perform the required tasks and functions it was created to perform, and to evaluate the system's decision-making capability. Objectives and metrics are the means used to measure the success of the system in performing its task.**

- **Test Infrastructure and Personnel – T&E of autonomous systems requires test methods, processes, strategies, physical ranges that may not currently exist, and the pool of personnel available to develop those tools and capabilities. Personnel may need to be recruited and trained.**

- **Design for Test – T&E of autonomous systems is made difficult by the system's internal decision-making process operating as a black box. Systems need to be designed in such a way that adequate T&E can occur even without full knowledge of the inner workings of the system.**

- **Test Adequacy & Integration – T&E of autonomous systems is inherently difficult due to a lack of information that would allow testers to adequately define and quantify risk and performance. The challenge is further compounded due to the dynamic nature of the system, which may present the need for a change in testing or additional testing.**

- **Testing Continuum – T&E of autonomous systems presents the challenge of not having a pre-determined test phase but requiring testing throughout the system life cycle. It is necessary to decide when, how, and what aspects of a system to test under initial assumptions; how to retest after learning has occurred; and how to use that information to make decisions.**

- **Safety/Cybersecurity for Autonomous Systems – T&E of autonomous systems will need to provide testers (and ultimately users) with assurance that the system will not perform actions that are deemed unsafe or undesired. It will also need to provide assurance that the system is reasonably resistant to physical and cyber attacks.**

- **Testing of Human System Teaming – T&E of autonomous systems will need to address the ability of any combination of humans and machines to perform as partners and determine how to measure the effectiveness of that team.**

- **Simulation for Developmental Mission Assurance - Due to the interplay of sensors, software, and human agents, autonomous systems represent a complex development environment. Modeling and simulation can be a critical component for mission assurance.**

- **Post Acceptance Testing – T&E of autonomous systems requires testing throughout the life of the system (see Testing Continuum challenge). Methods will need to be developed to test, evaluate, use, and update systems already in the field.**

For further clarification of any of the challenges described or any of the questions that follow, or to obtain a copy of the 2015 workshop report, please contact Dr. Steve Oimoen, Ms. Kaitlyn Jones, or Ms. Emily Divis (STAT COE contractors) at steven.oimoen.ctr@afit.edu, kaitlyn.jones.ctr@afit.edu, or emily.divis.ctr@afit.edu, respectively.

OBJECTIVE 1: Identify and develop methods and processes, and identify lessons learned needed to enable rigorous test and evaluation of autonomous systems.

OBJECTIVE 2: Refine current challenges/gaps in DoD methods, processes, and test ranges to rigorously test and evaluate autonomous systems.

O1 1. Identify, describe, and provide any methods, processes, or frameworks that you currently use to assure proper autonomous systems **development**.

O1 2. Describe the methods and processes which you currently employ to **test** autonomous systems and indicate which are most effective and informative.

O1 3. Please identify any gaps that currently exist in DoD methods or processes to test autonomous systems.

O1 4. Please identify any gaps that currently exist in DoD test ranges to test autonomous systems.

O2 5. Please identify any challenge areas (not listed above) encountered when testing autonomous systems.

For the following questions, prioritize (high, medium, low, N/A) the importance of each gap in each challenge based on how it affects your work. List and provide a short description for any missing gaps.

O2 6. Prioritize the challenges in testing and evaluating autonomous systems identified in the 2015 workshop that affect how you test and evaluate autonomous systems.

- Requirements and Measures
- Test Infrastructure and Personnel
- Design for Test
- Test Adequacy & Integration
- Testing Continuum
- Safety/Cybersecurity for Autonomous Systems
- Testing of Human System Teaming
- Simulation for Developmental Mission Assurance
- Post Acceptance Testing
- List other challenges in testing and evaluating autonomous systems that were not captured above and rate their importance to your work.

O2 7. Prioritize the following gaps in Requirements and Measures that affect how you test and evaluate autonomous systems.

- How to quantify success of decision-making
- How to measure perception, reasoning, and learning
- How to define unique metrics (i.e. trust, intent, system learning, perception, reasoning, distributed perception, distributed decision, doing the "right thing" for the "right reason")
- How to define negative requirements
- List other gaps in Requirements and Measures and rate their importance to your work.

O2 8. Prioritize the following gaps in Test Infrastructure and Personnel that affect how you test and evaluate autonomous systems.

- Community of Practice and mechanisms for sharing knowledge, M&S, instrumentation, environments, and data between the S&T and T&E communities
- Updated or new business model that adequately incentivizes all community stakeholders
- Sequence career path for "lifelong internships" across research, development, T&E, etc.
- Personnel gap analysis, resources for workforce development
- Structured approach to keeping the workforce current
- Alignment with adjacent disciplines
- Infrastructure for requirements, instrumentation, and measures
- List other gaps in Test Infrastructure and Personnel and rate their importance to your work

O2 9. Prioritize the following gaps in Design for Test that affect how you test autonomous systems.

- Standard framework or architecture with a T&E application program interface
- Formal models that bridge the gap from requirements to design
- Understanding of the human involvement with the framework so we know how to incorporate autonomy
- Direct impact of the Test Design gaps on requirements and their connection to the requirements challenge
- List other gaps Design for Test and rate their importance to your work

O2 10. Prioritize the following gaps in Testing Adequacy and Integration that affect how you test autonomous systems.

• How much testing is enough? Necessary and Sufficient

• Sequential test design in near real time

• Coverage of testing space given changes over time

• Process to identify integration points across the life cycle

• "Composability" of test results with other types of evidence

• New "statistical engineering" techniques; combine empirical with non-empirical

• How to assess swarm or team dynamics, including distributed perception, shared knowledge, and distributed decision making

• List other gaps in Testing Adequacy and rate their importance to your work

O2 11. Prioritize the following gaps in the Testing Continuum that affect how you test autonomous systems.

• Organizations and practical ability to coordinate and build on learning (i.e. cross pollinate testing, test data management system)

• Community of practice with agreed-upon goals and reciprocal sharing of data and outcomes to accelerate advancement

• Compositional analysis of data produced from incremental testing

• Processes, infrastructure, and data standards for the warehousing and sharing of test data

• Standards and agreements for reciprocity of performance/safety measures and licensing across government agencies and industry

• List other gaps in Testing Continuum and rate their importance to your work

O2 12. Prioritize the following gaps in Safety/CyberSecurity that affect how you test autonomous systems.

• New/updated Institutional Review Board policies/methods/processes

• Formal argument generators

• Tools to make tradeoffs between assurance and performance

• Cyber self-diagnostics (tests for them, introspection/system assesses own risk)

• Transparency of intent

• Classification level

• Graceful degradation and recovery

• Ability to turn various levels of autonomy on/off

• List other gaps in Safety/CyberSecurity and rate their importance to your work

O2 13. Prioritize the following gaps in Human System Teaming that affect how you test autonomous systems.

• Trust metrics & assessment

• Method for evaluating varying levels of autonomy

• List other gaps in Human System Teaming and rate their importance to your work

O2 14. Prioritize the following gaps in Simulation for Developmental Mission Assurance that affect how you test autonomous systems.

• Mission requirements changing as system adapts to its environment

- Inability of hardware to replicate complex operational environment for sensor testing
- Impossibility of modeling operational environment to anticipate all possible inputs in a virtual test
- List other gaps in Simulation for Developmental Mission Assurance and rate their importance to your work

O2 15. Prioritize the following gaps in Post Acceptance testing that affect how you test autonomous systems.

- Structured training events after fielding for each new environment and/or system
- Health maintenance system for field-based learning (i.e. periodic or trigger-driven; licensure (potentially need progressive licensure)
- Real-time data capture and analysis
- Infrastructure and methods for test community to tap experiential data stemming from deployed systems
- List other gaps in Post Acceptance testing and rate their importance to your work

Please add any continuations or additional comments here.

Appendix 2: List of Acronyms and Abbreviations

AFRL – Air Force Research Laboratory

AI – Artificial Intelligence

AS – Autonomous System

ASTC – Autonomous System Test Capability

ATEAS – Advancements in Test and Evaluation of Autonomous Systems

ATEC – Army Test and Evaluation Command

COE – Center of Excellence

COLREGS – Collision Regulations

CONOPS – Concept of Operations

D(DT&E) – Director, Developmental Test and Evaluation

DoD – Department of Defense

DOE – Design of Experiments

DRIVE – Digital RAS Integrated Virtual Environment

DTEP – Developmental Test, Evaluation, and Prototyping

ET CTF – Emerging Technologies Combined Test Force

FY – Fiscal Year

GPS – Global Positioning System

HPC – High Performance Computer

HWIL – Hardware in the Loop

JAIC – Joint Artificial Intelligence Center

JHU APL – Johns Hopkins University Applied Physics Laboratory

JMETC – Joint Mission Environment Test Capability

LVC – Live Virtual-Construction

M&S – Modeling and Simulation

MCM – Mine Countermeasures

MILS – Multiple Independent Levels of Security

MoEs – Measures of Effectiveness

MoPs – Measures of Performance

MRTFB – Major Range and Test Facility Base

NAVSEA – Naval Sea Systems Command

NCR – National Cyber Range

NIST – National Institute of Standards and Technology

OSD – Office of the Secretary of Defense

PM – Program Manager

PMS – Program Manager, Ships

RAPT – Range Adversarial Planning Tool

RAS – Robotic and Autonomous Systems

RSDP – Regional Service Delivery Points

SEER – Safety, Environment, Engagement, and Response

SLAM – Simultaneous Localization and Mapping

STAT – Scientific Test and Analysis Techniques

T&E – Test and Evaluation

TACE – Testing of Autonomy in Complex Environments

TEVV – Test, Evaluation, Verification, and Validation

TRMC – Test Resource Management Center

TTPs – Tactics, techniques, and procedures

UAV – Unmanned Aerial Vehicle

UGV – Unmanned Ground Vehicle

UMS – Unmanned Maritime Systems

USD R&E – Undersecretary of Defense, Research and Engineering

USV – Unmanned Surface Vehicle

UUV – Unmanned Underwater Vehicle

Appendix 3: List of Attendees and Contact Information

A list of attendees and contact information is available in a Distribution D report, which may be requested by contacting COE@afit.edu.