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

RESEARCH
AND ENGINEERING

JUL 20 2016

MEMORANDUM FOR TEST AND EVALUATION PROFESSIONALS

SUBJECT: Test and Evaluation of Autonomous Systems

Autonomous defense systems have gained great interest in recent years and the Deputy Secretary of Defense has identified autonomy as the critical technology in the Department's Third Offset Strategy. Along with the new capabilities that these systems offer, autonomous defense weapon systems pose new challenges in achieving safe, effective, and reliable performance in general, and in conducting adequate developmental and operational test and evaluation in particular.

The Test & Evaluation Center of Excellence for Scientific Test and Analysis Techniques conducted a workshop for the Test & Evaluation of Autonomous Systems 24-25 August 2015 addressing "How to conduct test and evaluation of autonomous systems and what specific testing methodologies and capabilities need to be addressed?"

In addition to addressing developmental testing, this workshop, through participation by its unique set of attendees, provided an initial framework in which to build a sequential assurance case for safe, effective, and reliable autonomous systems.

Developmental testing is a key component in obtaining accurate information to inform technical and acquisition decisions, and more generally, the systems engineering process. While many strides have been made in the last few years in the areas of developing; policy, establishing the chief developmental tester as a key leader position, creating workforce development efforts, and providing major acquisition programs access to scientific test and analysis technique experts, shortcomings in early-on assurance that our weapon systems work properly and provide the capability persist. Autonomous systems will only exacerbate any shortcomings. This workshop is an early step to address these autonomous systems challenges.

David Brown

Dr. C. David Brown
Deputy Assistant Secretary of Defense for
Developmental Test and Evaluation

STAT COE-Report-01-2016

Workshop Report: Test and Evaluation of Autonomous Systems

Darryl K. Ahner, PhD, PE

Carl R. Parson, PhD

STAT T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

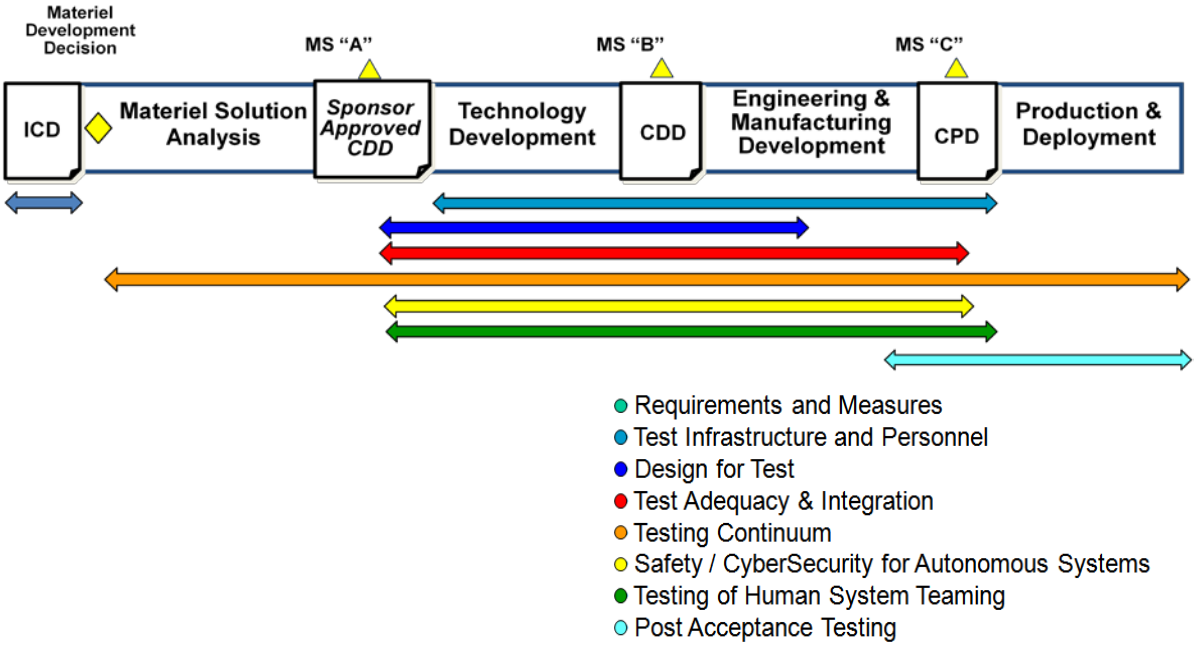
Autonomous defense systems have gained great interest in recent years. Additionally, autonomous defense systems most likely will need to operate in unstructured, dynamic environments. Along with the new capabilities that these systems offer, autonomous defense systems pose new challenges to conducting adequate developmental and operational test and evaluation. The Office of the Secretary of Defense (OSD) Scientific Test and Analysis Techniques in Test & Evaluation Center of Excellence (STAT in T&E COE) conducted a workshop for the Test & Evaluation of Autonomous Systems on 24-25 August 2015 to address "How to conduct test and evaluation of autonomous systems and what specific testing methodologies and capabilities need to be addressed?"

Several previous and ongoing studies provide a foundation from which to address the challenges of conducting adequate developmental and operational test and evaluation for autonomous defense systems. The 2012 Defense Science Board study "The Role of Autonomy in DoD Systems" highlights many major challenges with acquisition and operational use of autonomous systems. This study provides a significant stepping off point for exploring what requirements, processes, and methods may be required to adequately test autonomous systems.

Workshop participants identified several unique challenges that were subsequently categorized into eight challenge areas for test and evaluation of autonomous systems. These categories consist of both what makes the test and evaluation of complex systems difficult, in general, and what makes test and evaluation of autonomy, in particular, difficult. The challenges and current shortcomings of test and evaluation of complex systems may become more pronounced when coupled with autonomy. The challenge areas identified are:

- 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
- Post Acceptance Testing

No one challenge was identified as being more important than the other as they span the acquisition life on any autonomous system acquisition program as shown in the figure below.


Materiel Development Decision
MS "A"
MS "B"
MS "C"
ICD
Materiel Solution Analysis
Sponsor Approved CDD
Technology Development
CDD
Engineering & Manufacturing Development
CPD
Production & Deployment
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
Post Acceptance Testing

The workgroup developed a consensus on the eight challenge areas. Shortfalls in executing current acquisition programs for complex systems were identified as even more critical for future autonomous systems. Additionally, new challenges were identified that will need to be addressed through funding for T&E research, technology development, capability development, and infrastructure. Additionally, changes in policy will be required, along with workforce development efforts in order to adequately perform testing to inform key technical, programmatic and acquisition decisions throughout the lifecycle of the program. Among efforts requiring test research funding are:

- Improved statistical engineering methods are needed to support both developmental and operational testing of autonomous systems to address systems interacting with a dynamic environment in a non-deterministic manner. Improvement of these methods is a component of a larger need; these adaptive autonomous systems will require more stringent adherence to systems engineering principles throughout development.
- Processes and methods need to be developed to address Inputs – Process – Outputs of autonomous systems and human-machine element interaction and roles within the process.
- A test and evaluation continuum paradigm must be developed and adopted that requires testing start early and a more sequential progressive approach is taken that includes development and implementation of a comprehensive M&S strategy across the life cycle. There is no "test phase" with a beginning or end; it extends throughout the life cycle of the system.
- Measures must be developed to address state space adequacy, trust, and human-machine interaction.

- Design of experiment methods must be developed for defining test cases and expected results that overcome the difficulty of enumerating all conditions and non-deterministic responses that autonomy will generate in response to complex environments.
- Models and live virtual constructive (LVC) test beds are needed that support robust testing while minimizing risk and cost.
- Development of techniques that capture learning growth, possibly similar to reliability growth models is needed.

The primary challenge in testing autonomous systems is the broad scale and complexity of the systems, missions, and conditions. This is best addressed by breaking down the requirement, system, and/or mission into smaller pieces, which can then be readily translated into rigorously quantifiable statistical designs. One large comprehensive experiment is unlikely to inform technical, programmatic, and acquisition decisions adequately. A progressive sequential approach to testing is a better strategy that the workshop participants embraced.

While the workshop participants rigorously addressed testing of autonomous systems and research from a current perspective to advance the state of the art, unknown unknowns will surface as more advanced autonomous systems are developed. Test and evaluation lessons learned, best practices, and identification of testing challenges from early autonomous systems development will be critical in creating efficient and effective testing to understand and reduce the significant additional risks of autonomous systems.

1. Introduction

A. Background

Autonomous defense systems have gained great interest in recent years. In a April 19, 2011 memo, the Secretary of Defense designated Autonomy as one of the seven priority S&T investment areas in the FY 13-17 Program Objective Memorandum. In response, the Assistant Secretary of Defense, Research and Engineering (ASD(R&E)) set up the Autonomy Community of Interest resulting in much thought and study being conducted when considering the use of Autonomy in DoD systems. Acquisition of autonomous defense systems are considered for dirty, dull, or dangerous mission segments in addition to potentially decreasing personnel costs. Tomorrow, autonomous systems will perform a litany of undertakings, to include among other activities, augmenting human decisions, teaming with humans in the execution of missions, teaming with other systems for collaborative autonomous functions, and extended intelligence, surveillance, and reconnaissance (ISR). Autonomy enables a particular action of a system to be automatic or, within programmed boundaries, self-governing, while not necessarily replacing the human operator, but may extend and complement their capabilities to conduct complex, large-scale, long duration operations without sacrificing safety or effectiveness, and must allow Warfighters to focus on their primary mission, not on operating their tools. Additionally, autonomous defense systems most likely will need to operate in unstructured, dynamic environments. Along with the new capabilities that these systems offer, autonomous defense systems pose new challenges to conducting adequate developmental and operational test and evaluation. The OSD Scientific Test and Analysis Techniques in Test & Evaluation Center of Excellence (STAT in T&E COE) conducted a workshop for the Test & Evaluation of Autonomous Systems on 24-25 August 2015 to address “How to conduct test and evaluation of autonomous systems and what specific testing methodologies and capabilities need to be addressed?”

B. OSD Scientific Test and Analysis Techniques in Test & Evaluation COE

The Deputy Assistant Secretary of Defense for Developmental Test and Evaluation, DASD(DT&E), in collaboration with the U.S. Air Force Commander Air Education and Training Command, established the STAT in T&E COE in April 2012 under the stewardship of the Air Force Institute of Technology (AFIT). The STAT in T&E COE provides advice and assistance to designated major acquisition programs in the application of scientific test and analysis techniques in the development of test & evaluation strategies and plans. It performs this mission through personnel with PhD level technical STAT skills coupled with acquisition experience. Among the center’s functions are to work directly for Program Managers in supporting their efforts for more rigorous testing and evaluation, capture implementation of STAT best practices for wider dissemination throughout the acquisition community, develop case studies that exemplify appropriate use of STAT in achieving more rigorous testing and evaluation, provide technical assistance to the DASD(DT&E) as requested, identify STAT research needs and communicate them to the academic community, and provide training at the point of need to ensure program-led rigor in testing. The STAT in T&E Center of Excellence’s goal is to have all programs use scientific test and analysis techniques to ensure testing produces quality data that informs Better Buying Power decisions.

2. Previous Studies

Several previous and ongoing studies provide a foundation from which to address the challenges of conducting adequate developmental and operational test and evaluation for autonomous defense systems. The 2012 Defense Science Board (DSB) study "The Role of Autonomy in DoD Systems" highlights many major challenges with acquisition and operational use of autonomous systems. This study provides a significant stepping off point for exploring what requirements, processes, and methods may be required to adequately test autonomous systems. In addition, OSD AT&L has directed a DSB 2015 Summer Study on Autonomy. Recently, the OSD Autonomy Community of Interest (COI) published its Test and Evaluation, Verification and Validation (TEVV) Working Group Technology Investment Strategy. Below provides an overview of some key points from these documents as they apply to test & evaluation.

A. DSB 2012 Study on the Role of Autonomy in DoD Systems

Overall study Summary:

- Unmanned systems are having a worldwide impact (offensive and defensive) across the DoD, but we are operating in relatively benign conditions and at the initial stages of innovation of autonomy
- Main benefits of UxS* are to extend and complement human performance, not provide a direct replacement of humans
- Principal recommendation for capturing additional benefits of autonomous systems include:
  - Abandon definitions of levels of autonomy and replace with the autonomous systems reference framework. Use the framework to shape technology programs and to make key decisions for the design of future systems
  - ASD(R&E) should work with Services to establish a coordinated S&T program guided by feedback from operational experience and evolving mission requirements
  - Joint Staff and Services should improve the requirements process to develop a mission capability pull for autonomous systems
  - *USD(AT&L) to create developmental and operational T&E techniques that focus on the unique challenges of autonomy* (to include developing operational training techniques that explicitly build trust in autonomous systems)
  - DIA and Intelligence Community to track adversary capabilities for autonomous systems. Include these threats in war games, training, simulations and exercises

T&E Study Recommendations:

USD(AT&L) should establish developmental and operational T&E techniques that focus on the unique challenges of autonomy

- Coping with the difficulty of enumerating all conditions and non-deterministic responses
- Basis for system decisions often not apparent to user

- Measuring trust that the autonomous system will interact with its human supervisor as intended
- Expanding the test environment to include direct and indirect users (human supervisors, higher level command, etc.)
- Leverage the benefits of robust simulation

USD(AT&L) should task acquisition programs to assess vulnerabilities of U.S. systems to physical, jamming and cyber attacks.

B. Autonomy COI: TEVV Working Group Technology Investment Strategy, 2015-2018

In an April 19, 2011 memo, the Secretary of Defense designated Autonomy as one of the seven priority S&T investment areas in the FY 13-17 Program Objective Memorandum. In response, the Assistant Secretary of Defense, Research and Engineering (ASD(R&E)) set up the Autonomy Community of Interest and identified four challenge areas and designated a working group for each:

- Human/ Autonomous Systems Interaction and Collaboration
- Scalable Teaming of Multiple Autonomous Systems
- Test & Evaluation and Verification & Validation (TEVV)
- Machine Reasoning, Perception, and Intelligence

The purpose of the TEVV strategy is to lay the groundwork for a technology roadmap, identifying research objectives to further the state of the art in TEVV of Autonomous Systems. Definition of Automation and Autonomy from strategy –

Automation: The system functions with no/little human operator involvement; however, the system performance is limited to the specific actions it has been designed to do. Typically these are well-defined tasks that have predetermined responses (i.e., simple rule-based responses).

Autonomy: The system has a set of intelligence-based capabilities that allows it to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous systems have a degree of self-government and self-directed behavior (with the human's proxy for decisions).

The report highlights state-space explosion, unpredictable environments, emergent behavior, and human-machine communication as key challenges. The investment strategy identifies several gaps:

- Lack of Verifiable Autonomous System Requirements
- Lack of Modeling, Design, and Interface Standards
- Lack of Autonomy Test and Evaluation Capabilities
- Lack of Human Operator Reliance to Compensate for Brittleness
- Lack of Run Time V&V during Deployed Autonomy Operations
- Lack of Evidence Re-use for V&V

The investment strategy aligns DoD research in TEVV of Autonomy around the following goals:

TEVV Goal 1: Methods & Tools Assisting in Requirements Development and Analysis
TEVV Goal 2: Evidence-Based Design and Implementation
TEVV Goal 3: Cumulative Evidence through RDT&E, DT, & OT
TEVV Goal 4: Run Time Behavior Prediction and Recovery
TEVV Goal 5: Assurance Arguments for Autonomous Systems

3. Workshop Objectives

The STAT in T&E COE Workshop on Test & Evaluation of Autonomous Systems focuses primarily on TEVV Goal 3. This goal's emphasis is on progressive sequential modeling and simulation (M&S), and test and evaluation. M&S and T&E at each Technical Readiness Level (TRL) and product milestone, when planned and conducted properly, provide an invaluable resource not only to verify and validate that a system satisfies the user requirements, but also to aid in technology development and maturation. The development of effective methods to record, aggregate, and reuse T&E results remains an elusive and technically challenging problem. Methods must be developed to record, aggregate, leverage, and reuse M&S and T&E results throughout the system's engineering lifecycle; from S&T experimentation and TRL assessment, to requirements, to model-based designs, to live-virtual constructive (LVC) experimentation, to open-range testing, and finally post deployment.

The Deputy Assistant Secretary of Defense for Developmental Test and Evaluation, DASD(DT&E), has championed 'shift left' in order to improve DT&E to set the conditions for successful production and deployment. 'Shift left' techniques focus on performance, reliability, interoperability, and cybersecurity while conducting DT&E in a mission context. Tools of the 'shift left' initiative include constructing development evaluation frameworks to map testing to informing technical, program, and acquisition decisions and the use of scientific test and analysis techniques (STAT) to ensure test resources are used as efficiently and effectively as possible while providing decision quality information. The intent of 'shift left' is to set the conditions for improved production readiness and reduce the likelihood that major deficiencies get to the field.

With this framework in mind, the OSD STAT COE Workshop on Test & Evaluation of Autonomous Systems had the following objectives:

- Identify the challenges of testing and evaluating autonomous systems
- Identify existing test & evaluation requirements / processes / methods to address each challenge
- Identify gaps in the requirements / processes / methods which will need to be addressed for adequate test & evaluation of autonomous systems

Since these objectives are excessively broad for a two day workshop, the objectives were further scoped to focus on test and evaluation on autonomous systems over the next 10 years or until 2025.

4. Identified Challenges

A. Introduction to Challenges

Workshop participants identified several unique challenges that were subsequently categorized into eight challenge areas for test and evaluation of autonomous systems. These categories consist of both what makes the test and evaluation of complex systems difficult, in general, and what makes test and evaluation of autonomy, in particular, difficult. The challenges and current shortcomings of test and evaluation of complex systems may become more pronounced when coupled with autonomy. The identified 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 System Teaming
- Post Acceptance Testing

In the following sections, these challenge areas are explained further, with the specific challenges being enumerated explicitly. A discussion of the challenges is presented, followed by a discussion of the existing resources used to address those challenges. Finally, the technology and capability gaps are identified and detailed.

B. Requirements and Measures

I. Challenge Description

When quantifying the performance of autonomous systems, many difficulties arise. Test and evaluation must address the inputs, internal processes and states, and outputs for all autonomous functions employed by the system (e.g. perception, reasoning, learning, decisions, and behavior) as well as overall system performance to inform acquisition decisions and operational risk. For illustration purposes, decisions of a fully autonomous system may be viewed through decision theory. Using decision theory, autonomy addresses a decision problem

$$D = \langle A, X, \Omega, u, w, I, C(I) \rangle$$

where

A: action space

X: state space

Ω: outcome space

u : utility function

w: initial wealth (resources)

I : available information

C(I): information cost function.

Each of these elements must be understood in order to quantify performance of the decision engine, to inform acquisition decisions and operational risk, and ultimately to build trust from the warfighter that the system will perform as intended. This situation presents challenges to defining requirements in ways that are clear and testable, including in a well-defined and comprehensive operational mission environment. It also presents challenges to developing metrics to adequately test and evaluate these systems.

Although it may be necessary to evaluate the autonomous system's decision making capability, the top level requirement should always be for the system to successfully perform the tasks and functions it was built to perform. The top level test objectives and metrics need to be focused on the evaluation of the system's ability to accomplish the required tasks and functions (for a given set of input conditions, what will be the system's ability to accomplish its performance objectives). Starting with the top level requirement will help ground the test and evaluation activates with currently understood processes. The greater challenges will come with trying to tie the evaluation of top level performance to the systems performance relative to decision making.

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Requirements and Measures identified by the working group consist of:

- Requirements which are described with operational intentions
- Unsupervised learning methods
- Software requirements testing
- Communities of Interest/Critical Operational Issues and Criteria/Measures of Effectiveness
- JCIDS/Gap Analysis
- Fitness functions (weighting of potential decisions)
- Formal language requirements derived from natural language specs

Gaps for Requirements and Measures identified by the working group consist of:

- How to quantify success of decision making
- How to measure perception, reasoning, and learning
- Unique metrics
    - Trust
    - Intent
    - System learning
    - Perception
    - Reasoning
    - Distributed perception

  - Distributed decision
  - Doing the “right thing” for the “right reason”
- Negative requirements

Measures are needed to test and evaluate the adequacy of the perception global view that is commonly referred to as the state space in control theory. A parsimonious state space is often desired for memory storage and accessibility reasons; however, determining if the state space is adequate for the full range of situations and uncertain operational environments is difficult. Development of formal language requirements derived from natural language specifications is seen as a possible way forward.

Likewise, interoperability standards for communication can play an important role in enabling synergistic effects from autonomous systems. This will allow greater complexity of operations as autonomous systems and humans can coordinate activity and solve problems collaboratively. Examples from current DoD systems demonstrate that imposing interoperability standards before developing systems is key to prevent major challenges and costs for retrofitting or redesigning systems.

Autonomous software takes on various cognitive functions such as perception, comprehension, and projection and interacts with humans in ways that require evaluation methods that are not yet fully understood or developed. Understanding and evaluating the behavior of decision-making necessarily requires evaluating the allowable action space, the requirement of a functional but parsimonious world view, the understanding of commander's intent, coherence to rules of engagement, effects of system faults both of the autonomous and physical system, and the effect of resource utilization. This understanding and evaluation become even more complex when considering cooperative large numbers of homogeneous or heterogeneous autonomous systems performing tasks or missions.

C. Test Infrastructure and Personnel

I. Challenge Description

Two primary challenges were identified during the workshop that deal with the unique issues test of autonomous systems may have regarding infrastructure and personnel. Many of the necessary processes, systems, test infrastructure, and other capabilities simply do not exist, and may not have even been conceived of at this point. Additionally, the personnel needed to develop these tools must be professionally developed, trained, and recruited to ensure the DoD leads this effort.

First, there is a need to identify and develop the necessary skill sets within existing personnel, and recruit qualified candidates to become experts in test and evaluation of autonomous systems. This challenge has two key elements: training the existing and future workforce, and developing new T&E methodologies relevant to autonomous systems. Many of these skill sets are new or unknown, so not only does the T&E community need enhanced methodologies, but personnel are needed who will be capable of applying these novel methodologies on future autonomous systems. The advent and introduction of intelligent systems that collaborate with humans and other intelligent systems will drive the need for disciplines that are currently very scarce throughout the T&E community, to include anthropologists, sociologists, psychologists, etc. Resources will need to be allocated to allow for development of the current workforce, training, recruiting, and retaining a new base of experts in autonomous T&E, and development of curriculum will be needed to bridge the many technical disciplines which autonomous systems cover (e.g. Artificial Intelligence, Robotics, Machine Learning). Further, funding will be necessary for academic institutions, Federally Funded Research and Development Centers (FFRDCs), and other initiatives to expand the scholarly contributions.

The second specific challenge discussed is the identification and development of correct range requirements and instrumentation needed to perform valid test and evaluation of autonomous systems. In some cases these requirements and instrumentation may exist, but in many others, the community is early in the process of exploring the required T&E infrastructure, technologies, and capabilities for testing autonomy. This challenge ties in substantially with that of Test Adequacy (Section E), because T&E practitioners need to have systems and capabilities developed that ensure an autonomous system has been adequately tested. This becomes challenging because of certain aspects such as a system's ability to "learn" and how one is to measure or test "learning", both from the aspect of ability to learn as well as the current state of knowledge or wisdom resident within the system. Test adequacy is also difficult to ascertain due to the difficulty in quantifying test coverage when evaluating a non-deterministic, learning system with an almost infinite state space.

Because of a system's stochastic evolution, and the potential for unanticipated behavior, sequential progressive testing will likely be critical, and developing methods which allow for real-time test planning and conduct are needed. Subsequently, the ability to resource these types of dynamic methodologies is a unique challenge that autonomous systems will present. Tests will need to adapt in real-time as the autonomous systems under test adapt to their environment and or collaborating humans. Further, processes which allow for rapid implementation of changes are needed and, because of the likely volume of data, autonomous analysis will become imperative.

Another area of range requirements and instrumentation is the need for range world models. These models will likely need to be both physical and simulated, and should account for the unique aspects that autonomous systems will bring to the test community. The range models are required to provide ground truth, which will be compared to the world models generated by the autonomous systems under test. More M&S will likely be necessary for during the test cycle, which emphasizes the need to "shift left" as highlighted throughout this report. Though it is a challenge we currently face in T&E, development of physical test ranges may have cultural barriers that need to be overcome as different environments must be planned for. This challenge is only expanded when considering testing autonomous systems. Two examples where the attributes of autonomous systems will challenge current thinking, foundational principals and the culture of T&E are in the areas of repeatability and the desire to quantifiably bound the performance envelope while testing to the edges of those boundaries. First, an autonomous system equipped with machine learning will, by design, alter its decisions and ensuing behavior on a continual basis – thereby resulting in a system that never repeats its performance – even when immersed in the same environment and subjected to the same stimuli. Secondly, the non-deterministic and dynamic nature of autonomous systems makes it exceedingly difficult to predict and bound the software excursion paths as well as the possible decisions and behaviors exhibited by the system. This makes it impossible to define, let alone test all plausible operational conditions. It is likely then that performance evaluations and risk assessments will be reported in probabilistic terms versus definitive or absolute values.

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Test Infrastructure and Personnel identified by the working group consist of:

- Test Resource Management Center (TRMC) oversight and investments
- Defense Acquisition University (DAU) Training and NAVAIR University
- Advanced degrees, interdisciplinary statistical engineering, and academic partners
  - Naval Postgraduate School (NPS)
  - Air Force Institute of Technology (AFIT)
  - Arizona State University (ASU)
  - Virginia Polytechnic Institute
- MRTFBs, existing ranges, and the Range Commander's Council (RCC)
- STEM initiatives and internships, and career broadening opportunities
- T&E, Science of Test (SOT), and CTEIP investments, SBIRs
- FFRDCs
- Existing analytic methods

Gaps for Test Infrastructure and Personnel identified by the working group consist of:

- 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, and resources for workforce development
- Structured approach to keeping the workforce current
- Alignment with adjacent disciplines
- Infrastructure for requirements, instrumentation, and measures

The Test Resource Management Center (TRMC) "is focused on improving DT&E planning and execution, building the professional workforce, maintaining state of the art facilities, and providing data-driven support to the DoD Components" and is central to the existing test infrastructure and personnel capabilities. TRMC is congressionally mandated to provide Office of the Secretary of Defense (OSD) oversight of the national T&E infrastructure, which entails the mission to plan for and to assess the adequacy of the Major Range Test Facility Base (MRTFB) to provide adequate testing in support of the development, acquisition, fielding, and sustainment of defense systems. TRMC is additionally mandated to develop the Strategic Plan for DoD T&E Resources, to certify the adequacy of all DoD T&E budgets, to manage the OSD T&E technology and capability investment programs, to manage the OSD distributed test capability (Joint Mission Environment Test Capability) and to manage the National Cyber Range. Through their leadership, many resources exist from which to build upon the capabilities of the DoD testing community. On the academic side, numerous resources are in place to help train the community that will likely need to be inflated as autonomous systems become more commonplace. The Defense Acquisition University (DAU) currently has a robust workforce development capability consisting of online resources, online and in-residence professional development courses, and policies to ensure workforce currency such as biannual requirements for continual learning points (CLPs). The infrastructure DAU uses to educate the workforce presents a substantial benefit to long-term requirements as they are identified by the community at large.

Additionally, existing analytic methodologies such as the tool set known as scientific test and analysis techniques (STAT) provide a comprehensive basis from which new skills can be developed. The OSD STAT in T&E COE helps connect these two areas (education and research) by providing education to the community through a series of short courses, as well as providing research for novel T&E methods. Further, the capability of the many FFRDCs should be exploited in an integrated fashion with the various other technical resources available. Some of those technical resources are the various academic institutions present within the DoD who offer advanced technical degrees such as the Air Force Institute of Technology (AFIT) and the Naval Postgraduate School (NPS). Each of these institutions are likewise involved in the Science of Test (SoT) Consortium which partners with leading T&E civilian institutions such as Arizona State University (ASU), Virginia Polytechnic Institute and State University (Virgina Tech), amongst others in order to lead the effort for research and development of T&E methodologies. There are other resources such as the Naval Air System Command (NAVAIR) University College of Test and Evaluation which provide specialized training that should be integrated into the existing COI.

From an infrastructure perspective, there are numerous existing test ranges and test labs which make up the MRTFB that possess significant test assets., In addition resources such as the Range Commander's Council provide a venue and process to address common issues and establish standards across the test community. The facilities and assets which are currently in place are necessary for existing test needs, however they will require enhancements and associated investments to adequately perform T&E of autonomous systems. The operational footprint and duration of some autonomous systems will stress the physical limitations of the ranges. Investments will be required in infrastructure, instrumentation, spectrum management and data analytics. Additionally, the national cyber range will provide benefits relevant to Safety and CyberSecurity (Section G), and capitalizing off of their lessons learned will be critical to improving infrastructure as needs arise.

Given the many available resources, members of the workshop identified many resource, process, or methodology needs for the test infrastructure and personnel, most of which focus on workforce development. First, a sequenced career path should be developed which exposes people across research and T&E focus areas. Getting research, acquisition, system engineering and operational practitioners involved in various phases of the test process will improve long term coordination and cooperation across the community while also furnishing the T&E community with insight into the needs of the various consumers of test data. Further, a community of practice (COP) will provide a mechanism for sharing best practices, and lessons across the test community so that knowledge and resources can be shared and progress can be made across all acquisition programs. In order to provide this capability, a business model will need to be made that provides the resources to acquire and retain talented manpower willing to face these challenges. Another aspect to be considered is a structured approach to maintaining workforce currency and providing resources for workforce development once novel methods for T&E of autonomous systems are developed. Key to identifying the most appropriate skill mix and assessing the current state of the workforce relative to the skill mix is the conduct of a comprehensive Personnel Gap Analysis.

Identifying gaps in infrastructure becomes further challenging because of the significant unknowns in autonomous capabilities. One substantial gap in existing infrastructure is the ability to rapidly adjust resource allocations as tests are executed. The current structure of the systems engineering process, which requires a fairly static T&E plan early on, will likely not prove insufficient for the needs of the potential dynamically evolving systems. The ability to use resources for real-time test planning and conduct in an agile fashion presents a significant gap.

The other two gaps presented involve identification and development of proper requirements and measures (Section B), and instrumentation needed to adequately evaluate autonomous systems. Because each system will likely be vastly different and delivered as a complex 'black box', it is hard to identify what these need to be at this point, but allocating research into the potential for them will help bridge the gap in the near term, allowing for the DoD test community to be seen as leaders, instead of identifying these needs in a lagging fashion.

D. Design for Test

I. Challenge Description

One of the most common strategies today for improving our ability to test a system is test automation. The adoption of test first practices by the majority of agile teams demonstrates how test automation needs are addressed from the initial steps of system concepts. Design for Testability consists of the architectural and design decisions in order to enable us to easily and effectively test our system with decomposability of functions in mind, when practical.

The design for test challenge centers on the inherent opacity in many software systems and how they may relate to autonomous systems. Combine this with the dynamic nature of a learning system, the inability to bound that state space of a self-governing system, the challenge to predict or contain the behavior of a non-deterministic system, the fact that human cognition and decision making is essentially replaced with digital cognition and our inherent lack of insight into the internal states and processes of the autonomous functions and the T&E of autonomous systems begins to appear intractable at best. M&S will become increasingly important under the "shift left" mentality, and having a standard framework or architecture with a T&E oriented application program interface (API) is a central objective for this challenge. This challenge focuses on the need to be able to precisely stimulate the system under test and extract the internal states of key autonomous functions/components such that perception, reasoning, decision making and learning can be rigorously tested. Additional objectives for this challenge are to avoid constraining the architecture because of the potential unforeseen outcomes that a learning system may exude, while simultaneously avoiding the enactment of overly burdensome standards that stifle innovation and entrepreneurial motivations. The move towards the T&E API allows for a test interface to be integrated within the system in an optimal design vice bolted on intrusive

Features should be designed into systems so that the users and testers can understand the system's intent and the basis by which decisions are made by the system under known and emerging conditions. The key part of this challenge is that a fundamental paradigm of testing may change such that negative outcomes are tested against. That is, rather than testing to validate a performance metric or particular behavior, test to ensure an unacceptable behavior will not arise under a range of conditions. Some of the specific challenges identified by workshop members are that the test API should be

- Domain agnostic, with sufficient extensibility to accommodate domain specific attributes
- Abstracted to a sufficient level
- Able to expose any unanticipated semantics
- Able to support a business model that incentivizes broad adoption
- Able to precisely stimulate the system to simulate specific input conditions
- Able to extract sufficient data to provide insight or introspection of internal states during specified trigger events or time stamps
- Able to perform the stimulation and data extraction in an unobtrusive manner
- Able to perform the stimulation and data extraction during M&S events and during live test events

The system's world models and experiences will need to be:

- Discoverable prior to testing in order to baseline the system's level of intelligence,
- Extracted and measured during a test to provide a temporal tracking of intelligence,
- Uploaded to the system to reset the desired initial condition or state of the system to enable effective sensitivity and regression testing

Such manipulation, control and understanding of the system world models and experiential factors are necessary to ensure that test events can be injected with a certain level of uncertainty to be able to capture intent across a wide range of situations. Finally, metrics must be defined in order for designs to be developed that are testable. These potential MOEs or KPPs will be centered on behavior of the system, the ability of the system to learn and operate autonomously, the systems reasoning capabilities, the system's trustworthiness, and the system's perception. All of these metrics lead to their own unique challenges.

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Design for Test identified by the working group consist of:

- Simulation models and tools to stimulation a system
- Autonomous system predictive behavior technology
- Temporal logic/derived model
- Virtualized representation of complex information/cyber environment
- AVIA – Sim/Stim environment to stress and analyze autonomy
- Integrated reporting and diagnostic tools (e.g. orange wire B-2's)
- Statement of Work (SOW) defined requirement to use API, inject T&E, etc…
- Automated maneuvering systems

Gaps for Design for Test identified by the working group consist of:

- Standard framework or architecture with a T&E API
- Formal models that bridge the gap from requirements to design
- Understand the human involvement with the framework so we know how to incorporate autonomy
- The Test Design gaps directly impact requirements and should be tied to the requirements challenge

Many of the current capabilities for this challenge are the current systems with some level of autonomy and the experiences gained through their development. Many of our physical systems have automation currently embedded, along with integrated reporting/diagnostics tools. Improving these, capturing lessons learned, and presenting them to the community will become necessary. Other technology present is the robust M&S capabilities and the ability to simulate systems, stimulate them in various ways and characterize the outputs. The overall process for doing this with autonomous systems should utilize what we know, and build upon these methods.

Other potential resources include autonomous system predictive behavior technology, temporal logic definition, and enhanced LVC simulation capabilities. Our ability to simulate and test these various operational environments, including those in the cyber domain provide a starting point for future methodologies.

The gaps in existing capabilities are three-fold. First a standard framework or architecture with a test API should be developed. As discussed above, it should support a business model that can be incentivized for broad adoption and implementation across the DoD test community. Second, formal models need to be developed which bridge the gap from requirements to design, making this challenge interconnected with the requirements and measure discussion in Section B. Finally it will be important to understand and characterize the human involvement within the framework so we can quantify the incorporation of autonomy. This need also intersects the challenge of testing the human-system teaming (Section H)

A commonly accepted autonomy architecture that supports development of autonomous defense systems currently does not exist. Architectural factors will substantially affect the testability, costs, cross-platform reusability, and benefits of autonomous systems. The degree to which systems are modular will allow for the reuse and reconfiguration of hardware and software from different platforms and affect their testability. Inclusion of understanding of the human involvement within the autonomy framework within these architectures will be critical.

Additionally, the risk of deploying a large number of similar systems may create vulnerabilities across the force. Open architecture and interoperability standards, while useful in promoting modularity and reuse, may make this challenge more severe if they propagate vulnerabilities throughout the force. Alternatively, vulnerabilities often arise from the seams or interfaces that exist across components of a system or across platforms in a system-of-systems. The lack of common or standard architectures will negatively impact compatibility, interoperability, system-of-systems performance and correspondingly, expose exploitable vulnerabilities. This condition will become more acute as the community evolves to deployment of heterogeneous teams of autonomous systems.

E. Test Adequacy and Integration

I. Challenge Description

Test adequacy was the most broadly recognized challenge of the workshop, with the overarching challenge being the inability to collect an adequate amount of information to definitively quantify performance and risk for decision makers. Further, testing the integration of the system's physical and autonomous components presents a unique challenge not previously faced by the test community. Because a system is anticipated to learn and change over time, test adequacy is dependent upon all test events, across the entire systems engineering process. Additionally, test adequacy has the potential to change within a test event depending on the system's evolution.

In many cases, a test may need to be changed in real time, so the test community needs to consider

- How is the acquisition process affected, and how do we adapt it?
- What new scientific approaches for test design need to be developed to ensure adequacy
- What safety and security considerations (Section G) need to be adequately tested?
- What safety considerations need to be adequately addressed in design of the test environment?
- How will agile or elastic test planning and test conduct be performed to accommodate dynamic test conditions?

Test adequacy also relates to the requirements and metrics challenge, because many new metrics may need to be developed before being able to discuss the demonstration of their adequacy. This challenge is further complicated because unknown sources of variation are being introduced into the test environment. Currently, we account for such sources as environment, mission, operator, etc., but for autonomous systems variation may come from experiences and a systems individual learning, and the human-machine interaction in which different operators provide unique familiarity. A few other considerations identified by the workshop are:

- Negative learning presents a new challenge as a potential reliability metric, and may never really be testable until fielded because we hope the system doesn't devolve

- Most recently executed test may set the initial conditions for the next test unless learning can be “erased” or reset
- At what point is a system ready for fielding if the test units are continuously evolving and learning
- While we currently cover large and complex mission spaces, tests for autonomous systems must include a time domain to account for their ability to learn over time and experience
- Mission centric definitions such as high-level objectives versus performance centric definitions can drive larger tests

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Test Adequacy and Integration identified by the working group consist of:

- STAT, Experimental Design, Reliability growth, and other quantitative methods
    - Test coverage algorithms
    - Automated stepwise regression testing
- Emerging Developmental Evaluation Framework (DEF)
- C4ISR test environment and tools
- Test configuration capture tools
- Kinematic Analytics & Displays
- Sequential test design (not real time)
- Human Factors Design requirements, guides, standards
- Systems Engineering Processes/Principles
- Reliability Growth
- Tools/Methods for test coverage

Gaps for Test Adequacy and Integration identified by the working group consist of:

- How much testing is enough? What is MVP? Necessary and Sufficient
- Sequential test design in near real time
- Coverage of space dimension given temporal dynamics
- Process to identify integration points across the life cycle
- “Composeability” of test results with other types of evidence
- New “statistical engineering” techniques; combine empirical with non-empirical
- How to assess swarm or team dynamics, to include distributed perception, shared knowledge and distributed decision making.

Much of the existing systems engineering processes, principles, and framework for acquisition programs provide a beneficial starting point for T&E of autonomous systems. The T&E Master Plans (TEMPs) and required test plans, coupled with the emerging developmental evaluation framework (DEF) are used to detail the plans and activities for acquisition of the system. The TEMP requirement is a primary strategy for management of test resources ensuring the most effective information is obtained and provides a foundation from which new methods will likely be researched.

Other tools currently being used are the solution methodologies and model selection algorithms, such as automated step-wise regression, kinematic analytics and displays, and various human factor engineering design requirements, guides and standards. Further, the C4ISR test environment and tools already test certain factors such as desensitization and network or communications performance, and the test bed exists to provide quantifiable assessments of current and future technologies. The existing science behind human factors engineering, test, and evaluation will be of critical import when adding a system designed to make decisions that complement the warfighters capabilities.

Current DOE methods are and will continue to be valid and valuable. However, while necessary they may not be sufficient when evaluating autonomous systems. The Defense Science Board TASK FORCE REPORT: The Role of Autonomy in DoD Systems (2012) recommends that "The Under Secretary of Defense for Acquisition, Technology and Logistics (USD(AT&L)) should create developmental and operational test and evaluation (T&E) techniques that focus on the unique challenges of autonomy (to include developing operational training techniques that explicitly build trust in autonomous systems)." It goes on to state that "The Military Services should structure autonomous systems acquisition programs to separate the autonomy software from the vehicle (physical) platform." This necessarily would require application of DOE and other scientific test and analysis techniques (STAT) for the characterization of the inputs and responses of the physical system. Understanding of the performance of the physical system may be even more critical in autonomous systems than in our current systems, especially during integration with the autonomy software. The autonomous software must understand the capabilities and limitations of the physical system such that the system's decisions do not result in behavior that exceeds structural or aerodynamic limits.

Additionally, the role of software verification and validation will take on a more prominent role. For instance, software testing is the process of exercising software to verify that it satisfies specified requirements and to detect errors. Software testing is often conducted using combinatorial designs or risk based designs, the latter of which are advocated by DOT&E in a 2010 memo, Guideline for Operational Test and Evaluation of Information and Business Systems (Scientific Test & Analysis Techniques (STAT) for Software Testing Best Practice, http://www.afit.edu/images/pics/file/STAT%20for%20Software%20Testing.pdf). Static analysis is complementary to testing and involves examining the software instead of executing it. Static analysis includes techniques and approaches ranging from manual design and code reviews to fully automated source code scanners. Static analysis requires access to the software, which may be difficult in proprietary cases or remote or embedded systems. However, in theory static analysis may consider the effects of all possible inputs subject to imprecision from model abstractions and assumptions. For instance, testing could not find a "back door" in code through which the user gains full rights if the user enters a specific string, such as "JoshuaCaleb." However, testing may exercise an entire system end to end. Some tools combine testing and static analysis, gaining the best of both approaches. (Report on the Metrics and Standards for Software Testing (MaSST) Workshop 2012, http://samate.nist.gov/docs/MaSST_2012_NIST_IR_7920.pdf) An expansion of static analysis, formal methods involve the specification of requirements and the design of a system in a formal specification language or ontology that is semantically complete and allows for rigorous analysis. With the requirements and software expressed in the formal language, analysis approaches can broadly be separated into:

Axiomatic approaches: involve reducing the analysis to a mathematical theorem proof and provide mathematically rigorous statements on the ability of the program to achieve the specifications.

Semantic approaches: include model-checking, which utilize exhaustive search through all possible program executions while looking for behavior inconsistent with formally stated requirements. (Ewen Denney, and Ganesh Pai, "Evidence Arguments for Using Formal Methods in Software Certification," IEEE International Workshop on Software Certification (WoSoCer 2013), November 2013).

Existing test methodologies include sequential test design, test coverage metrics, and reliability growth planning, tracking, and projection models. Sequential test design is discussed in many sections of this report as being a critical existing technology, though its primary shortcoming is the need for a user in the loop to identify new design points and update the test. Existing tools for determining test coverage, such as statement coverage, may be useful in identifying the adequacy of a test, but coming up with test cases that achieve 100% statement coverage will be potentially challenging for non-deterministic systems with a nearly infinite potential state space. Finally, the concept of reliability growth provides a necessary tool for management of resources to help encourage system improvement over time. These ideas may be applied to autonomous systems in ways that are not yet apparent, such as the development of a “mean time to negative learning” metric as part of the reliability testing. Additionally the concept of reliability growth and associated analytical methods may be at least partially applicable (or analogous) to an emerging concept of trust growth.

Many of the capabilities currently used will provide a basis for improvement of autonomous systems, where the complexity is increased because of the potentially infinite design, decision, and outcome spaces. Not only will the practitioners need to sufficiently test the design space, they will also have to develop test cases that sufficiently test the decision engine, a concept introduced in Section B. However the T&E, acquisition and operational communities may be forced to accept that with autonomous systems we may never be able to test to 100% coverage of a performance envelope or a boundary of plausible behaviors, leading to more probabilistic test results and risk assessments vice definitive results that provide a quantifiable assessment of risk.

For certain human-system teams where experiential learning occurs, a need may exist to compose test results with any unique information such as user idiosyncrasies when determining test adequacy. If an exclusive human-system team has developed a certain level of trust, this type of information should be testable and maintained throughout the acquisition cycle. How do we assess a system’s ability to adapt to its human teammate and learn over time, much in the same way that Google algorithms can predict our search needs based on our behavior over the internet? How do we account of the diversity, deviation, competency, biases and non-deterministic nature of the human component of the human-machine team when establishing standard test cases or in our related M&S endeavors? This will require new statistical engineering techniques that combine empirical and non-empirical analyses.

Near real time, and adaptable or automated sequential test design was identified as a gap in many of the challenges discussed at the workshop. Because methods for testing "learning" for an autonomous system must be developed, we also face the challenge of increased complexity due to dynamically changing spaces (design, decision, and outcome) in the time domain.

Finally, a process which identifies points at which the autonomous and physical components of a system should be integrated to ensure their adequate testing is needed. This is additionally compounded by the need for integration with a human user during the developmental and operational testing phases.

F. Testing Continuum

I. Challenge Description

Many of the workshop attendees discussed the challenge that the current testing continuum framework under which physical systems are acquired will be insufficient as it is applied to T&E of autonomous systems. This challenge overlaps with many of the other identified challenges because it covers the current systems engineering process including DT and OT, and presents a real need for agile acquisition practices. The primary objectives of this challenge are to determine when to test, what to test, how to retest under the assumption that learning has occurred, and how the information from all tests should be used to inform decisions. Ultimately the desire is to establish practices (standards, guidelines, methods) that result in a continuum of T&E and V&V throughout the system lifecycle. That is from research, through development and throughout deployment. This continuum of testing will require testing start early, a more sequential progressive approach is taken, and implementation of a comprehensive M&S strategy across the life cycle. There is no "test phase" with a beginning and an end; it extends throughout the life cycle of the system.

One of the primary challenges that are presented under this category is the current inflexibility of test resources, funding, personnel, and the scheduling of test ranges. Much of the test planning is performed early in the acquisition cycle, and resources are allocated well in advance. Because the nature of autonomous systems is to experience emergent behaviors, funding should be flexible enough so that it may be applied to further testing or used in future testing without negative repercussion. Test personnel need to be embedded in the process from day one, should have a wide range of test expertise, and should be empowered to provide support during development.

Another challenge categorized under testing continuum is the need for a flexible concept of a requirement, due to the assumed stochastic nature of autonomous systems. For a physical system, a requirement may be the ability to accomplish a task with a certain level of confidence; for a software system, a requirement may entail a certain level of functionality. For an autonomous system, requirements may take on a unique form, such as the ability to reason in certain settings, or simply the negative requirement that it does not perform a certain set of tasks or exhibit particular behaviors given some controlled settings. Another unique challenge that requirements may face for T&E of autonomous systems is whether it is only the system being tested, or if it may be a human-system pair or a team that is being tested. If the assumption is that autonomous systems may acquire some of the idiosyncratic properties of its owner, then requirements testing will need to account for this.

Specific challenges identified at the workshop:

- Resource conflicts (ranges, personnel, funding)
- Need flexible concept of 'required' performance
- Next test depends on result of previous
- Not aligned with current requirements and processes
- Need to embed testers from day 1, with test support at development
- Test to evaluate utility, not rigid requirements

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Testing Continuum identified by the working group consist of:

- Sequential Test Design (Formal and Informal)
- Statistical Process Control (Characterized/Automated)
- Utility/Loss function analysis
- Constrained Optimization
- Bayesian Inference – combining info
- Real time analysis
- Block Upgrades/
- Existing Test Process "just" needs augmented: DT/OT/ Integrated DT-OT
- Rapid Capabilities Office
- Need to establish practices for a continuum of T&E across the life cycle

Gaps for Testing Continuum identified by the working group consist of:

- Organizations and practical ability to coordinate and build on learning
    - 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

Though it may not be sufficient in the future, the current systems engineering process is still a necessary tool for acquisition programs. To address the needs of autonomous systems, the process may simply need a detailed augmentation, though that is still to be determined.

The test strategy may require a more rigorous evaluation process than used today. Execution of DASD(DT&E)'s 'shift left' initiative is a good first step toward achieving more rigor, specifically, expansion and use of a developmental evaluation framework that specifies required technical, program, and acquisition decisions that support the program in understanding the current state of program development to include specific developmental risks. Concepts such as sequential test design and real time analysis, coupled with Bayesian methods provide ways to develop robust test plans which exploit data from across the acquisition phases.

The current Research & Engineering Autonomy Community of Interest (COI) Test and Evaluation, Verification and Validation (TEVV) Working Group Technology Investment Strategy 2015-2018, signed by ASD(R&E), states the need for

> "Cumulative Evidence through RDT&E, DT, & OT - Progressive sequential modeling, simulation, test and evaluation M&S and T&E at each Technical Readiness Level (TRL) and product milestone currently provide an invaluable resource not only to verify and validate that a system satisfies the user requirements, but also to aid in technology development and maturation. However, the development of effective methods to record, aggregate, and reuse T&E results remains an elusive and technically challenging problem. As an example, DoD Directive 3000.09 implies that autonomous weapons software, where possible, not be re-written but incrementally developed and verified by sequential, progressive regression testing. It is paramount that products, methods, tools, and capabilities developed in Goal 1 (Methods & Tools Assisting in Requirements Development and Analysis) and Goal 2 (Evidence-Based Design and Implementation) support the transition of autonomous systems to the DT and OT communities, to better define and, where reasonable, focus and increase the effectiveness of test and evaluation plans. Methods must be developed to record, aggregate, leverage, and reuse M&S and T&E results throughout the system's engineering lifecycle; from requirements to model-based designs, to live virtual construction experimentation, to open-range testing. This goal endeavors to research the development of standardized data formats and Measures of Performance (MOPs) to encapsulate experimental results performed in early research and development, ultimately reducing the factor space in final operational tests."

The paragraph goes on to state that "Additionally, statistics-based design of experiments methods currently lack the mathematical constructs capable of designing affordable test matrices for non-deterministic autonomous software. Software systems require a risk-mitigation methodology offering the same spirit as Design of Experiments (DOE) while not relying entirely on statistical approaches." It is important to note that this document tends to focus on the autonomous component of the system.

The use of scientific test and analysis techniques (STAT), including DOE methods, are and will continue to be valid and valuable. Any current shortcomings in using STAT by the acquisition community will be highlighted further when conducting test and evaluation of autonomous systems. In addition to the effective use of current STAT methods, these scientific test and analysis techniques must expand to include techniques that focus on the unique challenges of autonomy to include building trust in autonomous systems.

Another methodology highlighted by the working group is that of block upgrades, where an initial system is fielded with the anticipation that further capabilities will be systematically introduced that will include further testing. Statistical process control may also be applied in order to monitor the test process or the processes an autonomous system will be performing to ensure such factors as negative learning are identified and controlled.

Such agencies as the Test Resource Management Center (TRMC) may be used to assess budgets, manpower, and infrastructure for the T&E of autonomous systems, while investing in required technology and tools. Through studies, TRMC shares expertise, and lessons learned about certain facets of efficiently fielding weapon systems that can additionally be applied to autonomous systems. TRMC is currently conducting a study on T&E of Autonomy, which examines the projected state of autonomy at various time horizons, and then identifies the required test infrastructure and skillsets required to accommodate testing of that autonomy. The TRMC may be central to ensuring the right mix of T&E skill sets are recruited and retained, as well as the development of any infrastructure necessary to accomplish T&E for autonomous systems.

The gaps identified for this challenge focus on the need for a data management system and repository that will allow sharing of ideas, test designs, lessons learned, data, models, and data analytics so that organizations have the ability to coordinate tests and build on a broad learning base. Developing a centralized community of practice (or something similar) where test personnel develop unified goals and are encouraged to share data and outcomes will foster an environment that accelerates advancement of autonomous system T&E.

Additional needs center on the lack of agile infrastructure which will allow for rapidly changing test requirements and methodologies necessary to provide a compositional analysis of data through this unidentified incremental testing cycle. The processes and procedures by which TEMPs are currently created, updated, and resourced will need to be developed with unique characteristics of autonomous systems testing in mind. Additionally, the rigid definition of requirements may need to be addressed, along with how requirements are tested.

G. Safety/CyberSecurity

I. Challenge Description

When discussing the development of autonomous systems, safety is of critical importance. As with physical systems, the central objective is to quantify risk, thus testing of safety for autonomous systems becomes increasingly difficult. Autonomous systems will need to be demonstrably safe and effective. One specific question that will need to be answered is how one can ensure the system will never choose to perform an unsafe action, and further, how these tests can be constructed and results shared so that they are trusted by third parties. Because of the potential emergent behaviors “negative testing”, or testing that ensures some action is not taken, will likely be difficult so determining when a system has performed an undesirable behavior is complex. A factor which adds to the difficulty of testing safety is the need to not strictly impact or constrain the autonomous decision making capabilities of the system. This then calls for nonintrusive methods and instrumentation that will not corrupt or influence performance of the system and the ensuing test results. The autonomous systems emergent behavior coupled with its complexity leads to its classification as a complex adaptive system. Complex adaptive systems are characterized by system level behaviors that cannot be predicted given an understanding of the subsystems. These software intensive complex adaptive systems are difficult to test due to:

- Software components requiring combinatorial testing techniques;
- A system-of-systems possibly having ad-hoc participants and network connections or participants;
- A large if not infinite set of system states;
- Unknown, previously undefined environments

Another aspect of this challenge is that of CyberSecurity and the vulnerabilities which may be present with the physical components of the system, as well as potential reasoning capabilities or the algorithms which drive reasoning. Cyber and CyberSecurity testing, are more prevalent as subfields for autonomous systems for which more research is a focus. Characterizing the system’s attack surface, understanding the cyber kill chain, developing a coherent test planning, and testing and retesting autonomous systems will be critical. Individual stand-alone systems will be prone to attacks from local, network access, and adjacent network sources.

As stating above, when deploying large system-of-systems homogeneous or heterogeneous autonomous systems, the risk of deploying a large number of similar systems may create vulnerabilities across the force. Open architecture and interoperability standards, while useful in promoting modularity and reuse, may make this challenge more severe if they propagate vulnerabilities throughout the force.

The fact that autonomous systems ‘think’ differently than humans will open up new potential challenges to the resiliency of military forces, as adversaries may be able to take advantage of the shortcomings of machine perception and cognition. This is not to say that autonomous systems will necessarily be more vulnerable than human-controlled systems, but they will be vulnerable in different ways. If there is widespread adoption of a large number of autonomous systems with similar or identical perception and cognition systems, this raises the potential for one or a small number of weaknesses to endanger a large proportion of the force. This is typically a less serious problem with manned systems because each human operator is idiosyncratic. Quantification of decision making and learning is needed.

Test Challenges:

- Inability to test all cases – infinite factor space further complicated by a potentially infinite decision space
- System may be changing continuously as knowledge is gained and decisions are made
- How to tell "good" perception/reasoning from bad
- Awkward time scales: Cyber may be too fast, but long endurance tests will likely be too slow
- Potentially exploitable algorithms (decision processes) by adversaries
- Manage, mitigate, and define risks
- Real-time prediction
- "Negative learning" may not surface during any feasible test or acquisition cycles

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Safety/CyberSecurity identified by the working group consist of:

- Accelerated life testing (ALT)
- "Safety Case" methods + "safety bag"
- Run-time monitoring
- Real-time test termination/abort and geofencing
- Formal methods (state reachability)
- Risk based designed experiments (severity and rate of occurrence)
- Test coverage metrics and methodology
- Defined safety processes and procedures
- Adversarial testing (e.g. genetic algorithm)
- Safety clearance processes (flight clearance, weapon safety board, etc…)
- Combinatorial test designs
- System safety analysis (FMECA, Fault tree analysis, etc…)
- Formal Language requirements
- Institutional review boards (IRBs) for human research

Gaps for Safety/CyberSecurity identified by the working group consist of:

- New/updated IRB policies/procedures/methods
- 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

H. Testing of Human-System Teaming

I. Challenge Description

- Characterize and measure performance of human-machine partnership
- Measure shared situational awareness
- Testing human-machine tradespace
- Testing heterogeneous autonomous systems
- Measuring difference between human intent and machine actions
- Creating a normalized model of the human

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Testing of Human-System Teaming identified by the working group consist of:

- Human effectiveness measures
- Human factor design requirements
- Designed experiments that account for human factors
- Decision support system assessment processes

Gaps for Testing of Human-System Teaming identified by the working group consist of:

- Trust metrics & assessment
- Methodology for evaluating varying levels of autonomy

I. Post-Acceptance Testing

I. Challenge Description

The final challenge identified at the workshop was that of test and evaluation, use, and updating of systems which have been deemed acceptable and deployed. The objective of this challenge focuses on developing the methods that provide the ability to continuously monitor the status and performance of deployed systems in order to perform predictive analysis and to incorporate operational lessons/data into test methodologies. This will enable the capability to predict and mitigate anomalous performance/behavior while increasing the operational fidelity of the test environment and injecting current experiential knowledge into the systems under test. This capability will be enabled by the use of cloud based technologies and relatively ubiquitous reach back connectivity. Additionally, because there may be idiosyncratic behavior, testing a system’s ability to adapt and learn may be unique, and should occur after fielding.

Some specific challenges identified include the potential need for a periodic assessment of compliance with existing, newly introduced, or learned capabilities, and the inclusion of some sort of feedback and assessment of the system's ability to handle updates. This may call for some level of self monitoring to ensure negative responses don't occur, or development of a metric based on the last acceptable certified level for the system. Similarly, it will be important to identify and detect any trigger(s) which mean more testing (or system reversion) needs to occur.

Challenges:

- Recurring, periodic assessment of compliance with existing, newly introduced, or learned capabilities, rules, or constraints
- Assess value, robustness of learning (guard against negative learning or brittleness)
- Will require some level of self monitoring
  - Prevent negative responses from developing
  - Last acceptable certified level reversion capability
  - What is trigger that requires more testing?
- Assess autonomous system adaptation to an aging physical system
- Feedback & assessment of updates

II. Current T&E Processes, Methods, and Gaps

Current processes and methods for Post-Acceptance Testing identified by the working group consist of:

- Human effectiveness measures
- Short term: maintenance and inspections
- Long term: Schedule or event driven overhauls
- Assurance and acceptance sampling (reliability field)
- Statistical Process Control (SPC) to detect "out of control" systems
- Block/spiral updates

Gaps for Post-Acceptance Testing identified by the working group consist of:

- Structured training events after fielding for each new environment and/or system
- Health maintenance system for field based learning
  - Periodic or trigger driven
  - Licensure (potentially need progressive licensure)
- Real time data capture and analysis in order to use and feedback to broader database for system sharing
- Infrastructure and methods for test community to tap experiential data stemming from deployed systems

5. Conclusions and Recommendations

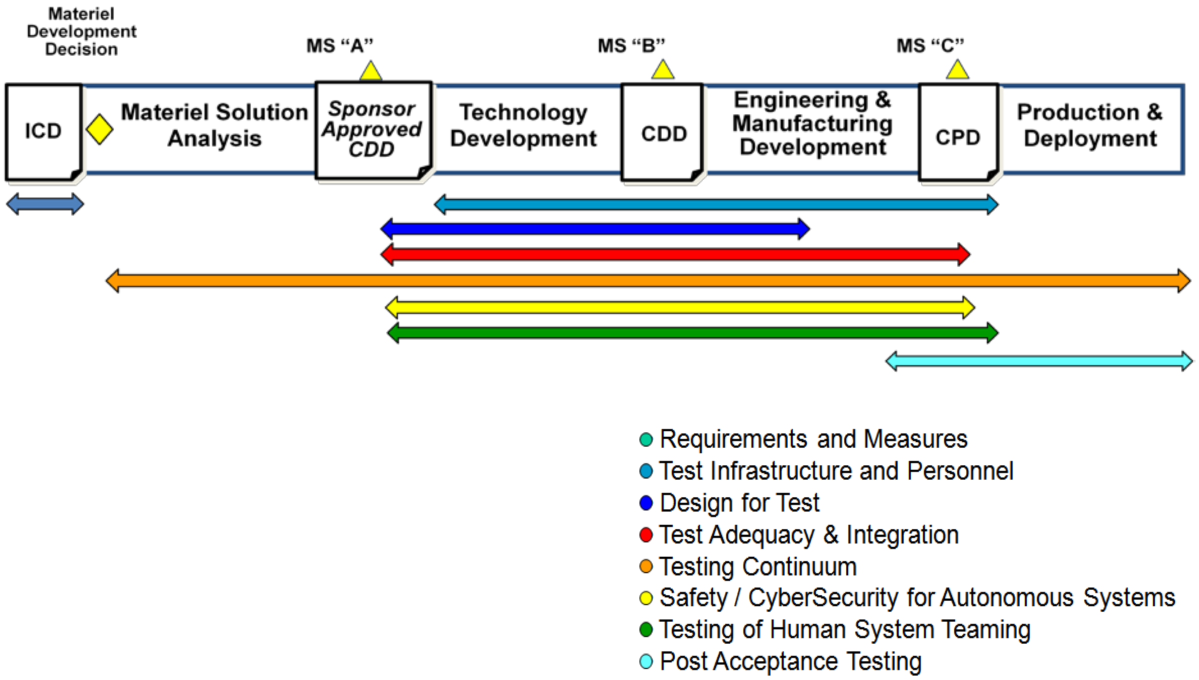
Autonomous defense systems have gained great interest in recent years. Autonomy enables a particular action of a system to be automatic or, within programmed boundaries, self-governing, while not necessarily replacing the human operator, but may extend and complement their capabilities to conduct complex, large-scale, long duration operations without sacrificing safety or effectiveness, and must allow Warfighters to focus on their primary mission, not on operating their tools. Additionally, autonomous defense systems most likely will need to operate in unstructured, dynamic environments. These systems therefore will have challenges in testing as they take the complex systems that are currently difficult to test and transforms them into complex adaptive systems that may learn and have emergent behavior.

The OSD STAT COE Workshop on Test & Evaluation of Autonomous Systems focused on identifying the challenges of testing and evaluating autonomous systems, identifying existing test & evaluation requirements / processes / methods to address each challenge, and finally, identifying gaps in the requirements / processes / methods which will need to be addressed for adequate test & evaluation of autonomous systems.

The workshop identified and developed eight challenges for test & 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 System Teaming
- Post Acceptance Testing

No one challenge was identified as being more important than the other as they span the acquisition life on any autonomous system acquisition program as shown in the figure below.


Materiel Development Decision
MS "A"
MS "B"
MS "C"
ICD
Materiel Solution Analysis
Sponsor Approved CDD
Technology Development
CDD
Engineering & Manufacturing Development
CPD
Production & Deployment
● 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
● Post Acceptance Testing

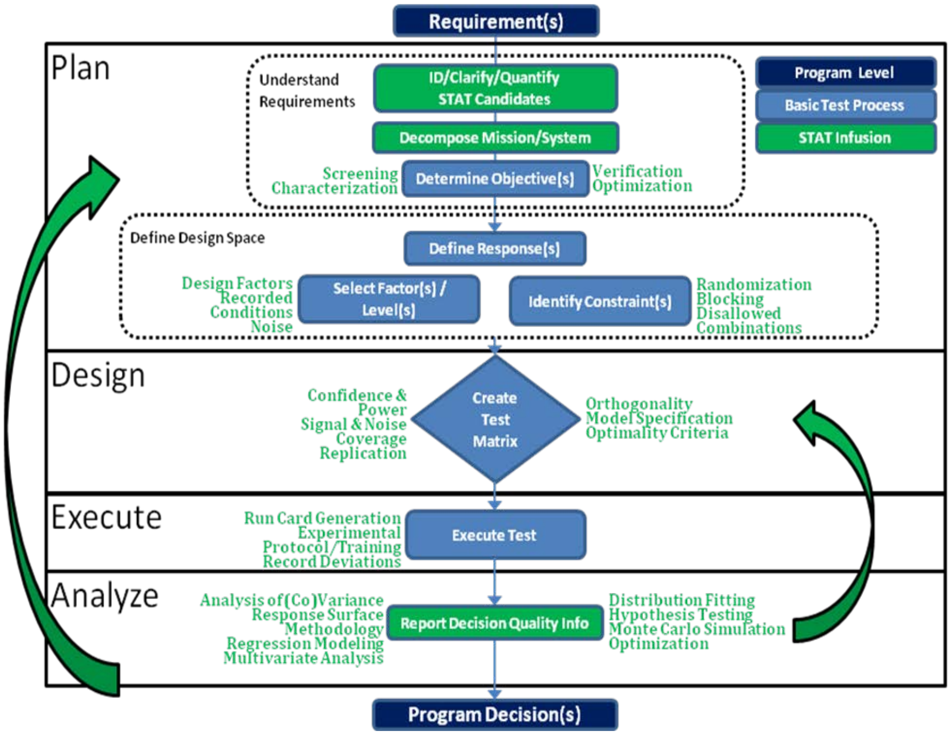
The workgroup developed a consensus on the eight challenge areas. Shortfalls in executing current acquisition programs for complex systems were identified as even more critical for future autonomous systems. Additionally, new challenges were identified that will need to be addressed through research funding for testing, changes in policy, and workforce development efforts in order to adequately perform testing to inform key technical, programmatic and acquisition decisions throughout the developmental lifecycle of the program. Among efforts requiring test research funding are:

- Improved statistical engineering methods are needed to support both developmental and operational testing of autonomous systems to address systems interacting with a dynamic environment in a non-deterministic manner. Improvement of these methods is a component of a larger need; these adaptive autonomous systems will require more stringent adherence to systems engineering principles throughout development.
- Processes and methods need to be developed to address Inputs – Process – Outputs of autonomous systems and human-machine element interaction and roles within the process.
- A test and evaluation continuum paradigm must be developed and adopted that requires testing start early and a more sequential progressive approach is taken that includes development and implementation of a comprehensive M&S strategy across the life cycle. There is no "test phase" with a beginning or end; it extends throughout the life cycle of the system.
- Measures must be developed to address state space adequacy, trust, and human-machine interaction.
- Design of experiment methods must be developed for defining test cases and expected results that overcome the difficulty of enumerating all conditions and non-deterministic responses that autonomy will generate in response to complex environments.

- Models and live virtual constructive (LVC) test beds are needed that support robust testing while minimizing risk and cost.
- Development of techniques that capture learning growth, possibly similar to reliability growth models is needed.

The primary challenge in testing autonomous systems is the broad scale and complexity of the systems, missions, and conditions. This is best addressed by breaking down the requirement, system, and/or mission into smaller pieces, which can then be readily translated into rigorously quantifiable statistical designs. One large comprehensive experiment is unlikely to inform technical, programmatic, and acquisition decisions adequately. A progressive sequential approach to testing is a better strategy that the workshop participants embraced.

To support such a breakdown, the STAT methodology can be used since it is an iterative procedure that begins with the requirements and proceeds through the generation of test objectives, designs, and analysis plans, all of which may be directly traced back to the requirement. Critical questions at every stage help the planner keep the process on track. At the end, the design and analysis plan is reviewed to ensure that it supports the objective that began the process. The following figure provides a concise process flow diagram that summarizes the application of STAT to the test and evaluation process and one in which the challenges identified clearly apply.



While the workshop participants rigorously addressed testing of autonomous systems and research from a current perspective to advance the state of the art, unknown unknowns will surface as more advanced autonomous systems are developed. Testing lessons learned, best practices, and identification of testing challenges from early autonomous systems development will be critical in creating efficient and effective testing to understand and reduce the significant additional risks of autonomous systems.

6. Workshop Contributors

Dr Darryl Ahner (Director, OSD STAT COE and workshop co-lead)
Mr Matthew Clark (Air Force Research Laboratory and workshop co-lead)
Mr Michael Badillo (Naval Surface Warfare Center – Corona)
Dr Luis Cortes (OSD STAT COE)
Dr Don Davis (Georgia Tech Research Institute)
Mr Christopher Eaton (412 Test Engineering Group)
Maj Jason Freels, PhD (Air Force Institute of Technology)
Dr Laura Freeman (Institute for Defense Analysis)
Mr Bob Grabowski (MITRE)
Mr Michael Harman (OSD STAT COE)
Mr Robert Heilman (VQuest supporting Test Resource Management Center)
Dr Raymond Hill (Air Force Institute of Technology)
Ms Kristen Kearns (Air Force Research Laboratory)
Ms Jennifer Lopez (Air Force Research Laboratory)
Dr Francisco Ortiz (OSD STAT COE)
Dr Carl Parson (OSD STAT COE)
Mr Rohintan Patel (Test Resource Management Center)
Dr John Raquet (Air Force Institute of Technology)
Dr David Sparrow (Institute for Defense Analysis)
Dr David Tate (Institute for Defense Analysis)
Dr Leonard Truett (OSD STAT COE)
Maj Brian Stone (OSD STAT COE)
Mr Daniel Sullivan, ctr (Defense Threat Reduction Agency)
Mr Reagan Woolf (412 Test Engineering Group)