Developmental T&E of Autonomous Systems Consolidated Challenges and Guidance

April 2025

Name: Charlie Middleton
Position/Title: STAT COE ctr support
Department/Organization: DTE&A

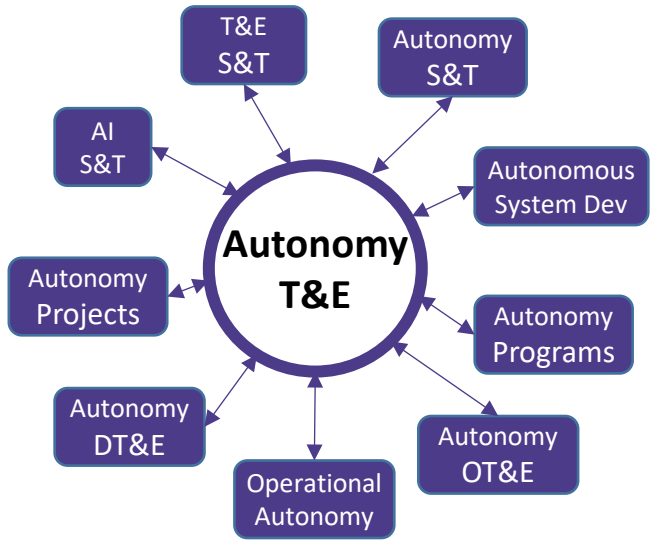
Overview

- Autonomous Systems T&E:
  - ➢Why are autonomous systems any different from traditional ones?
  - ➢What new or heightened challenges do these pose?
  - ➢What can we do about it?

- Overview of content in the pending DTE&A publication, “DT&E of Autonomous Systems Guidebook

Why Advance T&E of Autonomous Systems?

- We stand at the start of a new era in human culture and modern warfare. Machines may ‘soon’ do nearly all physical tasks that humans historically have
- AI-Enabled Autonomy is a revolutionary technology
  - Cheaper unit cost
  - Larger quantity
  - Less human capital/training
  - Faster execution
  - Safer to humans
  - Survivability
  - Endurance
- Challenges focused on determining its suitability, lethality, sustainability, adaptability, and ethical use

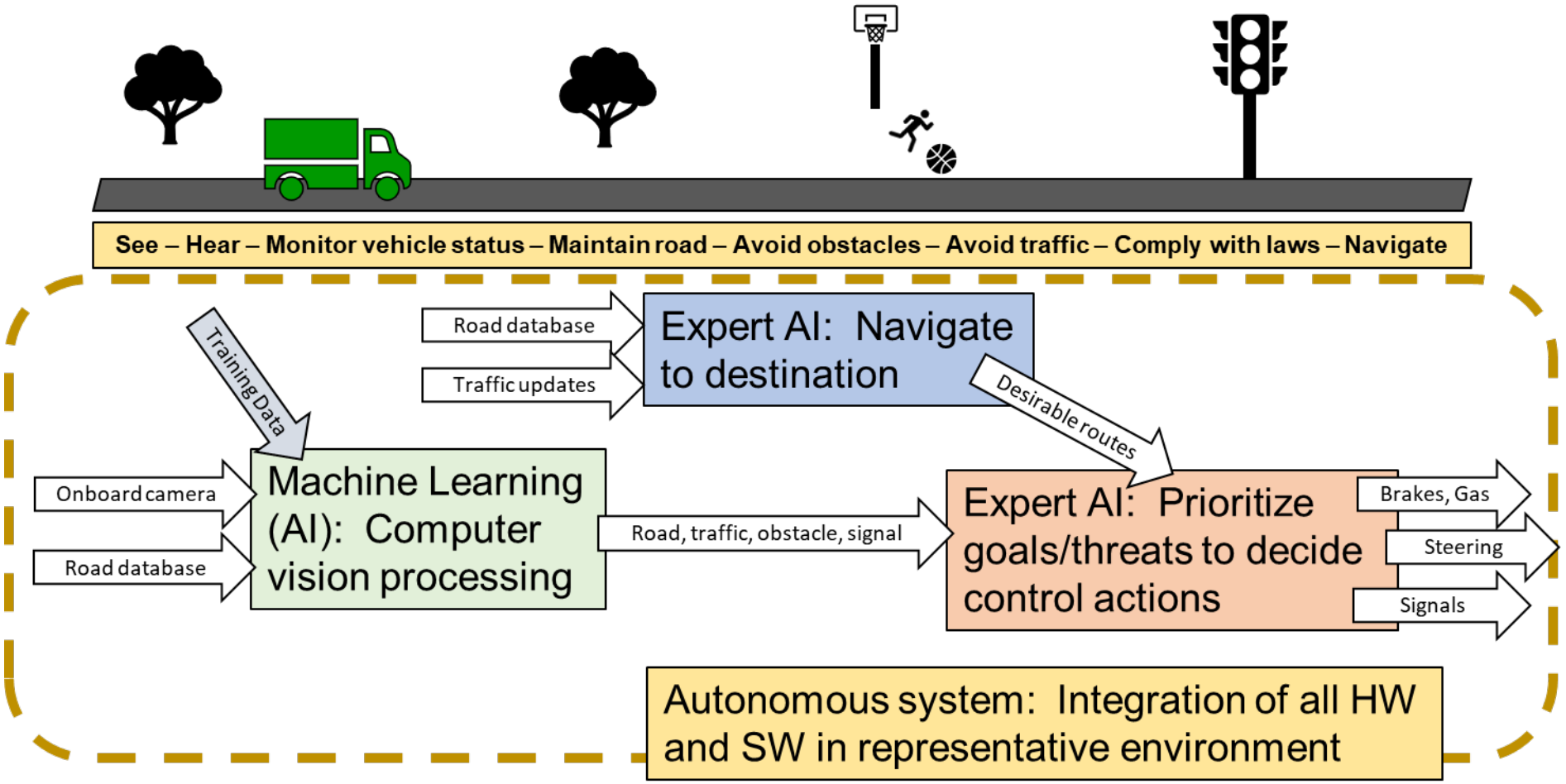


Test and Evaluation (T&E) is a critical linchpin between S&T research, tech maturation, system program offices, and warfighters.

Goal: Advance scientific practices for conducting T&E of Autonomous Military Systems that lead to mission assurance as demonstrated by effective, reliable, and ethical employment of modern, innovative autonomous capabilities

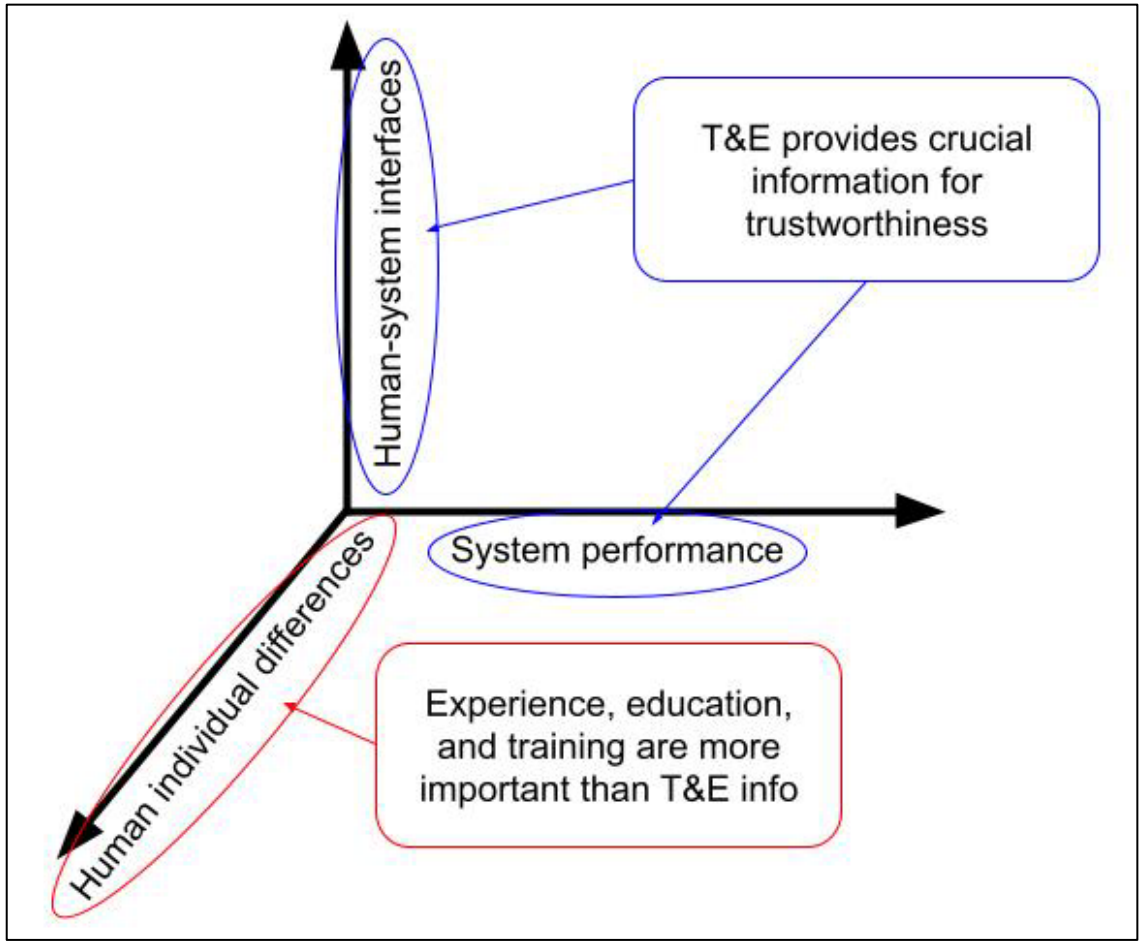
Autonomy Example with AI and ML

- NOTIONAL Autonomous Automobile System



Perspectives on Trust

- How can we trust the autonomous system?
  - **Trust** – a personal, human feature that will vary based on many factors
  - **Trustworthiness** – a system feature that can be more objectively measured
    - effective performance
    - safety, security, reliability, availability, maintainability …
    - **ability to be understood**
    - **ability to self-report problems it can't handle**
    - **ability to be controlled by humans when these features are violated**
- **T&E plays a crucial role** in establishing autonomous system **trustworthiness**
  - Key issue: Ability to be understood
  - Must have effective **human-system interfaces**



DT&E of Autonomous Systems

Challenges

- Overarching Challenges
  - Adapting to Developmental T&E as a Continuum
  - T&E of the OODA Loop

- Specific Challenges (13)
  - Requirements
  - Infrastructure
  - Personnel
  - Vulnerabilities
  - Safety
  - Ethics
  - Data
  - Human Autonomy Teaming
  - Black Box Components
  - Mission Evolution
  - Dynamic Learning
  - Test Adequacy and Coverage
  - Integration and Interoperability

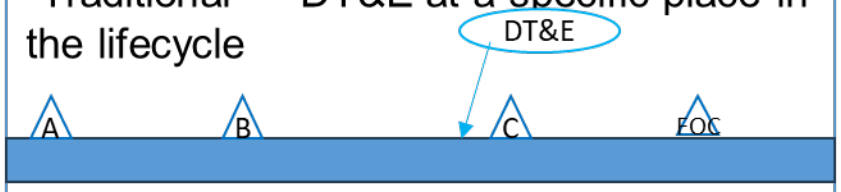
Overarching Challenge: Adapting to Developmental T&E as a Continuum

Developmental T&E across the continuum …

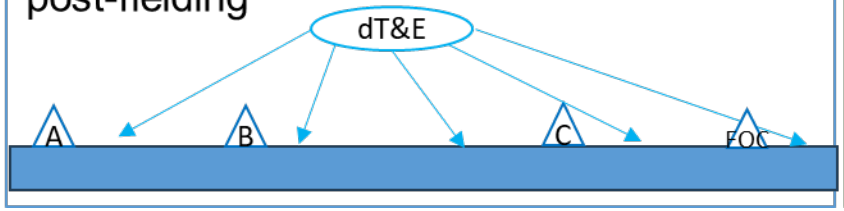
… the continuum of a ***system lifecycle.***

when the T&E occurs

"Traditional" – DT&E at a specific place in the lifecycle



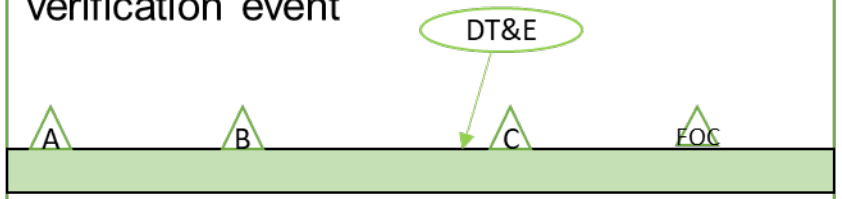
"dTEaaC" – extends 'left' and 'right' for both early S&T/R&D insights and iterative, evolving capability evaluation post-fielding



… the continuum of ***system change frequencies.***

how often T&E occurs

"Traditional" – DT&E as a one-time verification event

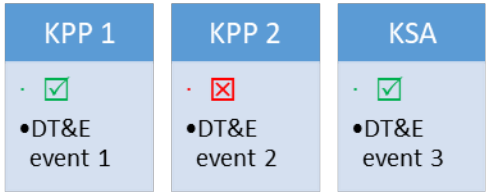


"dTEaaC" – iterative T&E supporting S/W Acq Path and Min Viable Product concept, iterative developments, & continuous testing

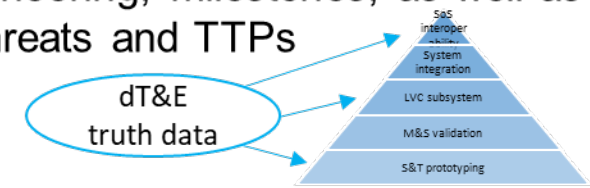
… the continuum of ***decision information needs.***

what needs the T&E feeds

"Traditional" – milestone driven, DT&E as a V&V of requirements/specifications

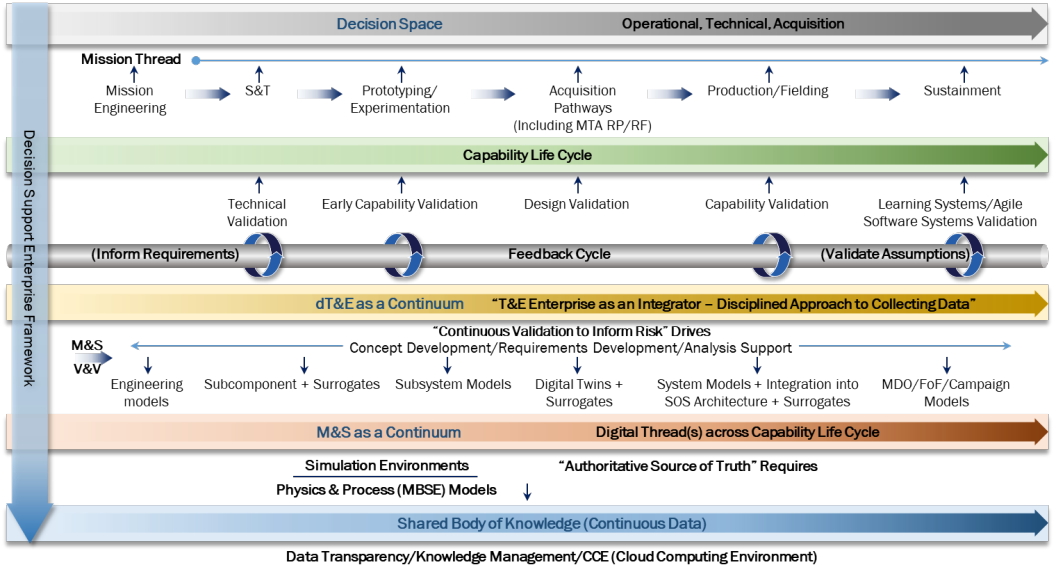


"dTEaaC" – truth data driven, feeding early S&T, prototyping, modeling & sim, digital engineering, milestones, as well as evolving threats and TTPs



dTEaaC Approach

- dTEaaC guidance addresses how data evaluation can occur across all analyses and studies, LVC and M&S, and test activities while being rooted in a common learning construct. By aligning all data-driven activities across the lifecycle, engineering and acquisition professionals can gather data more efficiently and evaluate data more holistically.



- dTEaaC methods are built on three core tenets:

1. **<u>Deliberately Executing a Campaign of Learning:</u>** Integrating our knowledge needs and learning opportunities across the entire engineering lifecycle

2. **<u>Data-Driven Decision Making:</u>** Embracing the assistance of decision support systems (models, AI, data dashboards, etc) to inform critical acquisition or make-buy decisions

3. **<u>Leveraging Digital Ecosystems:</u>** Exploiting readily available digital ecosystems instead of one-off or standalone tools, data repositories, and workflows

Overarching Challenge: T&E of the OODA Loop

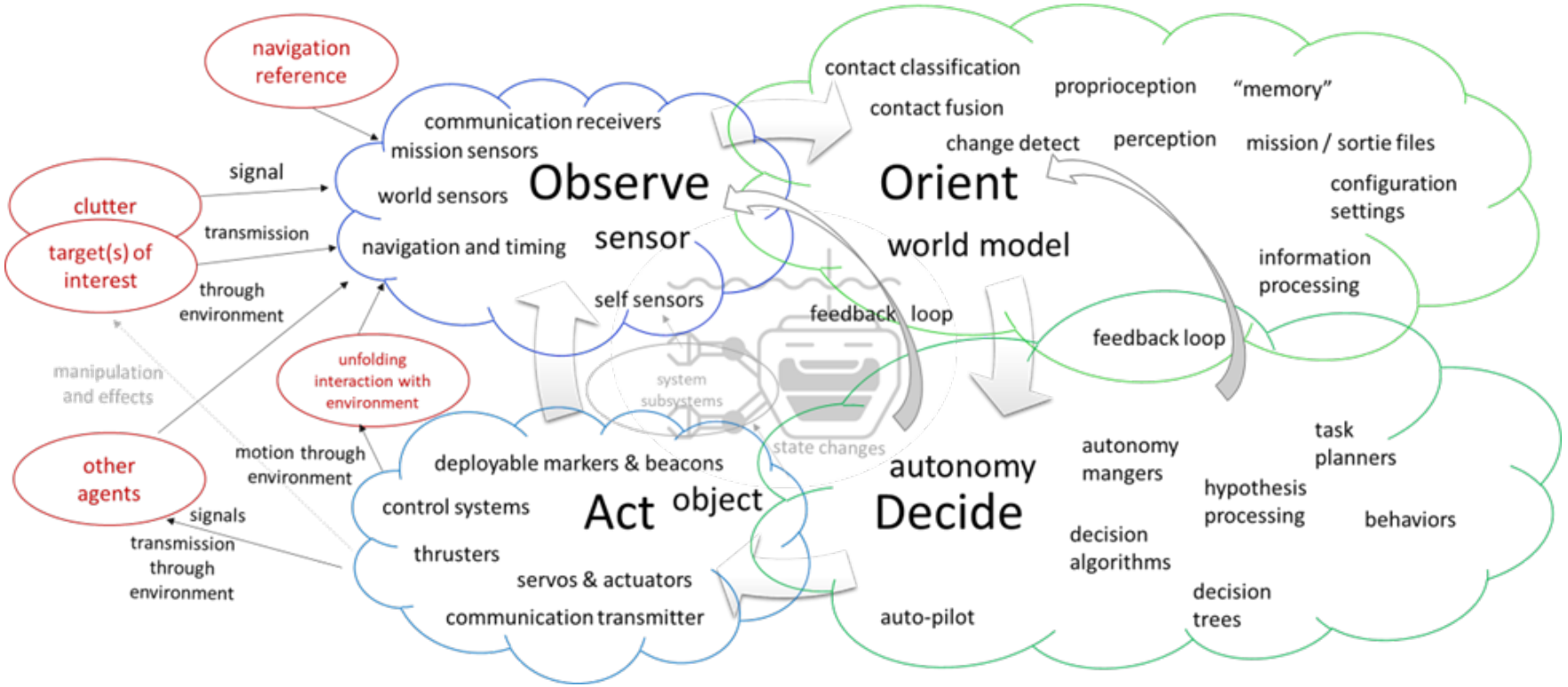
- Legacy T&E

Test point condition
e.g. Target/threat, range, speed, altitude/depth, control, EW

Single test point

Response
e.g. Miss distance, track time/accuracy, maneuver

- T&E of the OODA Loop



Mapping of Challenges vs Methods

Methods	Challenges														
	T&E as a Continuum	T&E of the OODA LOOP	Requirements	Infrastructure	Personnel	Exploitable Vulnerabilities	Safety	Ethics	Data	Human-Autonomy Teaming	Black Box components	Mission Evolution	Dynamic Learning	Test Adequacy and Coverage	Autonomy Integration and Interoperability
End-to-end Autonomy T&E Processes	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
STAT for Autonomous Systems	x		x				x	x	x	x				x	x
M&S for Autonomy T&E		x												x	
Operational Modeling		x	x				x			x					x
Small Scale Development	x		x				x		x						
Open System Architecture	x	x	x	x	x		x		x	x	x				x
Autonomy Requirements and Specifications			x				x			x					x
Continuous Testing	x	x			x	x			x		x		x	x	
Code Isolation	x					x	x								
Assurance Cases	x	x			x	x	x	x	x	x	x	x	x	x	x
LVC Testing	x	x	x	x	x	x	x	x		x	x	x	x	x	x
Experimentation T&E	x								x	x				x	x
Surrogate Platforms	x	x			x		x			x					
Formal Verification Methods			x			x	x							x	
AI Model Testing	x	x			x	x					x		x		
STPA for Autonomy			x	x		x	x	x		x				x	x
Adversarial Testing		x				x					x	x	x		
Post-acceptance Testing	x	x			x	x		x			x	x	x	x	
Cognitive Instrumentation	x	x			x	x		x	x	x	x	x	x		
Run Time Assurance	x			x	x	x	x	x		x	x	x	x	x	x
Test User Interface		x							x	x					x
Human-Autonomy Team Performance	x	x	x	x	x		x	x		x			x		x
Automatic Domain Randomization	x	x					x		x		x			x	
Automated Outlier Search / Boundary Testing							x				x			x	
Failure Path Testing							x							x	
Human Performance Standards	x		x		x		x		x	x		x			x
Task-based Certification	x		x		x		x	x		x					
Operational and Mission-Based Testing						x	x			x				x	x
Quantified Risk / Performance Growth Curves	x					x	x			x				x	x

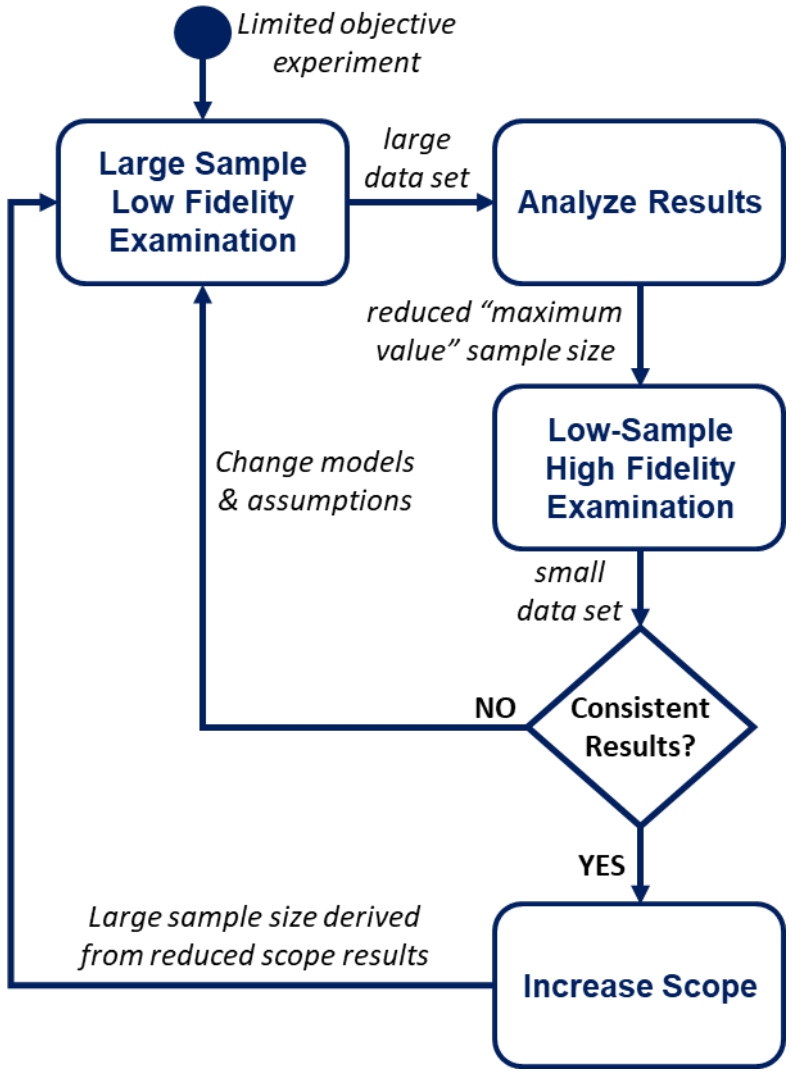
DT&E of Autonomous Systems

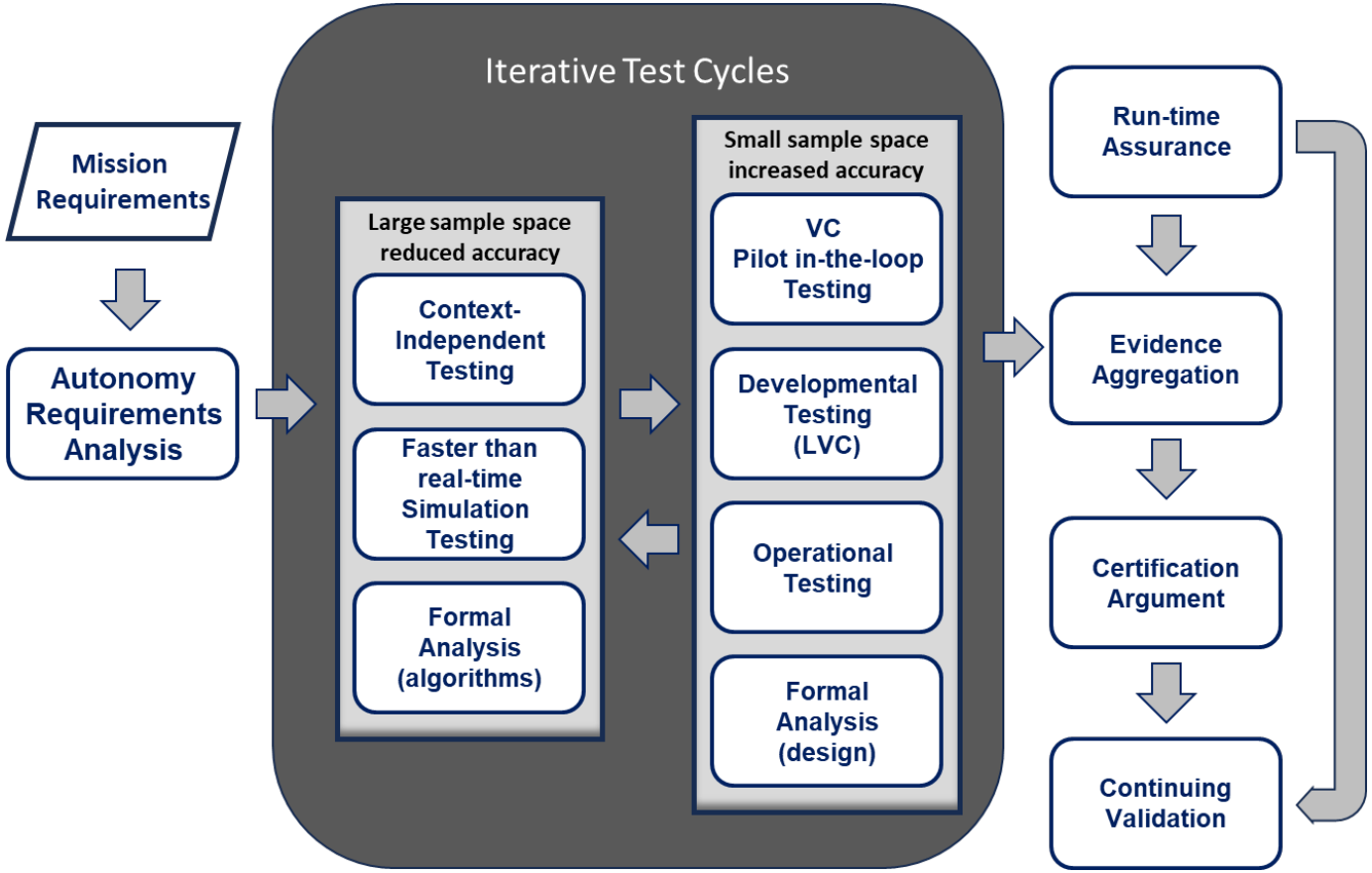
Challenges

- Overarching Methods
  - End-to-End Autonomy T&E Processes
  - STAT for Autonomous Systems
  - M&S for Autonomy T&E

- Specific Methods (26)
  - Broken up by phases
    - Acq and Dev Strategy
    - Test Strategy
    - Test Planning
    - Test Execution
    - Data Analysis and Evaluation

Overarching Method: End-to-End Autonomy T&E Processes





Overarching Method: STAT for Autonomy

- Scientific Test and Analysis Techniques (STAT): A structured, hypothesis-driven approach to T&E using scientific methods
  - Ensures that testing is objective, data-driven, and focused on drawing meaningful conclusions about system capabilities and limitations
- Key Benefits for Autonomous Systems
  - Accurate characterization of autonomous system performance and risks across the operational space - ability to measure and assess the breadth and depth of testing
  - Continuous improvement from data-driven test results
  - Understanding of how conditions and scenarios change system performance and trustworthiness
  - Reduced costs and development time from early identification of potential issues
  - Improved communication among stakeholders
  - Transparency and accountability for linking test data and evaluation to mission requirements and capability needs
- Use of STAT for T&E of DoD systems is required by DoD regulations
- Early STAT activities
  - Comprehensive requirements and mission analysis
  - Identifying measurable autonomy behaviors and test objectives
- STAT test planning activities
  - Generating comprehensive input and output data needs
  - Scientifically accounting for test constraints and limitations
- STAT test design
  - Understanding the combinatorics of conditions and state spaces
- STAT analysis
  - Statistical comparison of models vs real data, goodness of fit, and insights into where unknowns and risks still reside
- Best Practices:
  - Addressing complexity of autonomous system conditions and scenarios
  - Optimizing test efficiency with time and resources available
  - Sequential testing to maximize value of prior results and minimize risk
  - Bayesian statistics to scientifically account for prior knowledge
  - Collaboration to create a test team with diverse expertise
  - Early integration, flexibility and adaptability to evolving mission needs
  - Documentation and communication of plans, results, risks and unknowns

Overarching Method: Modeling and Simulation for Autonomy T&E

- M&S provides a powerful framework for analyzing, testing, and refining autonomous systems in controlled environments
- Key Benefits for Autonomous Systems
  - Early identification of performance issues, design flaws, and operational risks through controlled, repeatable test scenarios
  - Validation of autonomous system capabilities in a wide range of operational conditions
  - Facilitation of multi-domain testing by integrating land, air, sea, and space environments within a single simulation framework
  - Ability to model and test complex interactions, including human-autonomy teaming and multi-agent operations
  - Cost savings by minimizing real-world hardware testing
  - Mitigation of safety and/or security risks by simulating tests without endangering or compromising high-value assets
  - Acceleration of development timelines by allowing rapid iteration and evaluation of new system designs
- M&S is most effective as a complement, not a replacement, of live testing

- Types of Models
  - Cognitive models, System models, Environmental models
- Types of Simulations
  - Cognitive simulations, Faster than real-time simulations (FTRT), Real-time simulations, Virtual simulation, Constructive Simulation
- Uses of M&S
  - Requirements analysis, Formal analysis, Simulation-based testing, Bench testing, Human-in-the-loop testing, LVC testing
- M&S Impacts include
  - Interoperability, Human-system integration, Test optimization, Integrated test and training
- Best practices:
  - Clearly defining M&S goals and objectives
  - Independent, govt-owned simulation capabilities
  - Standardized & accredited M&S test environs, scenarios, and methodologies to provide credible results
  - High-fidelity to replicate complex operational scenarios
  - Modeling or integrating real-world uncertainty and variability
  - Use of actual system software within the simulation
  - Rigorous V&V of M&S with real-world data results, and regular updates to ensure relevance with evolving needs

Methods and Best Practices
Acquisition and Development Strategy

Method for Acquisition and Development Strategy	Short description
Operational Modeling	Provides a conceptual framework for understanding the roles, tasks, and behaviors an autonomous system will need to perform. Details use cases and procedures, helping ensure the intended functionality, enabling T&E to objectively evaluate performance of those tasks and behaviors.
Small scale development	Uses inexpensive, simple platforms and assets to accelerate testing and iteration, enabling rapid prototyping and scaling for larger systems. This approach supports efficient development cycles while minimizing initial costs and resources.
Open system architecture	Employs a modular design that uses modular system interfaces between major systems, major system components, and modular systems. Supports accurate T&E insight into internal system processes and portable T&E instrumentation and measures able to be re-used.
Autonomy requirements and specifications	Employs a deep understanding of autonomous behaviors and operational needs, ensuring that requirements are not based on assumptions or how humans perform tasks, but rather on what the autonomous system needs to achieve. Provides clear, concise, and testable requirements that accurately reflect the desired capabilities of the autonomous system.
Continuous Testing	Integrates testing throughout the development and operational lifecycle of autonomy technologies. Ensures that software and hardware components are reliable, safe, and capable of adapting to changing conditions and requirements. Helps track the growth and evolution of the autonomous capabilities over time, and is critical to establishing a pipeline of continuous integration and continuous delivery of capability.
Code isolation	Use of a software development framework separating safety-critical code, or mission-critical code, from non-critical code to ensure operational safety and prevent unintended consequences. Saves costs and time by allowing smaller scoped T&E for non-critical software changes.
Assurance cases	A structured argument, supported by evidence, that provides a compelling and valid case that a system is safe, secure, and fit for its intended purpose, and which is adaptable to the specific needs and context of the system being evaluated.

Methods and Best Practices
Test Strategy

Method for Test Strategy	Short description
LVC testing	Live, Virtual, and Constructive (LVC) testing is emerging as a proven method for T&E of autonomous systems. LVC offers a powerful and flexible approach to assess these complex systems across a range of operational environments and scenarios, optimizing resources and minimizing risks. Incorporates a mix of real-world, simulated, and emulated components.
Experimentation T&E	An iterative process of designing, executing, and analyzing experiments in operationally relevant conditions to assess early capabilities and limitations of autonomous systems. Exposes the system to a range of scenarios, including edge and corner cases, to understand its performance and to uncover unexpected behaviors and vulnerabilities. Unlike traditional T&E, it doesn't aim to confirm specifications but to reveal unknowns and inform future development.
Surrogate platforms	The use of substitute systems, simulations, or environments in place of the actual autonomous system or its intended operational environment, allowing testers to assess and refine autonomous system capabilities before testing with new assets,
Formal verification methods	A mathematically rigorous technique used to prove or disprove the correctness of a system's design with respect to a certain formal specification or property. This method is particularly relevant for autonomous systems where safety and reliability are paramount.
Adversarial testing	Uses simulated adversary forces and AI to identify system vulnerabilities and assess potential impacts. This method helps ensure that autonomous systems are resilient against threats and adaptable to hostile environments.
Post acceptance testing	Testing in operationally relevant environments after the system has been accepted for fielding. The complexity of their use cases often makes complete testing impossible before deployment, and many high consequence failures occur at very low frequencies. This necessitates a shift from the traditional DoD T&E paradigm of separate developmental and operational testing to a continuous evaluation process throughout the system lifecycle.

Methods and Best Practices
Test Planning

Method for Test Planning	Short description
AI Model Testing and Metrics	A systematic approach to evaluating the performance of AI models used in autonomous systems, involving the design of specific test cases, collecting performance data, and applying relevant metrics to assess the model's effectiveness and identify areas for improvement. It emphasizes continuous testing and monitoring throughout the system's lifecycle to account for the inherent uncertainties and complexities associated with AI.
System Theoretic Process Analysis (STPA) for Autonomy	A hazard analysis method grounded in Systems Theory. STPA is a modern accident causation model that asserts safety is a dynamic control problem. Uses a top-down approach for analysis and delivers qualitative results that can be used to guide the design of today's complex sociotechnical systems, including autonomous systems and their test safety critical issues, risks and processes.
Human-Autonomy Team Performance Methods and Measures	Methods focus on evaluating key factors between humans and autonomy, such as situational awareness, role clarity, communication, and collaboration to ensure effective and reliable team dynamics.
Automatic Domain Randomization	Leverages algorithms to automatically create variations in tests or simulations, including environmental conditions, sensor parameters, and the physical characteristics of objects and surroundings. Incorporates both environment and agent parameter randomization, providing a comprehensive approach to testing significantly exceeding the practical limits of manual scenario generation.
Automated Outlier Search and Boundary Testing	A process to identify where model behavior is at or near the limits of its operating conditions or exhibits changing performance. These conditions can be related to environmental factors, sensor performance, decision-making capabilities, or other constraints that the system is designed to handle. These regions are also important because they can identify critical areas for real-world testing for validation.
Failure path testing	Identifies and analyzes the potential paths where a system may fail under specific conditions. Autonomous systems' high dependence on complex software creates a need for testing of the many potential ways that software faults, bugs, or poor designs could cause unexpected system failures or deficiencies.

Methods and Best Practices
Test Execution

Method for Test Execution	Short description
Cognitive Instrumentation	A method to gain insight into the internal state and decision-making processes of autonomous systems. Imagine being able to see inside the "mind" of an autonomous system. Monitoring and analyzing data related to perception, reasoning, and planning, helping testers understand why a system behaves in a particular way. This "internal workings" refers to the machine's ability to perceive, reason, decide, and team in its dynamic OODA loop.
Run Time Assurance	A continuous process of monitoring an autonomous system's performance, detecting anomalies, and initiating appropriate responses to maintain safe and effective operation. It acts as a deterministic "wrapper" around the autonomy under test, with the authority to intervene and guide the system to a fail-safe condition if necessary. This allows for the safe exploration of complex autonomous behaviors without the risk of catastrophic failures.
Test User Interface	Provides testers with tools to interact with, manipulate, and evaluate the autonomous system in a safe and repeatable environment. Allows for the injection of various scenarios, environmental conditions, and system disturbances to assess the system's robustness and ability to handle unexpected situations. Enables the collection of valuable data on system performance, human-machine interactions, and even operator workload for tasks involving human partners

Methods and Best Practices
Data Analysis and Evaluation

Method for Data Analysis and Evaluation	Short description
Human operator performance standards	Applies specific measures of performance, suitability, and effectiveness based on established training and proficiency standards. These standards ensure that autonomous systems can achieve mission effectiveness by meeting or exceeding human performance baselines, enabling reliable interaction with human operators.
Task-based Certification	A capability-focused assessment method that shifts the focus from traditional pass/fail verification of individual requirements to evaluating the system's ability to perform mission-essential tasks in its intended operational environment, acknowledging the complex and adaptive nature of autonomous systems. An iterative certification process supporting the certification of autonomous systems for limited operations with specific tasks, with the expectation of expanded capabilities over time.
Operational and Mission-Based Testing	Focuses on evaluating autonomous systems in realistic mission settings alongside manned and unmanned assets. Assesses the system’s effectiveness, adaptability, and resilience under collective full-spectrum threats, to ensure that autonomous system capabilities meet the demands of dynamic, real-world mission conditions.
Quantified Risks and Autonomy Performance Growth Curves	Quantifies various types of risks for an autonomous system by using statistical techniques similar to reliability growth curves. By measuring relevant metrics over time, testers can use statistics to measure improvement and to make justified predictions of future capabilities.

DT&E of Autonomous Systems Guidebook

Summary and Future

Chapters:

1. Introduction
2. Policy
3. Background and Vision
4. Challenges
5. Methods and Best Practices
6. Conclusion
7. Glossary
8. Acronyms
9. References
10. Acknowledgements

Planned Expansion:

- Resources for T&E of Autonomous Systems
  - DoD Test Labs and Test Ranges, other test facilities
  - Hardware resources – testbeds, surrogates, platforms, etc.
  - Software resources – assurance cases, requirements analysis, runtime monitoring, etc.
  - Simulation environments
- Examples and Case Studies
  - Land autonomy
  - Sea autonomy
  - Air autonomy
  - Space autonomy
  - Swarm autonomy
- Intent to update this guidance on a relatively frequent basis
  - Always looking for future collaborators
  - Seeking more methods and best practices!

Questions?

Visit www.cto.mil/DTEA

Visit www.AFIT.edu/STAT

Email AFIT.ENS.STATCOE@us.af.mil

STAT COE | Delivering Insight to Inform Better Decisions