Performance Measures, Environments, Actuators, Sensors (PEAS) for Testing Autonomous Intelligent Agents

Dr. Laura Freeman, Dr. Jitesh Panchal
Adam Dachowicz, Melanie Grande, Andrew Lang (Purdue)
Dr. Ivan Hernandez, Scott Welch (Virginia Tech)

12th Annual SERC Sponsor Research Review
HOSTED VIRTUALLY ON: November 18, 2020
www.sercuarc.org

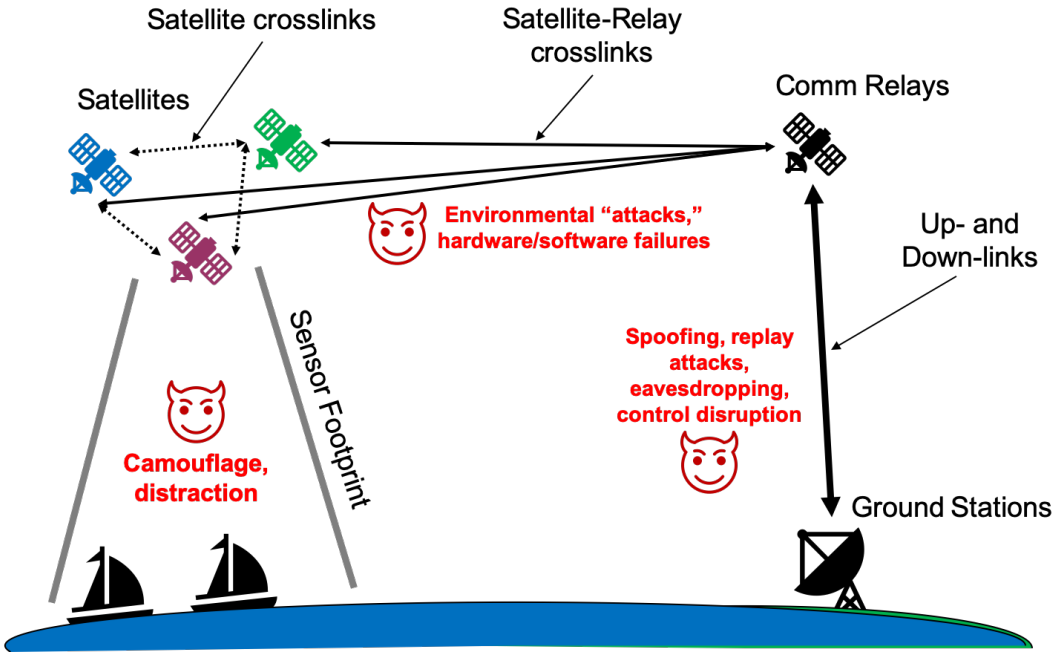
VIRGINIA TECH

Outline

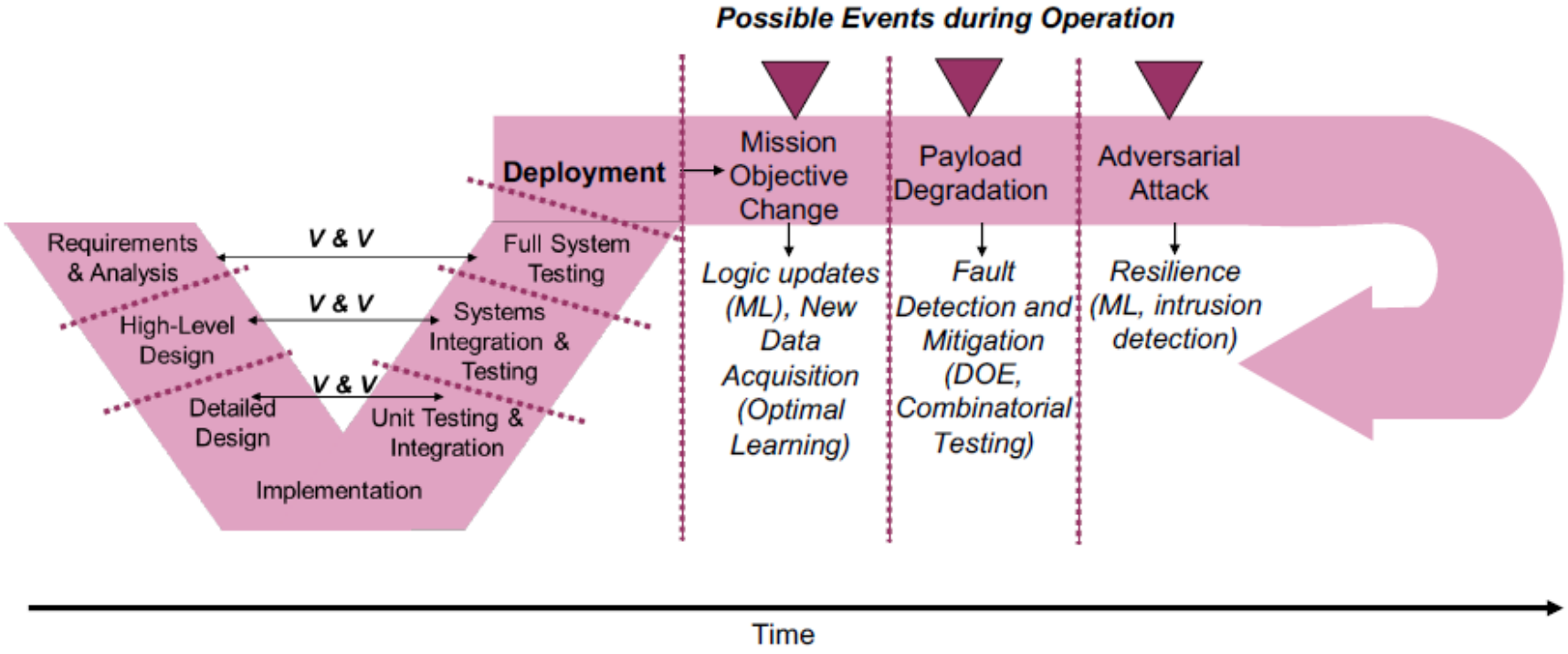
- Phase 1: Test and Evaluation Framework
- Phase 2: Performance Measures Case Studies
- Summary & Future Work

Test and Evaluation Framework

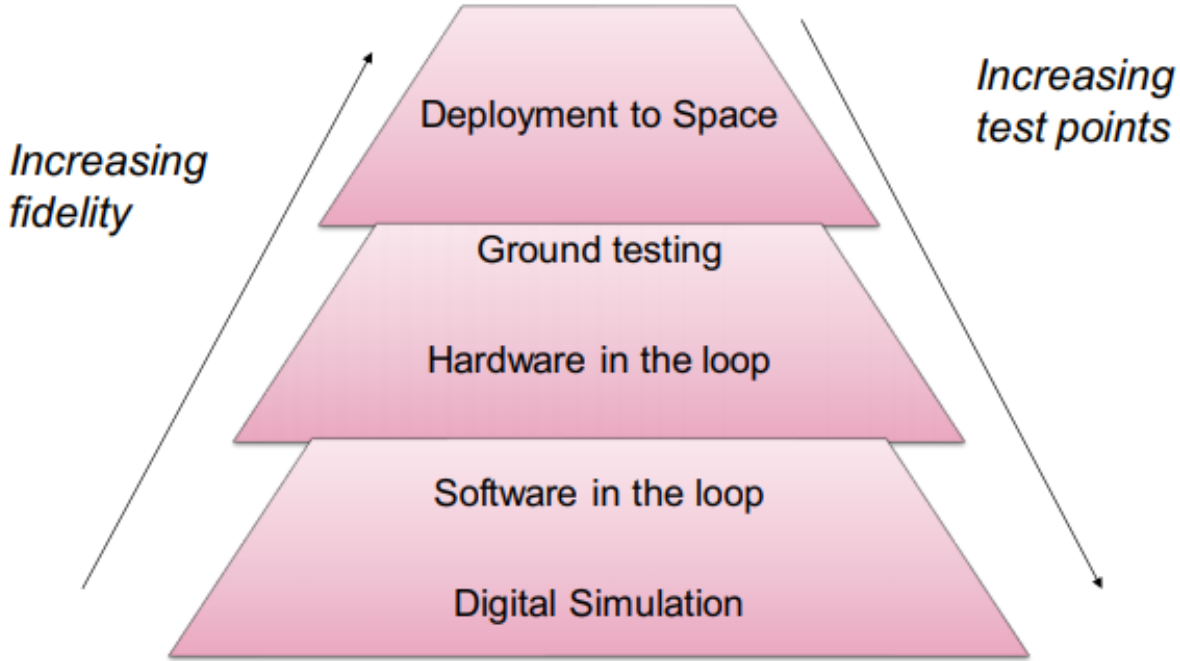
- Multi-Agent Systems of Autonomous Intelligent Agents



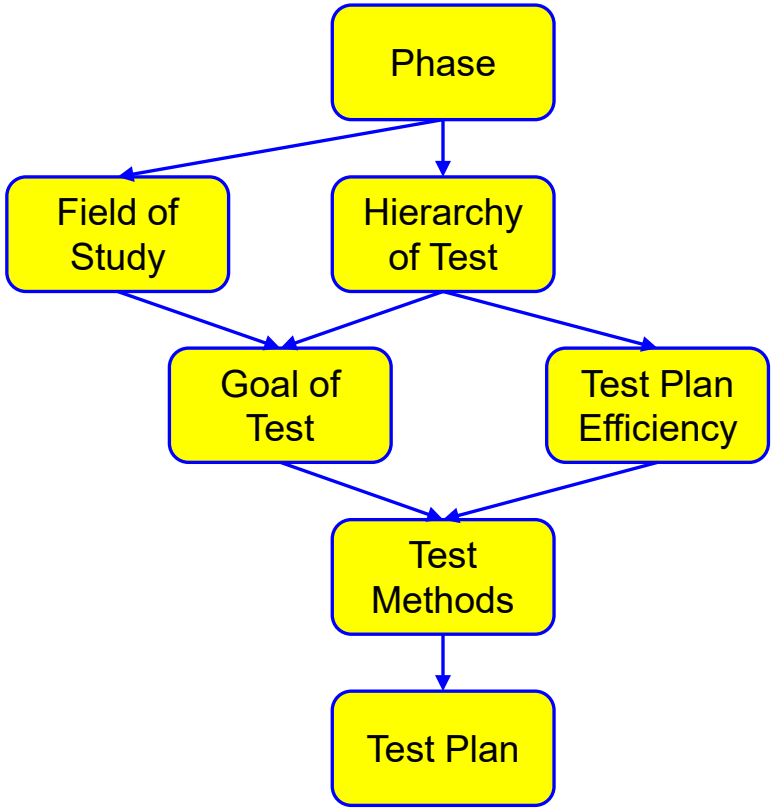
VTP Model


Possible Events during Operation
Deployment
Mission Objective Change
Payload Degradation
Adversarial Attack
Requirements & Analysis
V & V
Full System Testing
High-Level Design
V & V
Systems Integration & Testing
Detailed Design
V & V
Unit Testing & Integration
Implementation
Logic updates (ML), New Data Acquisition (Optimal Learning)
Fault Detection and Mitigation (DOE, Combinatorial Testing)
Resilience (ML, intrusion detection)
Time

Hierarchy of Test



Framework Process



Execution Illustration

System Under Test	Phase	Field of Study	Hierarchy of Test	Test Plan Efficiency	Goal of Test	Test Methods
AI Algorithm	Unit Testing	AI Mission Environment Knowledge	Components: - Functions in code - Program integrates functions - Simulated inputs	- Blackbox testing as scripts on simulated program inputs - Focus on bugs or misunderstood requirements to stress all possible types of inputs and conditions	Assess AI reliability, robustness, performance, and biases	CIT combined with optimal learning to ensure input space coverage and minimize unknown areas of performance

Table shows a high-level overview of how each framework concept contributes to guiding test plan development

Execution Illustration

System Under Test	Phase	Field of Study	Hierarchy of Test	Test Plan Efficiency	Goal of Test	Test Methods
Satellite	Systems Integration Testing	Above Plus: - Software - Electronic Hardware - Mechanical Systems	Components: - AI Algorithm, Sensors, Actuators, Control Software - Integration considers inputs from/outputs to components to other subsystems	- Integration testing of full system in simulated space environment - Focus on coverage of interactions between systems to identify faults or performance drop-offs	Assess system performance, reliability, cybersecurity, and interoperability	DOE using both known software integration areas and environment factors

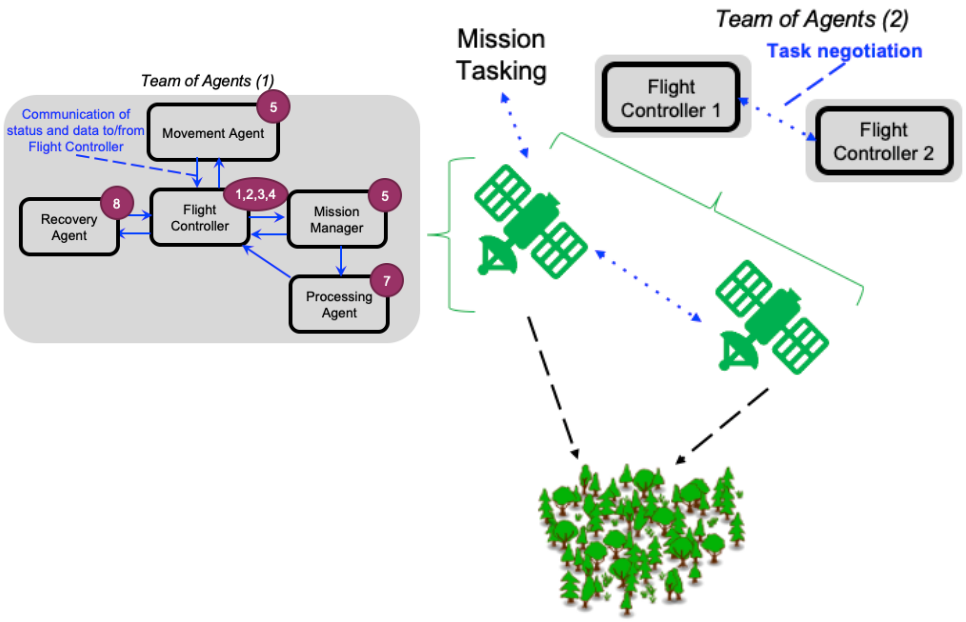
Execution Illustration

System Under Test	Phase	Field of Study	Hierarchy of Test	Test Plan Efficiency	Goal of Test	Test Methods
Constellation with Ground Station Operators	Operation	Above plus: - Psychology - Human-Computer Interaction - AIA teaming	Components: - Satellites - Ground station - Operators - Environmental conditions	- System in operational environment - Focus on achieving mission objectives, detecting changes in performance	Assess mission accomplishment	DOE with human's as factors (within-subject design)

Outline

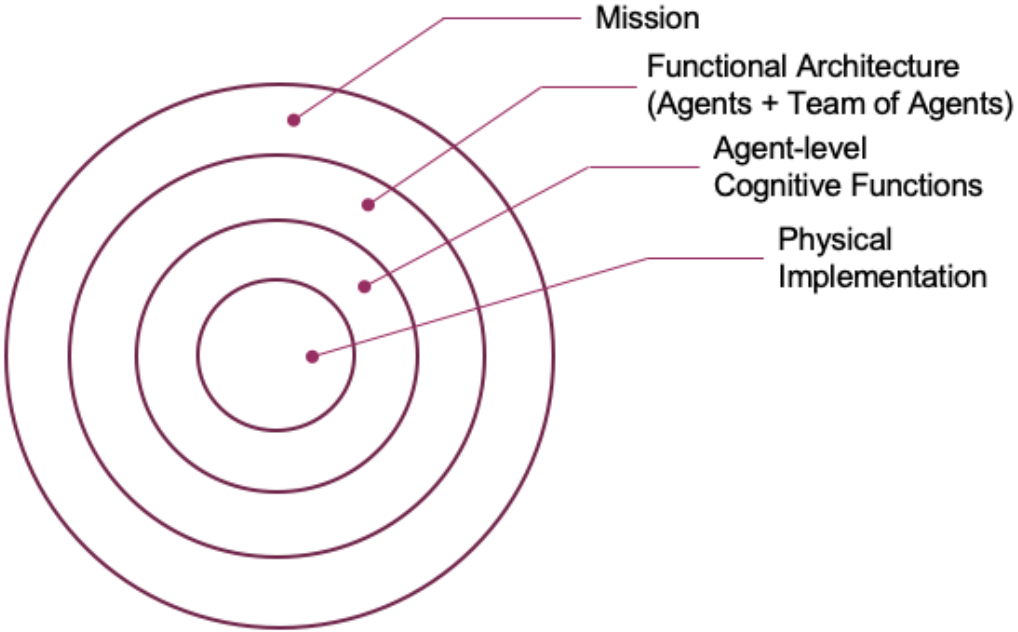
- Phase 1: Test and Evaluation Framework
- **Phase 2: Performance Measures Case Studies**
- Summary & Future Work

Case Study of System of Cognitive Agents



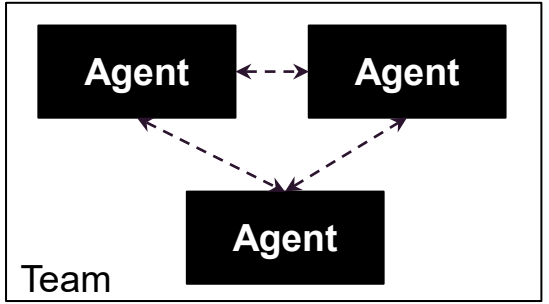
Mission	System-Level Task (Goal)	Responsible Agents
Look for forest fires in southwestern United States during summer	1. Observation plan (locations, orientations, and time) created	Flight Controller
	2. Task schedule created for all spacecraft	Flight Controller
	3. Data packed and sent to appropriate agents/ stations	Flight Controller
	4. Data and commands received and unpacked	Flight Controller
	5. Spacecraft in position at next observation location	Movement Agent
	6. Observation data taken of desired location	Mission Manager
	7. Fires recognized and characterized	Processing Agent
	8. Subsystems operating to accomplish their goals, with all faults responded to	Recovery Agent

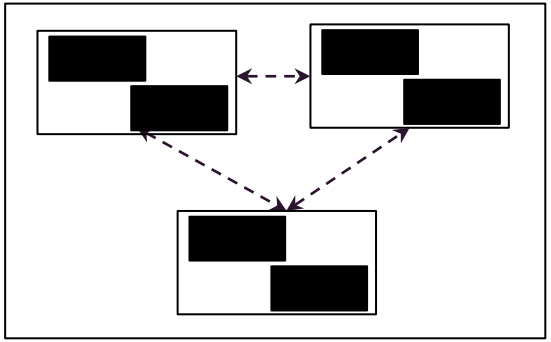
"Onion" Model of Performance Measures



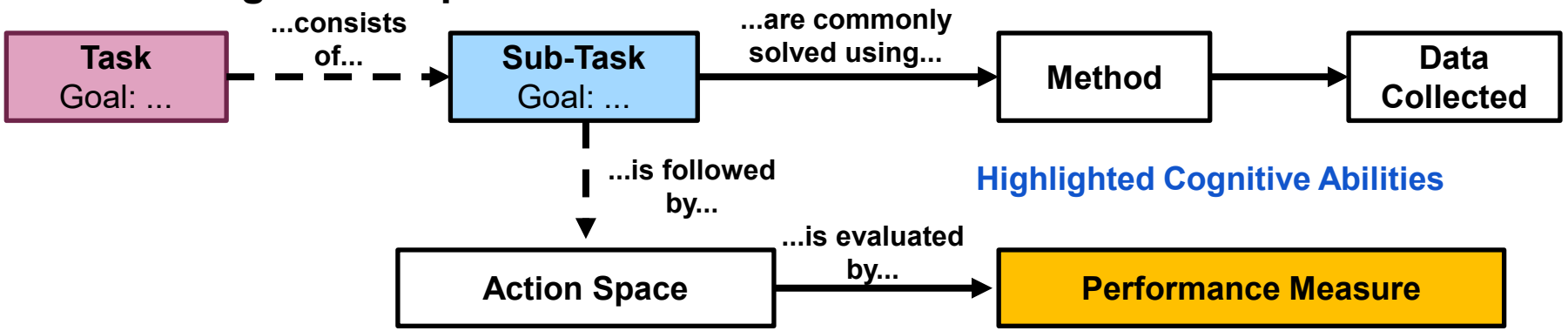
- **Black box** evaluation examines observable outcomes
- **White box** evaluation examines the implementation algorithms
- **Grey box** evaluation represents a place along the continuum where there some access to the system's processes

Mission Objectives:
- Provide canopy coverage over the United States
- Detect and track fire plumes in the United States





Block Diagram Template



Task: The work a system will undertake to achieve the mission.

*A **Task** must be comprised of 3 parts:*

1. **Goal:** a desired discrete end state
2. **Objective:** a criterion describing how well the goal was achieved
3. **Constraint:** a restriction or condition placed on the variables needed to achieve the goal

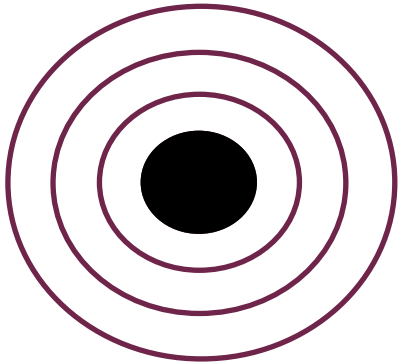
Sub-Task: A task necessary to complete *prior to* completion of the higher-level task.

Method: Can be algorithms, models, etc. that are commonly successful at performing the task.

Data Collected: Data or measurements that need to be collected, usually required to compute some performance measure at the end.

Summary of Results

"Onion" Model of Performance Measures



Measures of Effectiveness

- Mission Success - % Capture
- Canopy Coverage Completeness
- Fire Detection - Precision
- Fire Detection - Recall
- Probability of Tipping Success
- Fire Track Completeness
- System Availability
- System Recovery Timeliness

Agent	Agent Level PM	Team Level PM
Flight Controller Agent (part 1 – tasking agent)	Degree of satisfaction of preference constraints; utilization efficiency measure	Aggregate utilization efficiency
Flight Controller Agent (part 2 – communications agent)	Error rate (of command data being unpacked)	Aggregate data sets transferred to ground control over time
Flight Controller Agent (part 3 – observation agent)	Percentage of ground observed Percentage of ground observed, location-based coverage frequency	Aggregate revisit time, aggregate number of fires and properties identified, number of visits per location (or length of tracking time)
Mission Manager Agent	Duty cycle of sensors, error (MAE/RMSE) of predicted vs. actual power draw, error (MAE/RMSE) of area observed vs. desired area observed	N/A
Spacecraft Movement Agent	Propellant usage, reaction wheel saturation, error (MAE/RMSE) of predicted vs. actual propellant usage	N/A
Spacecraft Processing Agent	Number of true/false positives, number of true/false negatives, number of fires and properties identified	Aggregate number of fires and properties identified
Spacecraft Recovery Agent	Downtime of each system over time frame	N/A

Outline

- Phase 1: Test and Evaluation Framework
- Phase 2: Performance Measures Case Studies
- **Summary & Future Work**

Summary

- This year we:
  - Developed an overarching framework for thinking about test and evaluation
  - Developed case study specific performance measures for mission/tasks/agents
  - Initiated research into Agent-Team performance measures and analysis methods that would enable performance prediction in previously un-tested environments/tasks
- Deliverables
  - T&E Framework Report
  - Performance Measures for Multi-Agent Systems of Autonomous Intelligent Agents in Satellite Systems
    - Case Study Report

- Test and Evaluation Program should:
  - Establish a method for collecting data
    - This data should include performance measures useful for evaluation and prediction and other descriptive data that characterize the mission, tasks, and operating environment and provide diagnostic capabilities for explaining changes in performance measures
  - Establish a method for determining sufficient data collection to meet algorithm assurance requirements
  - Be useful for performing all the following decision-making functions:
    - Guide system and algorithm design by collecting data on the evolution of system capabilities, allowing for the iteration of multiple system architectures and algorithm selections
    - Understand system performance growth (or degradation) as a function of system maturity
    - Characterize performance as a function of operating environments and tasks
    - Make predictions about future environments/tasks
    - Quantify risk and make informed decisions based on risk reduction or risk tolerance

Future Work Vision

- Expect there is a high priority in getting fielded capabilities into space over the next 5 years
- Test and evaluation is a fundamental capability required to make informed decisions on fielding and operational use
- Future work should address both the direct application to real systems and expanded investigation into what new measurement and analysis capabilities are needed
- Proposed three strategic areas:
  - Implemented and integration of performance measures with programs
  - Expand and refinement of performance measures for:
    - Agent Teams
    - Cognitive architectures – expand performance measure development with SOAR architecture specific measures
  - Close the “Design – Test – Evaluation – Design” cycle
    - Supporting design of cognitive agents via trade studies of engineering configurations
    - Develop an online test and evaluation