

SERC RESEARCH REVIEW 2023 | NOVEMBER 15, 2023

Digital Engineering for Test and Evaluation

WRT-1070 and WRT-1071

Director Operational Test and Evaluation

Laura J. Freeman



SYSTEMS
ENGINEERING
RESEARCH CENTER

Mission Statement

- Transform T&E state-of-the-art to address:
 - Rapidly changing and technologies and systems that continually evolve over their lifespan
 - Support the DoD in rapidly providing warfighting capabilities to counter advanced threats and new technologies
- Areas of emphasis:
 - Digital Transformation
 - Transforming current processes (e.g., digital TEMP's), and
 - Developing new methods for T&E that leverage digital transformation.
 - Speed to Fielding
 - Middle Tier Acquisition - “is used to rapidly develop fieldable prototypes within an acquisition program to demonstrate new capabilities and/or rapidly field production quantities of systems with proven technologies that require minimal development.”
 - Speed of need.
 - Theme: data & models as the universal translator of information

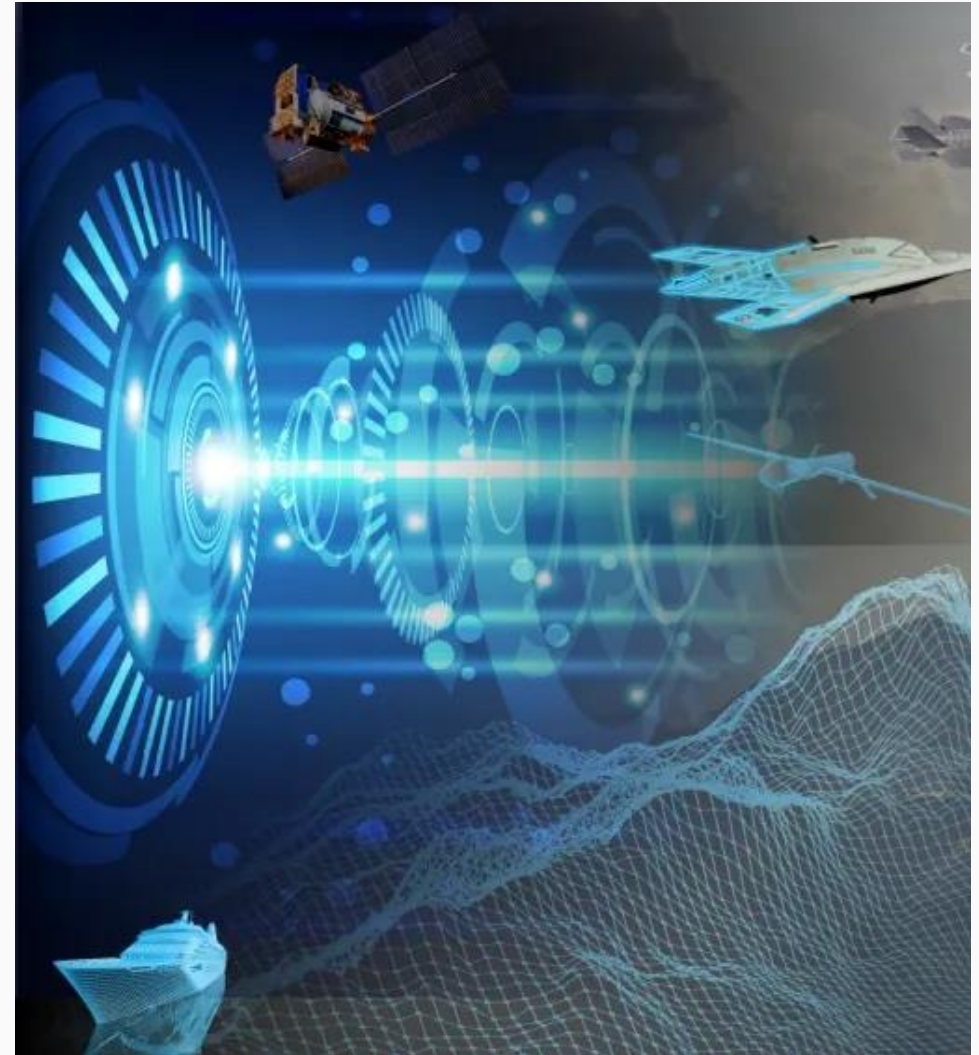
Agenda

- Motivation & Building Blocks
- Digital Transformation in T&E
- Deep Dive: T&E of Artificial Intelligence Enabled Systems
- Deep Dive: Digital Engineering and T&E
- Framework for Accelerating
- Concluding Thoughts

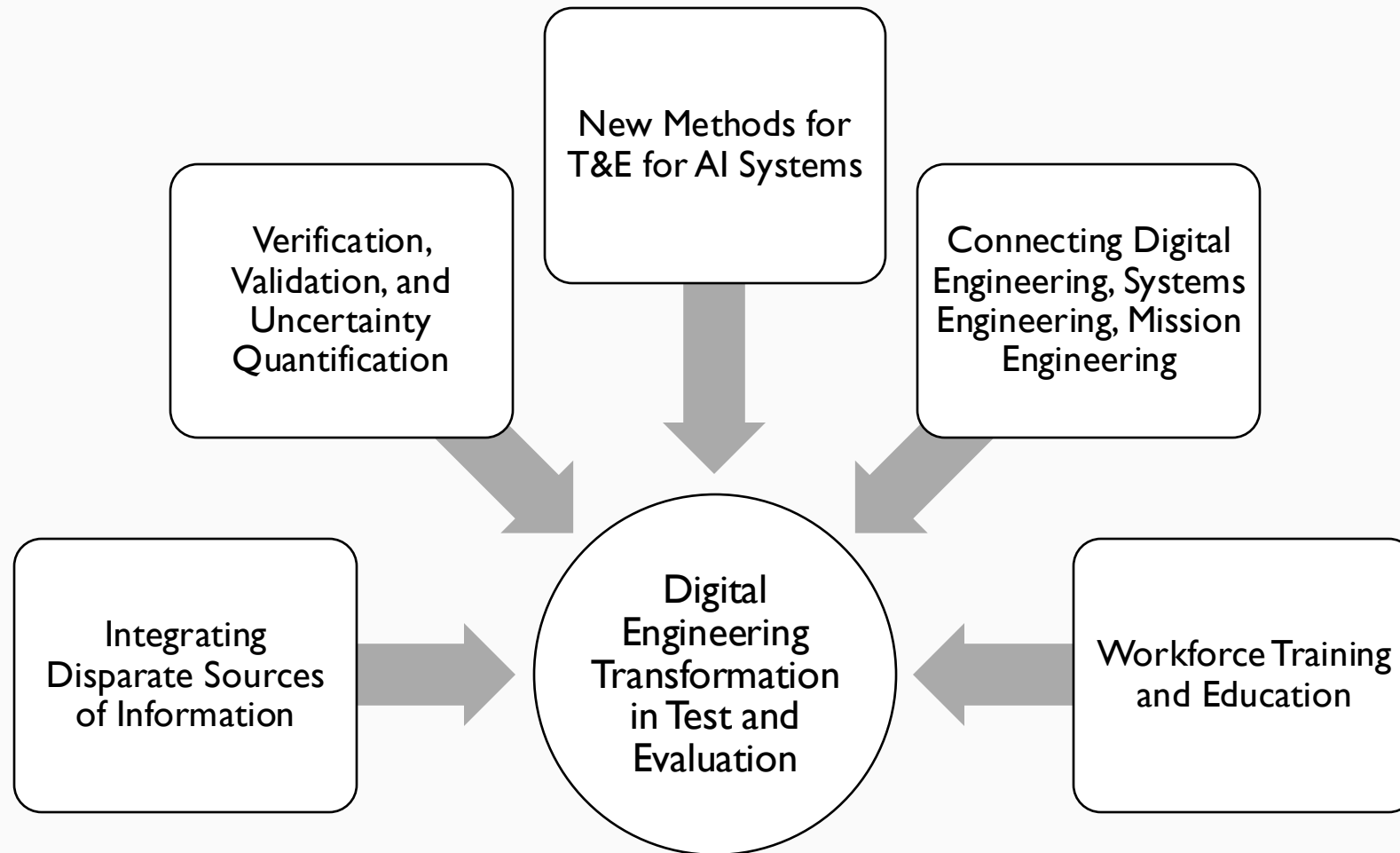
Changing Landscape

Strategic Drivers

- ✈️ Engineering of Software-Reliant Systems
- ✈️ Artificial Intelligence / Machine Learning
- ✈️ Joint All-Domain Operations
- ✈️ Data
- ✈️ Speed to Field
- ✈️ Culture
- ✈️ Talent Management

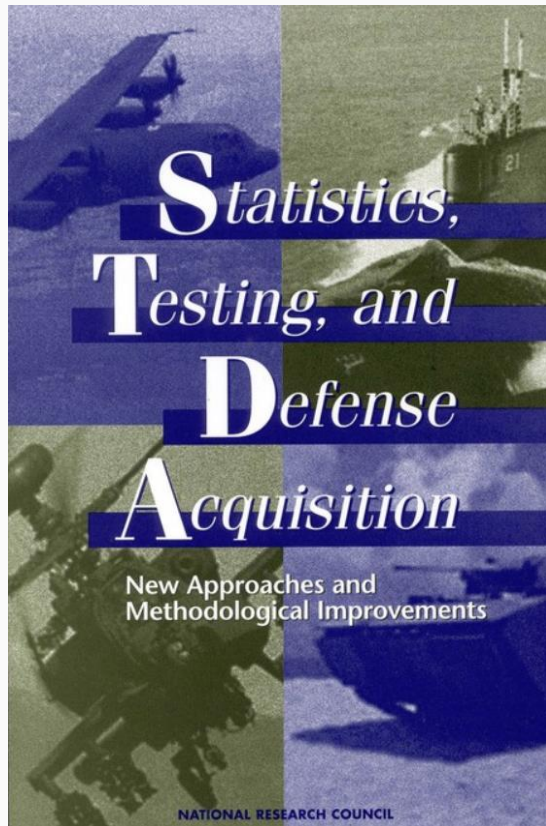


Transformation Building Blocks



Foundational Knowledge

Integrating Disparate Sources of Information



Improving Reliability Estimates with Bayesian Hierarchical Models
Kassie Fronczyk, Rebecca Dickinson, and Laura Freeman

THE PROBLEM
The reliability of a weapon system is an essential component of its suitability for operational deployment. Yet, in an era of reduced budgets and limited testing, verifying that reliability requirements have been met can be challenging, particularly using traditional analysis methods that depend on a single set of data coming from a single test phase.

In the Department of Defense (DoD), test data are often collected in several phases. The two broad types of testing are developmental testing and operational testing. The primary goal of a developmental test (DT) is to verify that a system meets its design specifications. This testing can occur as contractor testing, government testing, or a mixture of both and is usually carried out in a controlled environment that often lacks the realism of combat scenarios and trained users. The purpose of an operational test (OT), on the other hand, is to determine whether the system is effective and suitable in a combat scenario. OT data are collected under test conditions that replicate, as much as possible, field use.

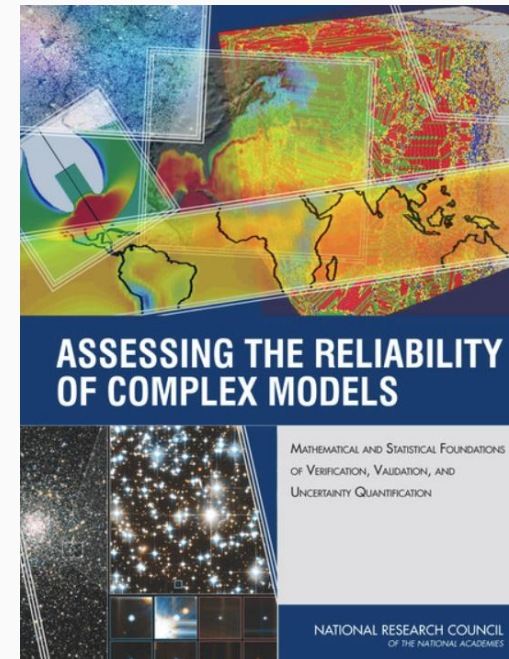
Reliability is one of the primary aspects of a system's operational suitability. It is important that a system perform as intended under realistic operating conditions for a specified period of time without failures. Reliability requirements for ground vehicles are often based on the mean number of miles between failures. A serious equipment failure that occurs during mission execution and results in the abort or termination of a mission is scored as an Operational Mission Failure (OMF). A less critical failure of a mission-essential component is scored as an Essential Function Failure (EFF). For example, an engine failure would be scored as an EFF if a vehicle took multiple attempts to start but eventually succeeded. If the vehicle could not be started, it would be scored as an OMF.

Requirements are typically written in terms of OMFs. Verifying whether the reliability requirements of a system have been met by looking at only a single test phase, however, can be challenging. Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates. The National Academies, in three

Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates... DoD [should] employ statistical approaches to capitalize on all available data from multiple test periods and not limit the reliability analysis to a single test period.

28 IDA | RESEARCH NOTES

Verification, Validation, and Uncertainty Quantification



Approved for public release; distribution is unlimited.

IDA INSTITUTE FOR DEFENSE ANALYSES

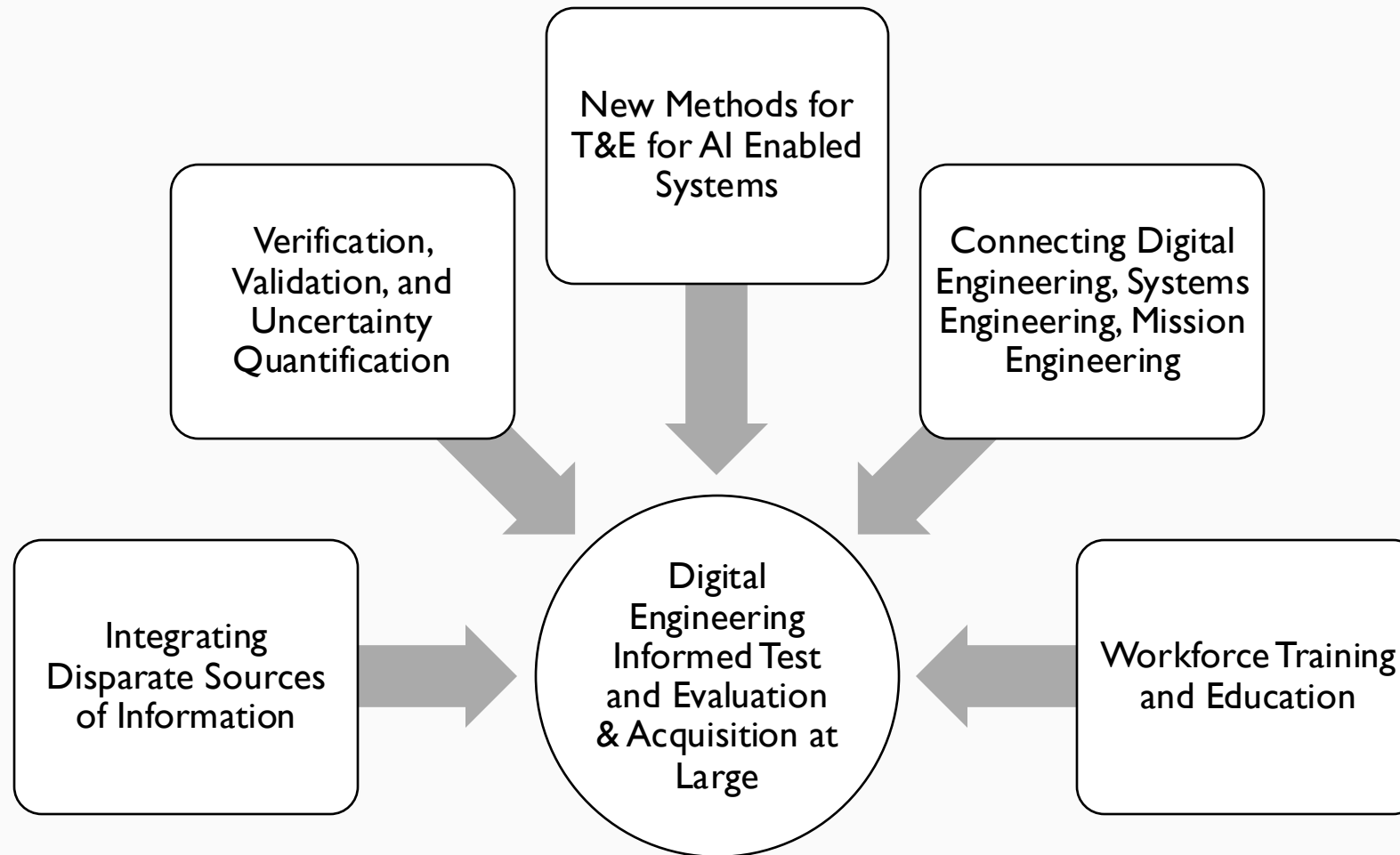
Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

February 2019
Approved for public release.
Distribution is unlimited.
IDA Document NS D-10455
Log: H 2019-000044

Transformation Building Blocks



T&E of AI-Enabled Systems: Best Practices (so far)

1. AIES require new measures for evaluation.
2. Data coverage measures are a component of test adequacy.
3. T&E programs for AIES must learn the critical factors (features) for each level of T&E.
4. T&E continuum for AIES should include data validation and model, system, and operational testing.
5. Experimental design provides methods for efficiently testing AIES across the continuum.
6. T&E programs must include robustness/adversarial test sets.
7. Side-by-side operational testing can quantitatively capture the impact of AI in AIES.
8. Data management is essential.
9. The T&E doesn't end until the system is retired and we need methods for monitoring and triggering events for independent tests.

The ITEA Journal of Test and Evaluation 2022; 43: 174-180
Copyright © 2022 by the International Test and Evaluation Association

Best Practices for Addressing New Challenges in Testing and Evaluating Artificial Intelligence-Enabled Systems

Laura Freeman, Ph.D.
Justin Kauffman, Ph.D.
Daniel Sobien
Tyler Cody, Ph.D.
Erin Lanus, Ph.D.

Virginia Tech National Security Institute, Arlington, VA

Introduction

The integration of artificial intelligence (AI) and statistical machine learning (ML) into complex systems exposes a variety of challenges in traditional test and evaluation (T&E) practices. As more decisions at varying levels are handled by AI-enabled systems (AIES), we need T&E processes that provide a basis for ensuring system effectiveness, suitability, and survivability. This involves methods for assessing the component ML models and AI algorithms, including the ability to show how they result in repeatable and explainable decisions, as well as an understanding of any failure modes and failure mitigation techniques. Moreover, there is a need for AI assurance to certify that AI algorithms operate as intended and are free of vulnerabilities arising either from faulty design or from adversarially inserted data or algorithm code. T&E needs new processes for characterizing the training data sufficiency for ML models, algorithm and model performance, system performance, and operational capabilities. Freeman (2020) outlined challenges facing current T&E methods for complex software-enabled systems, key challenges exacerbated by embedded AI, and 10 themes for how T&E will need to change for AIES [1].

In order to sufficiently test AIES, the T&E community needs to tackle the following challenges:

- determine testing requirements when state space size makes testing all cases infeasible or the open world problem makes enumerating all cases impossible;
- address the potentially invalid assumption that these emergent systems can be decomposed; and
- deal with dynamically varying systems that are potentially never in a "final" state during deployment [1].

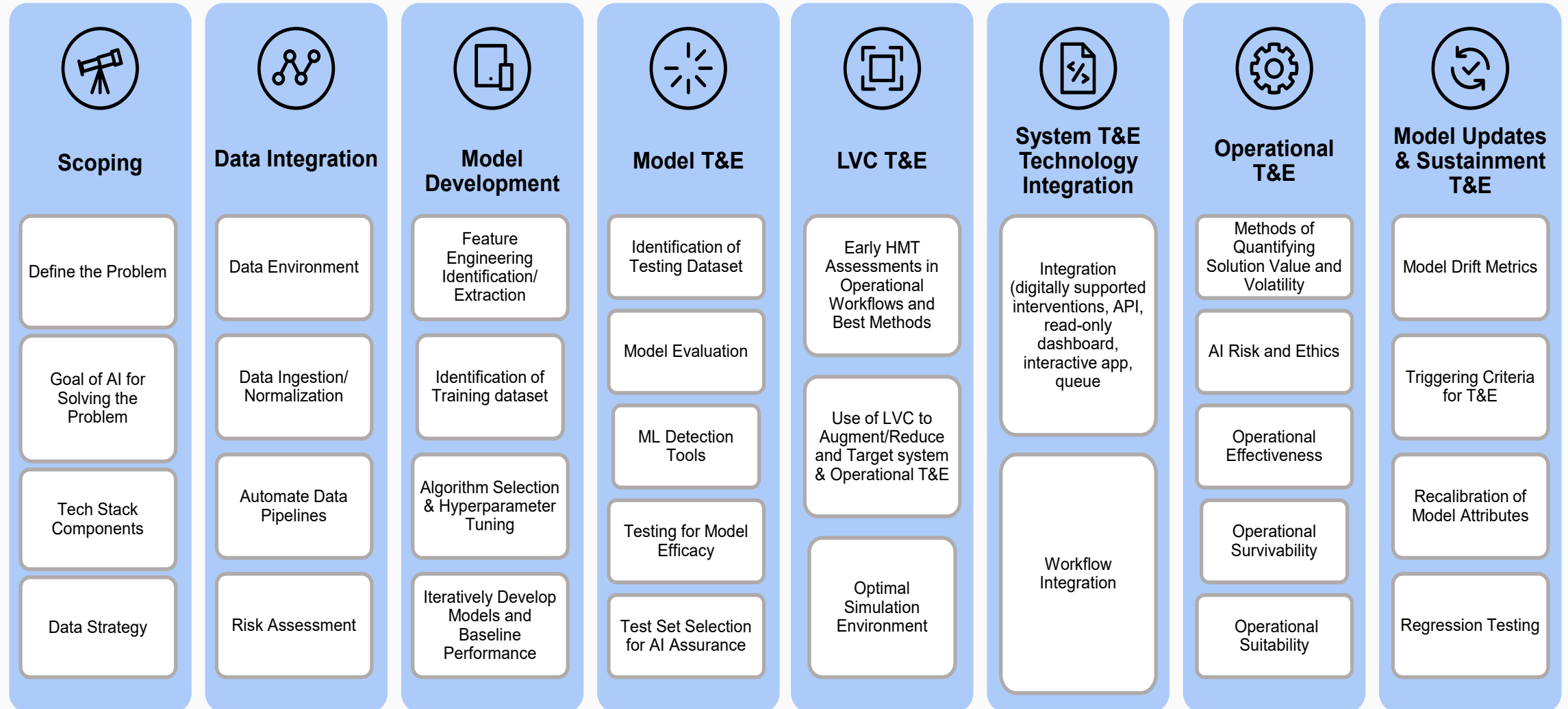
Figure 1 summarizes the 10 different themes outlined to enhance T&E in order to address the challenges with adequately testing and evaluating AIES. Over the past year, Virginia Tech has worked to test and evaluate a variety of AIES. This best practice guide adds further refinement and context to the themes in Figure 1. The best practices contained in this article translate these themes into executable T&E practices. In developing the guide, we leverage our experience working in T&E for both AI systems development and our work with the wider AI community. The best practices captured here reflect an initial attempt to make T&E for AIES tractable. These practices need to be tested against a variety of AIES to ensure they are truly best practices. One highlight carried through many of the best practices is the important role of data. Data is no longer just a product of T&E. It is now an input to the development of the AI system itself. This notable change drives new requirements and practices for T&E of AIES. Additionally, this list is far from complete and should be seen as a living documentation of practices. As more AI systems become available for testing, new practices will evolve and this list will need to be updated. However, each of the practices in this document has proven useful in testing DoD AIES.

T&E Best Practices to Address AI Challenges

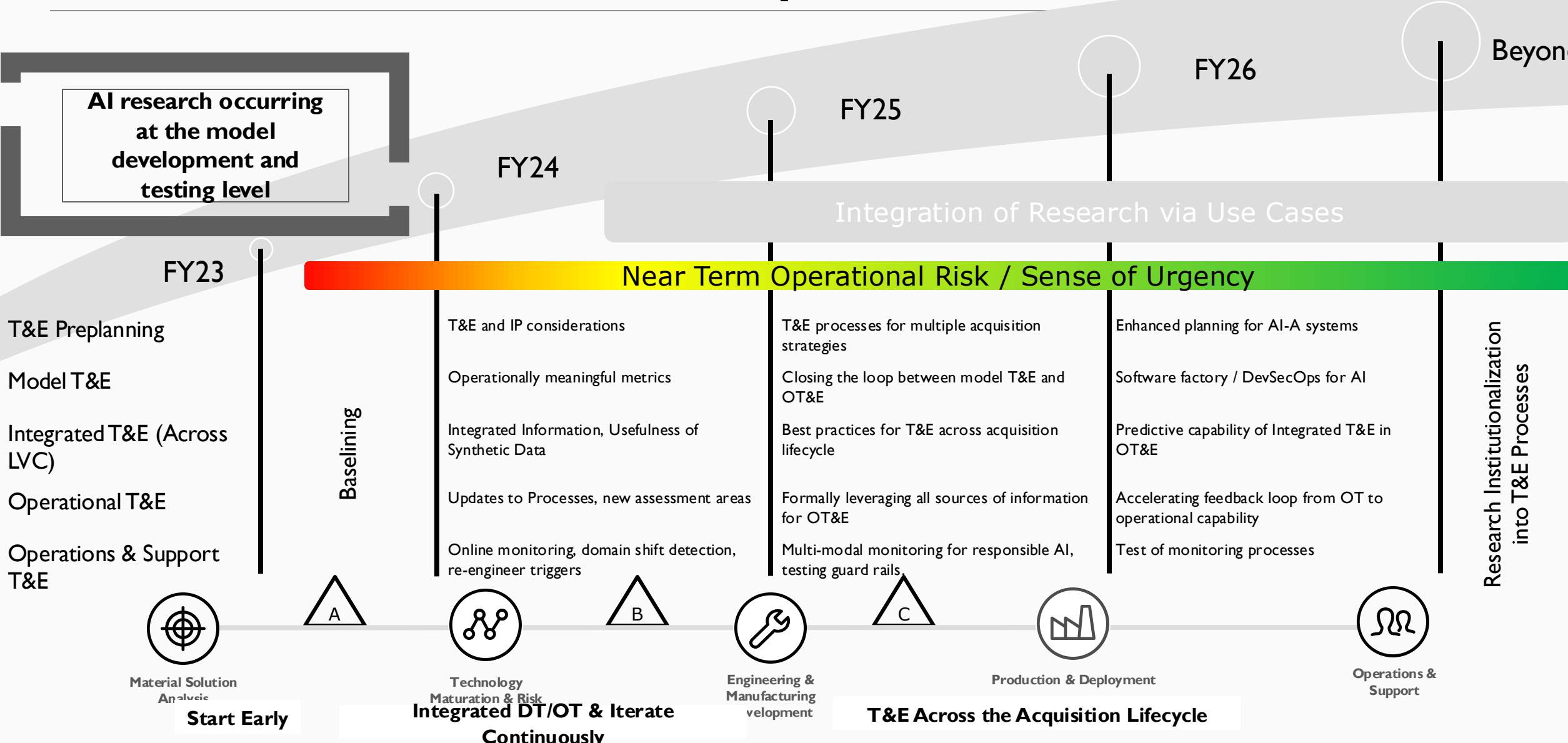
Best Practice 1: AIES require new measures for evaluation.

As mentioned in Freeman's 2020 article, AIES still require legacy evaluation measures for effectiveness, suitability, and survivability. However, these measures

T&E Research Ontology



Research Review & AI T&E Roadmap



T&E Research Needs & Responsible AI

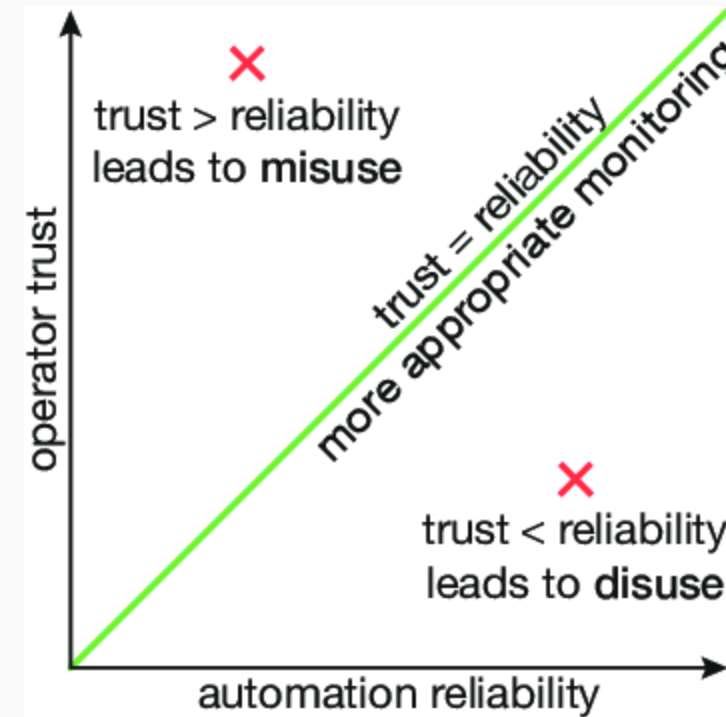
- **Responsible AI** – DoD personnel will exercise appropriate levels of judgement and care, while remaining responsible for the development, deployment, and use of AI capabilities.
- **Equitable AI** – The Department will take deliberate steps to minimize unintended bias in AI capabilities.
- **Traceable AI** – The Department’s AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies , data sources, and design procedures and documentation.
- **Reliable AI** – The Department’s AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
- **Governable AI** - The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

T&E Research Needs & Ethical AI Principles

- **Responsible AI** – DoD personnel will exercise appropriate levels of judgement and care, while remaining responsible for the development, deployment, and use of AI capabilities.

T&E Research Need

Calibrated trust –
can we verify that user's reliance on AI-
capabilities matches the capabilities.



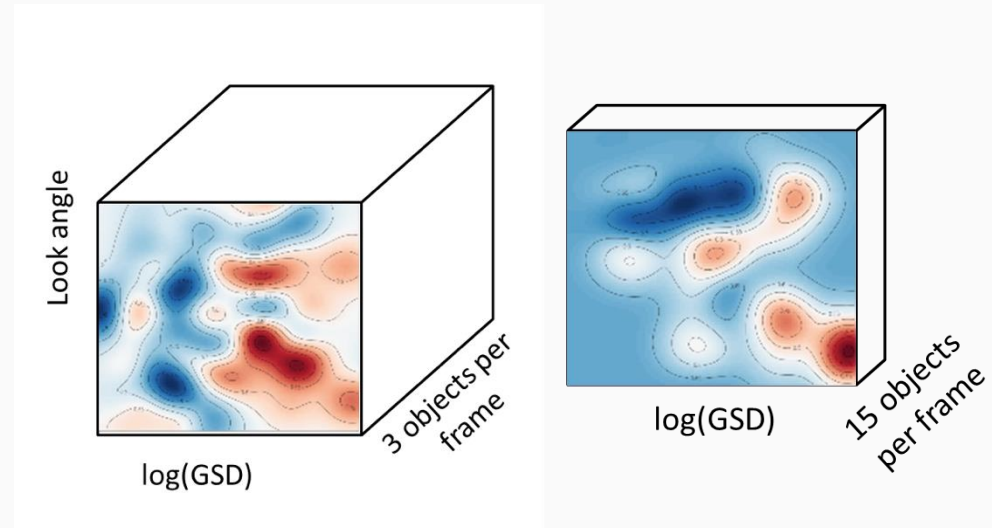
T&E Research Needs & Responsible AI

- **Equitable AI** – The Department will take deliberate steps to minimize unintended bias in AI capabilities.

T&E Research Need

Performance mapping –
across key factors for AI capabilities to detect
potential biases.

Training data adequacy –
assessment of training data coverage – to
predict / eliminate performances bias due to
skewness in the training data.



T&E Research Needs & Responsible AI

- **Reliable AI** – The Department’s AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

T&E Research Need

Model Robustness – How to adequately test generalizability in terms of robustness / adaptation.

Model Resilience / Security – cyber-T&E and new requirements for AI red teams for AI enabled systems.

Model Safety – how to test AI safety to include disengage and/or deactivate processes.

Test Efficiency - integrate information on effectiveness, suitability, survivability, and lethality across multiple test events spanning the system life-cycle.

AI Assurance - how to craft assurance cases for critical concerns (safety, security, etc.) that leverage various types of evidence

Defining test adequacy for AI enabled systems, specifically

- How to define the validated operating region for the AI
- How to learn what factors matter to the AI component
- How to best test black-box AI
- How to characterize the AI contribution to operational mission outcomes

T&E Research Needs & Responsible AI

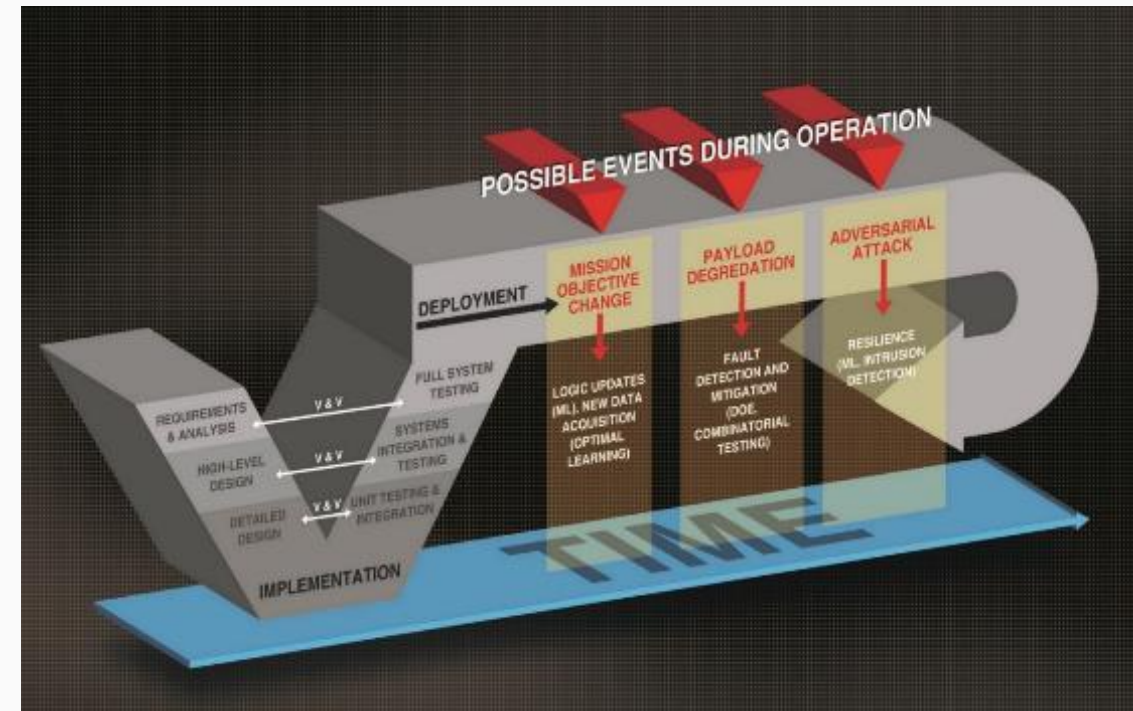
- **Governable AI** - The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

T&E Research Need

Drift Detection - How to best test if instrumentation for detection of drift domain shift is effective?

Predicting Unintended Consequences - How to predict (outside of domain shift), when a system using AI might produce unintended behavior?

Disengage/Deactivate – How to test disengage and/or deactivate processes?



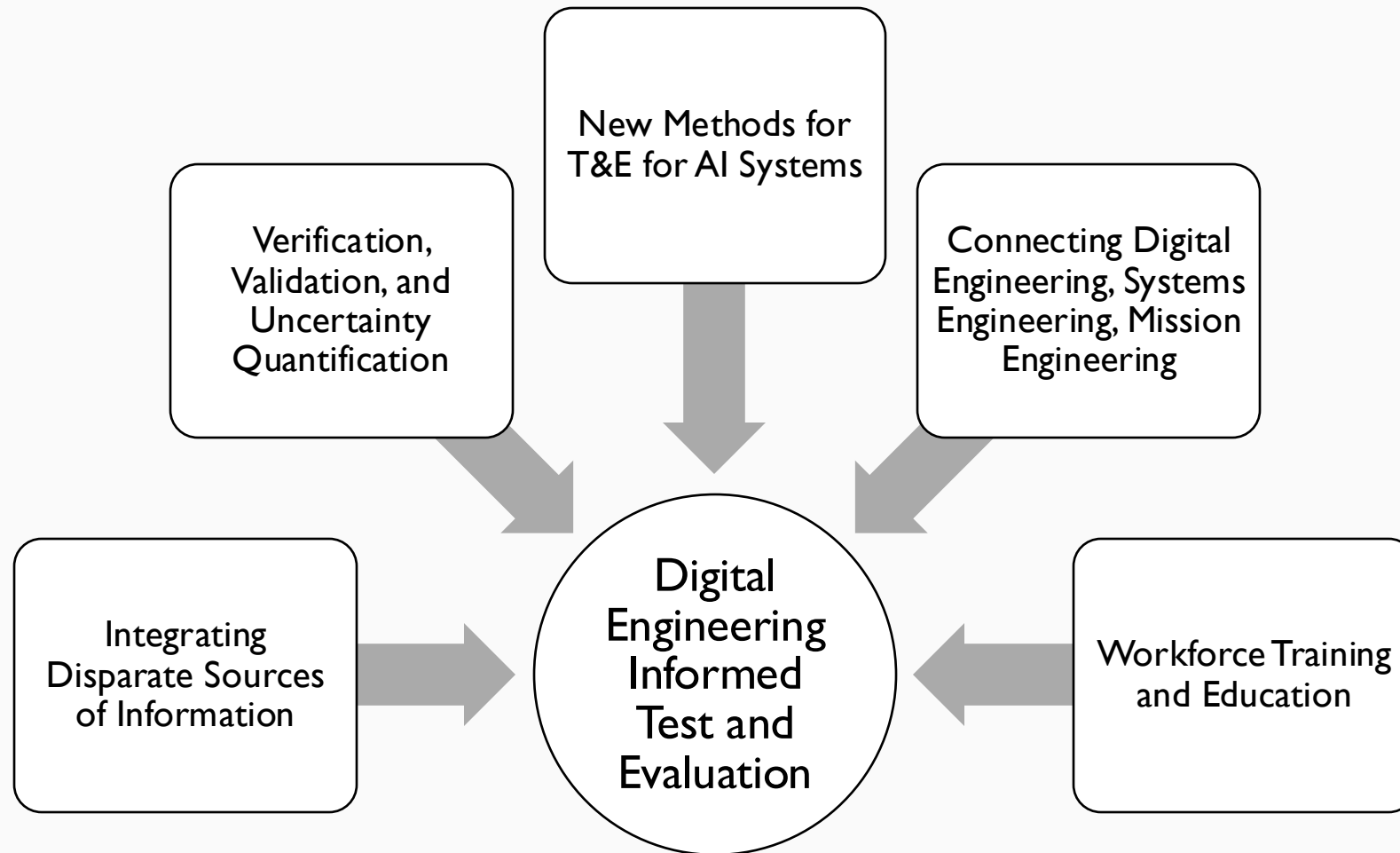
AI T&E Harness

Objectives:

1. Develop a T&E AI prototype environment
Accelerate the transition of research and methods into T&E tools
2. Prototype policy, standards, metrics, and risk frameworks
3. Accelerate education and training of T&E practitioners in prototype digital environment

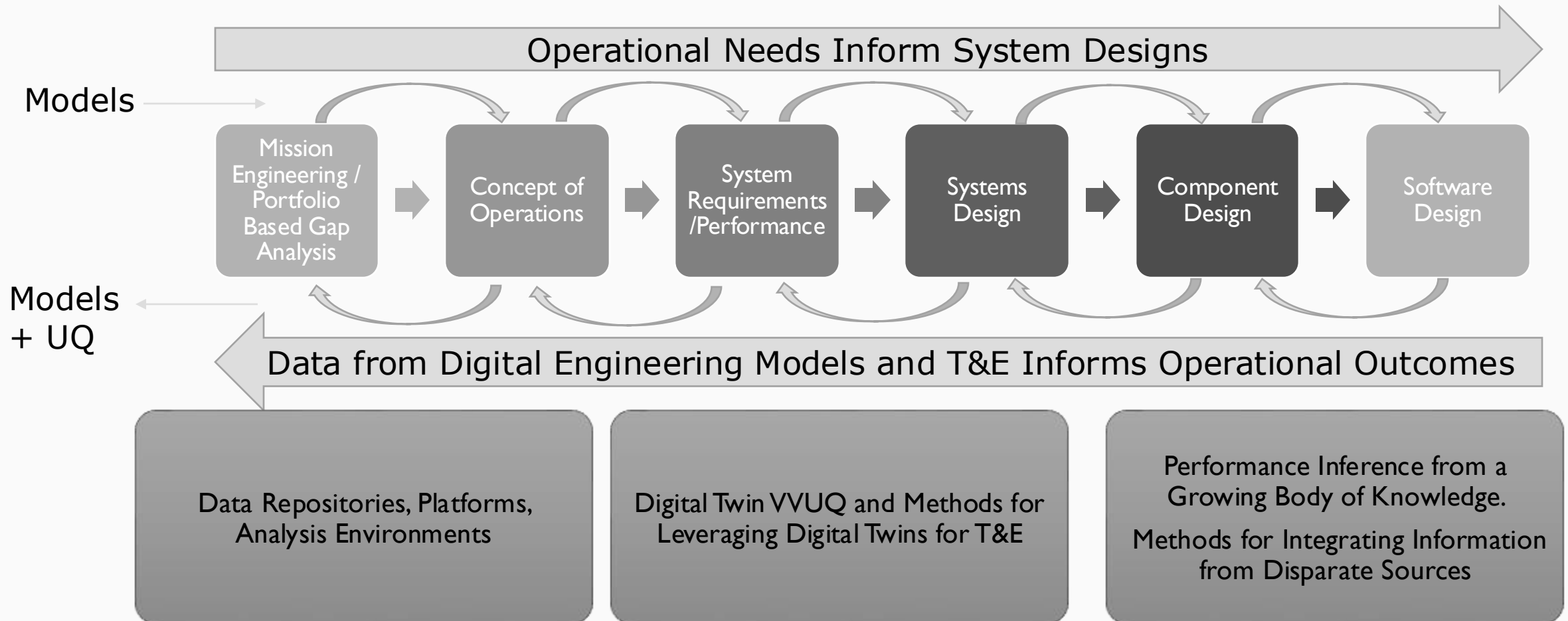


Transformation Building Blocks



Top-Down and Bottom-Up Approach

- Data, models, tools that inform from mission to system. Information can flow both ways and T&E results can inform both systems engineering more directly and mission engineering



DE FOR T&E: CURRENT STATE OF MBSE FOR T&E

MBSE for T&E is the use of MBSE to support test planning, test execution, and the analysis of test results to include integrating information from model-based data sources as appropriate.

- Overall Observation: government and industry are widely adopting and implementing MBSE tools and methods to support early system design and development
- On the contrary, the integration of MBSE with T&E has been limited
- Three Key Enablers for MBSE for T&E:

Culture / Workforce		Methods / Process		Interoperability	
Challenges	Recommendations	Challenges	Recommendations	Challenges	Recommendations
<ul style="list-style-type: none"> • Change from traditional methods • Lack of trust in tools • Expensive tooling • Unfamiliarity with MBSE 	<ul style="list-style-type: none"> • Train development team on model-based practices + benefits • Establish plan for level of modeling detail to be accomplished • Implement metrics for effectiveness of your model-based test planning 	<ul style="list-style-type: none"> • No modeling language standards for test planning • Different base system modeling languages throughout different organization 	<ul style="list-style-type: none"> • Adopt a standard ontology and modeling guide • Connect test planning to the mission objective and system design • Implement TEMP for digital connectivity with system designs and dynamic test planning 	<ul style="list-style-type: none"> • Tooling providers have unique data standards • Data compatibility issues due to versioning and configuration of tool installations 	<ul style="list-style-type: none"> • Use a mathematical representation of test strategy for verification and validation using a system modeling language • Promote and adopt open standards and data structures

The Undersecretary of Defense, Research and Engineering defines MBSE as “an integrated digital approach that uses authoritative sources of system data and models as an acquisition life-cycle across disciplines to support life cycle activities from concept through disposal .”

DoD Digital Engineering Strategy
June 2018

DE FOR T&E: 3-PHASE ROADMAP

Replacement of traditional paper-based T&E artifacts with digital artifacts that include descriptive and executional models

Benefits: Streamlined workflows, improved access and sharing of information, increased T&E information consistency, more efficient and effective change management, and enhanced knowledge management

The modeling framework established in Phase 2 is integrated with mission and systems engineering modeling frameworks, both of descriptive models and of executional models

Benefits: Ability to perform total system trade-off's (T&E drives ME and SE and vice versa)

Phase 2: Dynamic T&E Planning and Execution

Phase 1: Model-Based T&E Planning and Control

Phase 3: Coupled Mission, System, and T&E Decision Making

The modeling framework established in Phase 1 is augmented in this phase with quantitative methods that enable dynamically updating a test plan as results become available during its execution

Benefits: Higher efficiency of test planning and management due to improved use of current observations and past experience (only activities that are truly necessary are executed), increased accuracy of test planning due to earlier identification of potential technical risks and gaps

Our recommendation: make Phase 1 of the “MBSE for T&E” Roadmap the new baseline for all acquisition programs with respect to the adoption and implementation of MBSE and DE tools and methods

Levels for Model-based Test and Evaluation Master Plans (MB-TEMPs)

1. Digitize & digitalize TEMP

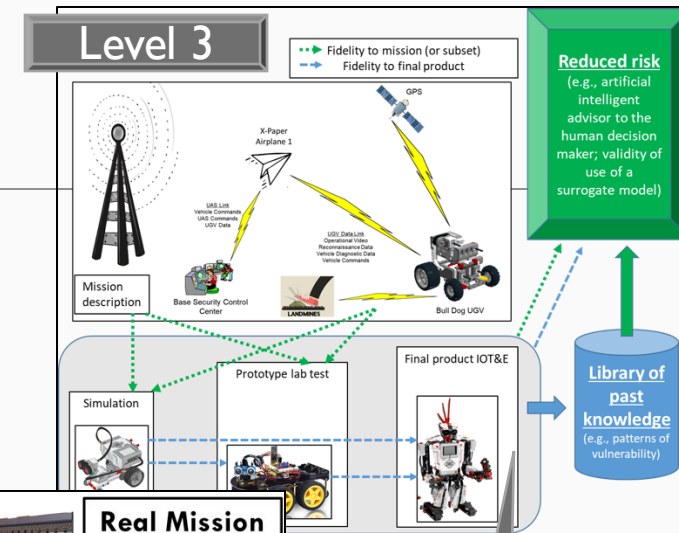
- Ex: Wiki-page with text, linked to schedule/cost info, model-based systems engineering (MBSE) tools
- Value: Begins conversion from text-based to data-driven; clarify test process; support complex concepts like test as a continuum

2. Condition-based decision analysis

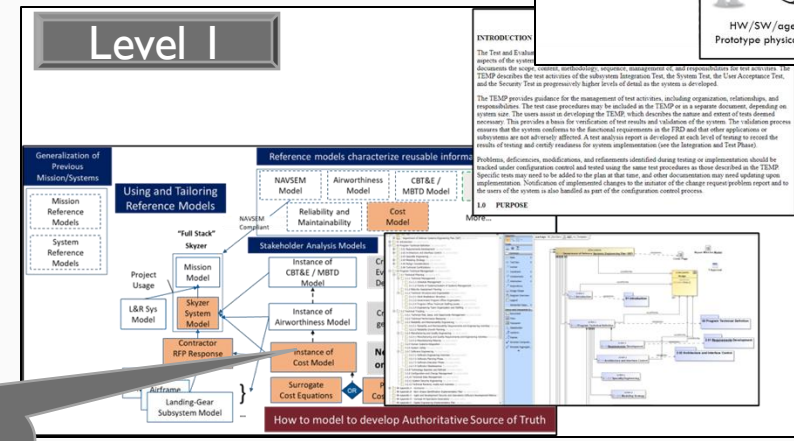
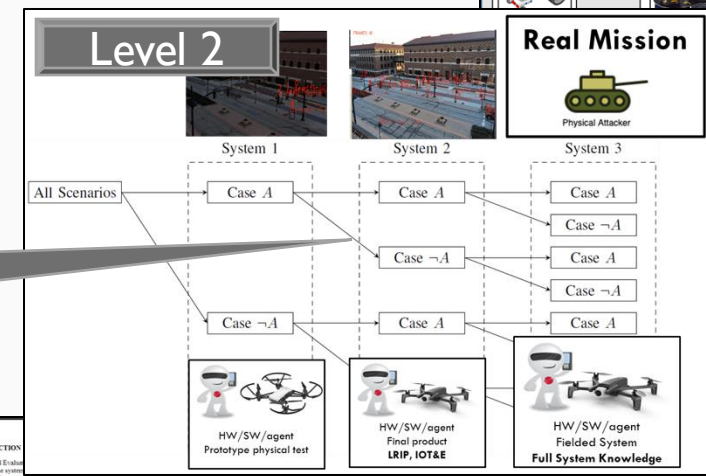
- Ex: Bayesian networks, utility theory
- Value: Metrics to assess risk/opportunity based on past experience; test efficiencies

3. Systems-based structure of data

- Ex: Mathematical definition of models
- Value: Reuse of knowledge; (artificial) intelligent advisor to human decision maker, characterization of fidelity & surrogate systems



Decision dependence



Set of reference models and visualization for decision makers

Reuse, artificial intelligence and machine learning (AI/ML), maps between validity & fidelity

Challenges and Opportunities

- Data standards for digital engineering
- Repositories of data and models for systems, missions, and environments
- Interoperable DE environments (platform centric versus mission centric)
- Balancing system complexity, model fidelity, and cost effectiveness
 - UQ – all models are wrong, some are useful
- Workforce and culture
- Policy and guidance

Concluding Thoughts: Keys to success

- Models linking system design, capability to mission outcomes have immense power to inform decision making at multiple levels
 - ✓ Data and statistical models (emulators) linking disparate sources of information are key enablers
- Shifting thinking in T&E from acceptance and compliance to formative evaluation focus supports rapidly evolving system design to support operational missions through digital representations
- Model based capabilities not only better inform operational evaluations, but also can prioritize testing requirements
- Early user involvement & feedback is key
- Progressive, sequential testing in LVC environments
- Ability to integrate all credible information in evaluations

Thank you

Stay connected with SERC Online:



Email the presenter: Laura Freeman

✉ laura.freeman@vt.edu

Email the research team:

✉ [Peter Beling \(beling@vt.edu\)](mailto:beling@vt.edu)
[Geoff Kerr \(geoffreykerr@vt.edu\)](mailto:geoffreykerr@vt.edu)



SYSTEMS
ENGINEERING
RESEARCH CENTER