INSTITUTE FOR DEFENSE ANALYSES

What Counts as Progress in the T&E of Autonomy? (Conference Briefing)

David M. Tate

October 2019
Approved for public release;
distribution is unlimited.
IDA Document NS D-10922
Log: H 19-000548

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About this Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, project number AX-1-3542, "Policy, Guidance, and Resource Analysis for the Director of Developmental Test and Evaluation," for the Director of Developmental Test & Evaluation (D,DT&E). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments

Thank you to David A. Sparrow for performing technical review of this document.

For More Information:

David M. Tate, Project Leader
dtate@ida.org, (703) 575-1409

Steve Warner, Director, System Evaluation Division
swarner@ida.org, (703) 845-2096

Executive Summary

There is widespread interest and concern within the Department of Defense (DoD) regarding the test, evaluation, verification, and validation (TEVV) of military systems with autonomous capabilities. For such systems to be fielded and used, it will be necessary that senior decision-makers be sufficiently confident in the systems' dependability (e.g., safety, security, reliability) to authorize deployment. It will also be necessary to understand any operational limits needed to ensure dependability, such as restrictions on geographic locations, weather conditions, or other environmental factors. To support these decisions, developers and testers will need to produce effective *assurance cases*.

An assurance case is a structured argument that a system is sufficiently dependable to permit fielding in a specific range of operational contexts. Existing standards and regulatory bodies already require explicit assurance cases for complex systems, with regard to safety, cybersecurity, and reliability. Researchers at the Institute for Defense Analyses have been working with various offices in DoD to develop a framework for structuring and executing assurance cases for systems with autonomous capabilities, and to understand the implications of this framework for TEVV. In particular, we consider systems that feature one or more of:

- Perception
- Reasoning
- Planning
- Course of action selection
- Learning
- Self-organizing (or emergent) behavior
- Human-machine teaming

IDA identified ten recurring challenges associated with TEVV of these capabilities, and mapped those ten challenges to applicable existing test methodologies and potential novel test methodologies that could help to address them. This analysis was the basis for the Defense Acquisition University (DAU) continuous learning module CLE002 "Introduction to Test and Evaluation of Autonomous Systems", which became available to acquisition workforce personnel in September 2019.

This briefing describes the IDA framework and CLE002 contents in detail, tracing TEVV requirements from the desired properties of the assurance cases backward through effective arguments, evidence needed to support those arguments, measurements needed to provide that evidence, and instrumentation needed to take those measurements. It closes with a discussion of high-priority research areas that would support development of these TEVV tools.

What Counts as Progress in the T&E of Autonomy?

David Tate

Institute for Defense Analyses

Work sponsored by OUSD(R&E)

October 2019

Cleared for Open Publication, Oct 15, 2019, DoD Office of Prepublication and Security Review; Case# 20-S-0051

The goal is assured effectiveness and dependability

Autonomous capabilities don't help if we're not sufficiently confident to field and employ the systems

There will always be *some* kind of certification or licensure or acceptance testing process

May have multiple certification authorities (e.g. Safety, Cybersecurity, Effectiveness, Reliability…)

Aside: Performance vs. Robustness

Historically, T&E has usually been more about establishing what the system can do than about confirming what it won't do

For safety-critical or high-assurance systems, it's the other way around – which has caused problems for T&E

Autonomy is more like the latter than the former

This has important implications for requirements

State of the Art: Assurance Cases

An **assurance case** is a structured argument that the system is sufficiently dependable to permit fielding in a specific operational context.

Existing standards and regulatory bodies already require explicit assurance cases for complex systems:

- Safety cases
- Software assurance cases (cybersecurity)
- Reliability cases

Currently, these are stovepiped requirements.

Example: SAE JA1002 (2004) Software Reliability Program Standard

7.2 Role of the Software Reliability Case—The Software Reliability Case **presents arguments and evidence that the requirements can be achieved, will be achieved, and have been achieved**. For maximum effect, the Case should be developed and witnessed as development decisions are made. It is not intended to be a retrospective justification of the solution.

The Software Reliability Case should be a readable overview of the evidence that the software meets its reliability requirements, with references to project development records and the results of analyses of software components as appropriate. The Case provides significantly more than proof that the Plan has been executed as it provides evidence about the products of the development process. This evidence should address the direct evaluation of the reliability of the software elements (e.g., from analysis of the design and reliability tests and trials), and also the suitability of the software architecture and the software engineering process. The Case should be a living document and its development should proceed through a number of stages of increasing detail during the project.

Assurance cases require both *evidence* and *arguments*

A pile of evidence is not an argument

An argument without evidence is unconvincing

The wrong evidence doesn't help

The outputs of TEV&V must provide the evidence that supports convincing assurance cases

Where does the evidence come from?

Traditional assurance cases are based on:

Exhaustive testing

Formal verification

Design of experiments

Run-time monitors

Human in the loop + training

Autonomy breaks this model

Can’t test exhaustively – state space is too large.

Can’t rely solely on Design of Experiments – we don’t know what all of the factors are, and we can’t assume smooth behavior between design points

Interactions among run-time monitors and core functions add complexity (and a need for additional testing)

Human-Machine Teaming (HMT) explodes both state space and factor set

We don’t need to define autonomy formally

Focus not on what autonomy is, but on the system attributes that make behavior *unpredictable*:

- State space explosion
- Non-smooth or fractal response
- Lack of transparency
- Changing system behaviors over time
- Emergent behaviors

Autonomous systems that don’t have these features don’t pose any new problems for TEV&V in terms of assurance

The specific culprits are:

Perception

Reasoning

Planning

Choosing

Learning

Self-organizing behavior

Human-machine Teaming

I could easily spend several slides on each of these and how they make assurance harder – but not today

This isn’t about “levels of autonomy”

“Level of autonomy” is not the relevant metric for how hard TEV&V will be – partial autonomy often requires more complicated human-machine teaming

Assertion: real-time teaming with humans requires more sophisticated assurance cases than full autonomy would

We identified 10 specific challenge areas for TEV&V

Transparency	Instrumenting machine thinking
	Linking system performance to autonomy
	Comparing AI models to reality
Training data	Quality of inputs to machine learning
Hazards	Elevated safety concerns
	Exploitable vulnerabilities
	Emergent behavior
	Post-fielding changes
Human-machine interaction	CONOPS and training as design features
	Human trust

Four primary categories

Ten specific challenge areas

All driven by the autonomous capability areas listed above

Feedback welcome

We mapped these to specific autonomous capabilities…

		Perception	Reasoning	Deciding	Learning			Emergence	Human-Machine Teaming			
					Supervised Learning	Unsupervised Learning	Reinforcement Learning	Self-Organizing Behavior	Shared Situational Awareness	Mutual Understanding of Goals	Human-Machine Comms	HMT CONOPS, TTPs, and learning
Transparency	Instrumenting machine thinking	X	X	X		X			X	X		X
	Linking system performance to autonomy	X	X	X			X	X	X	X	X	X
	Comparing AI models to reality	X	X		X				X	X		
Inputs to ML	Quality of inputs to machine learning				X	X	X					X
Hazards	Elevated safety concerns		X	X				X	X	X		X
	Exploitable vulnerabilities	X	X	X	X	X	X	X			X	
	Emergent behavior			X		X	X	X			X	X
	Post-fielding changes					X	X					X
Human-machine interaction	CONOPS and training as design features								X	X	X	X
	Human trust								X	X	X	X

X indicates a need for TEV&V tools

...and to existing or proposed T&E tools and methods

		Standard tools and methods				Novel tools and methods				
		Design of Experiments	Observational Studies	Surveys	Modeling and Simulation	Formal Methods	Explainable AI	Cognitive Instrumentation	Adaptive Red-Teaming	Rapid Sequential Test Design
Transparency	Instrumenting machine thinking				X		X	X		
	Linking system performance to autonomy	X	X	X	X		X	X	X	X
	Comparing AI models to reality	X			X		X	X		
Inputs to ML	Quality of inputs to machine learning	X			X			X	X	
Hazards	Elevated safety concerns				X	X	X	X	X	X
	Exploitable vulnerabilities	X			X	X	X	X	X	
	Emergent behavior	X	X		X	X		X	X	X
	Post-fielding changes				X	X			X	X
Human-machine interaction	CONOPS and training as design features	X		X	X				X	X
	Human trust		X	X	X		X	X		

The old tools aren’t useless…

Design of Experiments

Observational Studies

Surveys (esp. for human-machine teaming)

Modeling and Simulation

So what does progress look like?

Claim:
Things that improve our ability to make convincing assurance cases for systems with autonomous capabilities count as progress in the T&E of autonomy

(There is probably a lot of overlap with what will be needed for diagnosis / debugging during development.)

Everything else… Not so much, at least not immediately.

…so new specialized methods will be needed as well

Formal methods

Instrumenting cognition / explainable AI

Intelligent adversarial testing

Rapid automated sequential test design

Example: Formal Methods

Formal Verification of Human-Automation Interaction

Asaf Degani, NASA Ames Research Center

Michael Heymann, Technion

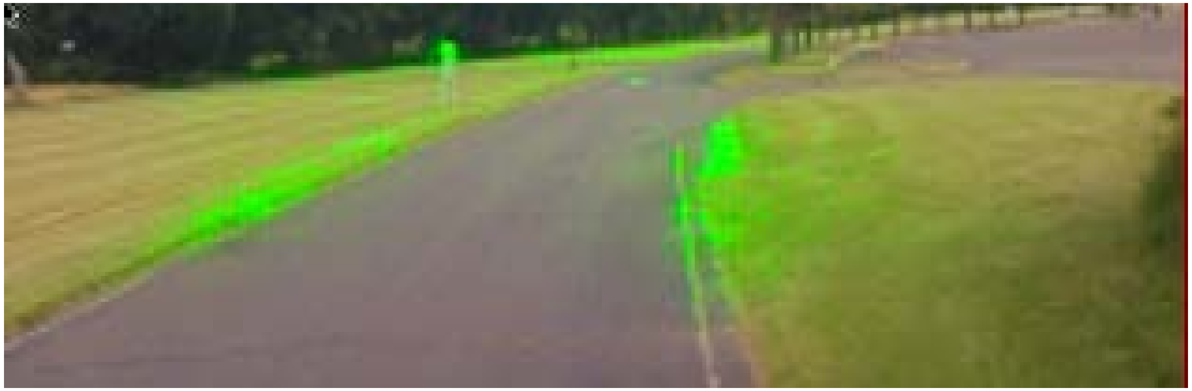
HUMAN FACTORS **44** #1, Spring 2002

Using Formal Verification to Evaluate Human-Automation Interaction: A Review

Bolton, Bass, and Siminiceanu

IEEE Transactions on Systems, Man, and Cybernetics: Systems **43** #3, May 2013

Example: Instrumenting cognition



Salient pixel analysis of the NVIDIA PilotNet self-steering system shows that the system all but ignores the road surface itself, focusing instead on features that indicate not-road. This system does not maintain an internal representation of the terrain; the neural net generates steering commands based on the real-time camera inputs.

Image from Bojarski et al., *Explaining how a deep neural network trained with end-to-end learning steers a car*. arXiv:1704.07911v1 [cs.CV] 25 Apr 2017

Example: Intelligent adversarial testing

The **Range Adversarial Planning Tool** (RAPT) developed at Johns Hopkins Applied Physics Laboratory generates test-planning products by using data from simulations of the autonomy software.

Gaussian Process Regression (GPR) is used to form a model of the autonomy performance and to identify regions of the configuration space that show steep gradients in response, indicating possible edge cases.

This information is then used to generate test designs that balance coverage with oversampling of the high-interest regions.

This characterization of the challenges is being used to train testers

CLE002

“Introduction to Test and Evaluation of Autonomous Systems”

Now available at DAU online Training Center

We expect it to evolve as the field advances

CLE 002 course outline

1. Introduction (includes discussion of DoD policies)
2. How autonomous capabilities affect T&E

    Perception, Reasoning, Planning, etc.

3. How human-autonomy interactions challenge T&E
4. Familiar methods that apply to T&E of autonomy

    STAT, Modeling and Simulation

5. Novel methods to address remaining challenges

    Formal methods, cognitive instrumentation, etc.

6. Resources available to support T&E of autonomy
7. Where challenges fall within the acquisition life cycle

Implications for instrumentation

It will be difficult if not impossible to argue convincingly for the dependability of Perception, Reasoning, Planning, etc. without evidence that they are working as intended.

“Working as intended” is about process, not just results. We care about getting the right behavior **for the right reasons**, because the reasons extrapolate to situations that will not be explicitly tested.

The stovepipes aren’t helping

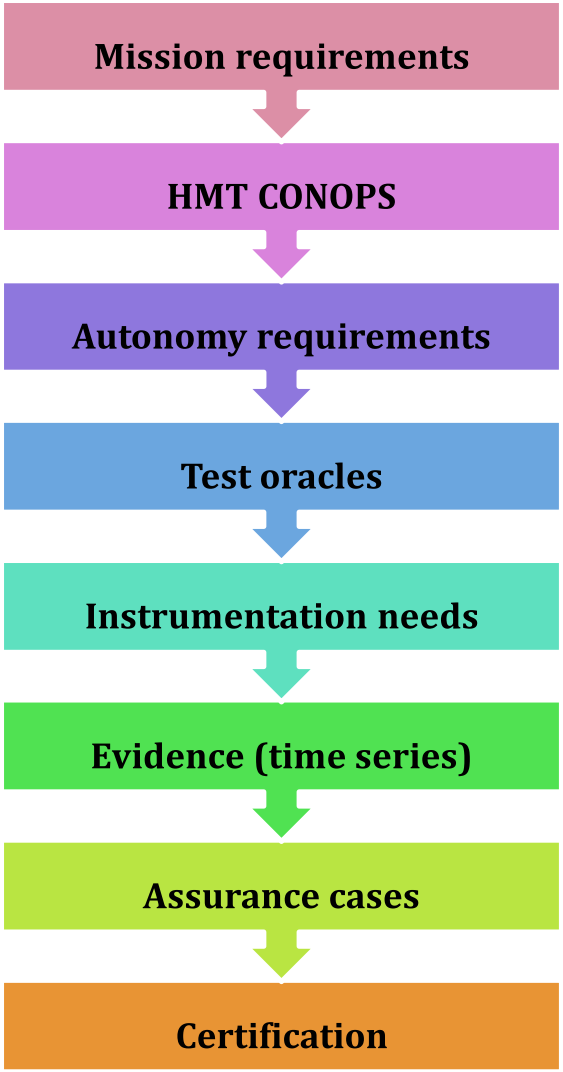
I assert without proof:

It will be more difficult and less efficient to make separate cases for safety, security, reliability, etc. than it will be to make unified cases for overall dependability

Rationale:

The stressing characteristics of autonomous systems are the same for all of these dimensions, and the “attack surfaces” are mostly common across them. Safety, security, reliability, and mission effectiveness are coupled at a level well below the “black box”.

Immediate research priorities



Specifying testable autonomy requirements

- Perception, Reasoning, …
- Teaming and self-organizing
- Negative requirements

Test oracle specification

- Technology (in)dependent
- Human-machine teaming
- Automated generation of oracles

Formal verification methods

Instrumenting cognitive functions

- Aligned with oracles
- Support both assurance and trust

Virtual test environments

Logic of assurance cases

- Multi-legged arguments
- Combining formal and empirical
- Composability criteria

V&V of ML training and training data

Intelligent adversarial testing

Regression testing criteria

IDA

Backup

Recurring Challenges

1. **Instrumenting machine thinking**
   In order to be able to diagnose the causes of incorrect behavior or inadequate performance, it will be necessary to be able to tell whether the problem lies in the Perception, the Reasoning, or the Deciding functions of the autonomous system. It will also be necessary to distinguish coding errors from inadequate algorithms or bad data. Without the ability to instrument and monitor internal states of the autonomy, diagnosing problems will be slow at best and impossible at worst.
2. **Linking system performance to autonomy**
   In complex collaborative activities, it can be very difficult to figure out what is enabling (or hindering) success. For example, on a soccer or basketball team it can be very difficult to pinpoint which players (and which behaviors) are leading to wins and losses. To design and improve autonomous systems, it will be necessary to figure out how the system's various autonomous capabilities interact to enable (or hinder) mission execution.

3. **Comparing AI models to reality**
   Autonomous systems represent reality through stylized internal models. Perception provides inputs for these models; Reasoning allows them to be expanded and corrected. The ability of an autonomous system to do its mission will depend on the degree to which the internal modeling of reality supports accurate Perception, valid Reasoning, and effective Deciding. This will not generally be a function of how detailed the models are ("high resolution"), or even of how closely the models mirror reality ("high fidelity") – it will be a function of whether the <u>right kind</u> of information is incorporated into the model, and that the resolution and fidelity be enough to support the mission needs. Test and Evaluation will necessarily include prototyping and experimentation to figure out what kind of internal model, using what kind of representation, is needed to achieve both performance and dependability.

Recurring Challenges (continued)

4. **CONOPS and training as design features**
   To date, the paradigm for designing systems has been to make a reasonable guess about how the operator will use that system, and what would be a good user interface, and to work out the details of CONOPS, TTPs, and training long after the basic design has been decided. For autonomous systems, where the system operates itself and interacts autonomously with humans, the details of CONOPS and TTPs (and corresponding training) are part of the system design, and will have to be identified, verified, and validated much earlier in the development process. This will pose organizational and personnel challenges to T&E, in addition to methodological challenges.
5. **Human Trust**
   In human-machine teaming (HMT) contexts, how the humans behave (and thus how well the team performs) depends in part on the humans' psychological attitudes toward the autonomous systems. "Trust" is the term generally used to describe those attitudes, though in practice those attitudes are generally more complex and nuanced than simply "how much do I trust it?". In order to design, debug, and improve HMT performance, T&E will need to be able to measure the various dimensions of Trust, to support understanding of how Trust affects team performance.
6. **Elevated Safety Concerns**
   Traditionally, T&E personnel have relied on the training and common sense of equipment operators to provide many kinds of safety assurance, both in the field and on the test range. Autonomous systems potentially take many of the decisions underlying routine safety out of the hands (and minds) of operators, and depend instead on complex software that allows the system to 'operate' itself. During Developmental Test and Evaluation, and on into Operational Test and Evaluation, it is likely that this software will still contain major bugs, and that the algorithms and training data being used might not be the final best choices. This creates a potential for various kinds of mischief – especially for weapon systems, highly-mobile systems, or other systems that could be dangerous in the hands of an unreliable operator.

Recurring Challenges (conclusion)

7. **Exploitable vulnerabilities**
   When systems operate themselves, they can be vulnerable to modes of attack – cyber, electronic, or physical – that would not be as much of a concern for a human-operated system. For example, a cyberattack that compromised the ability of an autonomous UAS to recognize other aircraft, or a physical proximity attack that repeatedly triggered the UAS's collision avoidance routine, might be much more effective than against a human-piloted aircraft. AI based on machine learning has its own set of potential vulnerabilities, both during training of the AI and in operation. T&E of autonomous systems will need to be aware of this expanded attack surface.
8. **Emergent behavior**
   DoD Directive 3000.09 specifically warns against the possibility of "unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semi-autonomous systems". Developing T&E methods to analyze the potential for emergent behavior in order to avoid it will be central to providing adequate verification and validation of autonomous systems.
9. **Post-fielding changes**
   Systems that employ unsupervised learning during operations will continue to change their behavior over time. This creates a need not only for periodic regression testing, but also for predictive models of how post-fielding learning might affect system (or team) behavior. Traditional Operational Test and Evaluation (OT&E) is concerned with the effectiveness and suitability of the system as it is today. Adding a requirement to be able to predict the effectiveness and suitability of the system it might become is a new challenge.
10. **Quality of inputs to machine learning**
    Machine Learning – especially supervised or reinforcement learning – depends critically on the data used to train the AI. Supervised learning data must not only be representative of the range and type of data the system will take as input during operations, but must also be correctly and completely labeled. This leads to a need for verification and validation of the data used to train the AI that is similar to the need for verification, validation, and accreditation (VV&A) of modeling and simulation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
xx-10-2019

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
Sept - Oct 2019

4. TITLE AND SUBTITLE
What Counts as Progress in the T&E of Autonomy?
(Conference Briefing)

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Tate, David, M.

5d. PROJECT NUMBER
AX-1-3542

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Document NS D-10922
H 19-000548

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Deputy Director, Policy Support
Engineering Policy & Systems
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
USD(R&E)/DDR&E(AC)/EP&S

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Organizations across the Department of Defense have recognized that autonomous systems will pose significant new test and evaluation challenges. IDA researchers have been exploring the implications of autonomy for instrumentation, facilities, measurement, and assurance case development. Dr. David Tate was invited by the Scientific Test and Analysis Techniques Center of Excellence to present at the upcoming Advances in the Test and Evaluation of Autonomy workshop. This is the briefing he presented. It lays out a framework for thinking about what kind of assurance cases for autonomous capabilities would be convincing to certification authorities, what kinds of arguments those assurance cases would use, what kinds of measurements or other evidence could support those arguments, and what kind of instrumentation or computational tools would be needed to collect that evidence. It also describes how that framework is currently being used as the basis for Defense Acquisition University online continuous learning module CLE002, Introduction to Test and Evaluation of Autonomous Systems.

15. SUBJECT TERMS
instrumentation; reasoning; assurance case; perception; autonomy; evidence, measurement; adversarial testing; planning; learning; modeling and simulation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tate, David, M.
Unclassified	Unclassified	Unclassified	Same as Report	35	**19b. TELEPHONE NUMBER** *(Include area code)* (703) 575-1409

Reset