INSTITUTE FOR DEFENSE ANALYSES

T&E of Cognitive EW: An Assurance Case Framework (Conference Presentation)

David M. Tate

June 2020
Approved for public release; distribution is unlimited.
IDA Document NS D-14269
Log: H 20-000256

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

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

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 a technical review of this document.

For More Information

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

David E. Hunter, Director, Cost Analysis and Research Division
dhunter@ida.org, (703) 575-4686

Copyright Notice

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-14269

T&E of Cognitive EW: An Assurance Case Framework (Conference Presentation)

David M. Tate

Executive Summary

There is widespread interest and concern within the Department of Defense (DoD) regarding the test, evaluation, verification, and validation (TEV&V) of military systems with autonomous capabilities. Autonomous systems will not be approved for fielding unless senior decision-makers are sufficiently confident in the systems' dependability (e.g., safety, security, reliability, and effectiveness). Commanders in the field must also 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 supporting the claim 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 TEV&V. In particular, we consider systems that feature one or more of these autonomous capabilities:

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

Cognitive Electronic Warfare (EW) relies on several of these autonomous capabilities. This briefing, developed for the 2020 inaugural workshop of the Cognitive Electronic Warfare Test and Evaluation Working Group, describes a framework for planning successful assurance cases for Cognitive EW systems. This framework includes the specification of appropriate assurance arguments, identification of key evidence needed to support those arguments, measurements that can produce that evidence, implied instrumentation needs, and resulting test infrastructure requirements. The briefing also discusses quantitative and analytical methods and tools to support these activities.

T&E of Cognitive EW:
An Assurance Case Framework

David Tate

Institute for Defense Analyses

June 2020

The goal is assured effectiveness and dependability

Advanced 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.

There may be multiple certification authorities (e.g., Safety, Cybersecurity, Effectiveness, Reliability).

State of the Art: Assurance Cases

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

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

- Safety cases (oldest, most mature literature)
- Software assurance cases (cybersecurity, reliability)
- Robustness cases

Currently, these cases are generally stovepiped.

Example: ISO/IEC 15026-2 (2011)

Systems and Software Engineering—
Systems and Software Assurance—
Part 2: Assurance Case

1 Scope

This part of ISO/IEC 15026 specifies minimum requirements for the structure and contents of an assurance case. An assurance case includes a top-level claim (or set of claims) for a property of a system or product, systematic argumentation regarding this claim, and the evidence and explicit assumptions that underlie this argumentation. Arguing through multiple levels of subordinate claims, this structured argumentation connects the top-level claim to the evidence and assumptions.

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 should 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

Autonomous capabilities (such as Cognitive EW) can break this model

We can't test exhaustively – the state space is too large.

We can't rely solely on DoE – we don't know the factors and can't assume smooth response everywhere.

Interactions between run-time monitors and core functions **add** complexity (and need additional testing).

Human-Machine Teaming (HMT) explodes both the state space and the set of potentially relevant factors.

Each aspect of dependability generates an "assurance attack surface"

If your system can be made unsafe, you lose…

If your system can be made unreliable, you lose…

If your system can be made to fail the mission, you lose…

If control of your system can be lost, you lose…

…regardless of whether it's the adversary, the environment, or your teammate that is doing it to you.

The technologies that enable Cognitive EW include:

Supervised learning (Perception)

Sensor fusion (Perception)

Knowledge representation (Reasoning)

Inference engines (Reasoning)

Reinforcement learning (Planning)

Expert systems (Planning)

Unsupervised learning (adaptive threat recognition)

HMT CONOPS

⋮

What are the assurance attack surfaces then?

They arise from the **inputs** to the enablers:

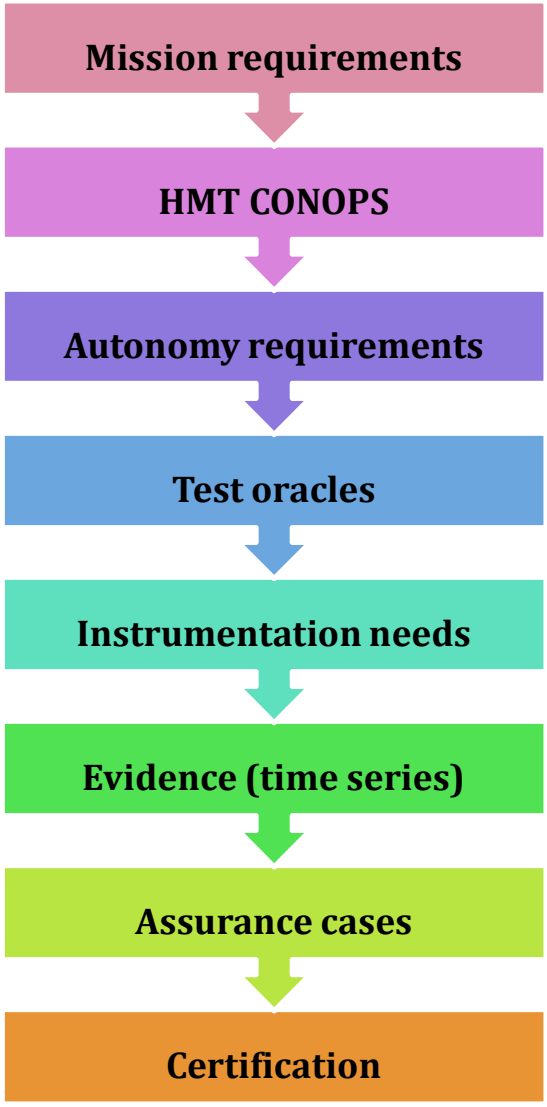
Perception: sensors, algorithms, stored data, training data
Reasoning: world model, ontology, data, algorithms
Planning: game model, reinforcement algorithm, rules
Machine learning: training data, models, architecture
Expert systems: ontology, world model, stored data
HMT: messages sent/received, world model, CONOPS

The attack surfaces drive T&E planning

Convincing assurance cases for Cognitive EW systems must address all of the cited attack surfaces.

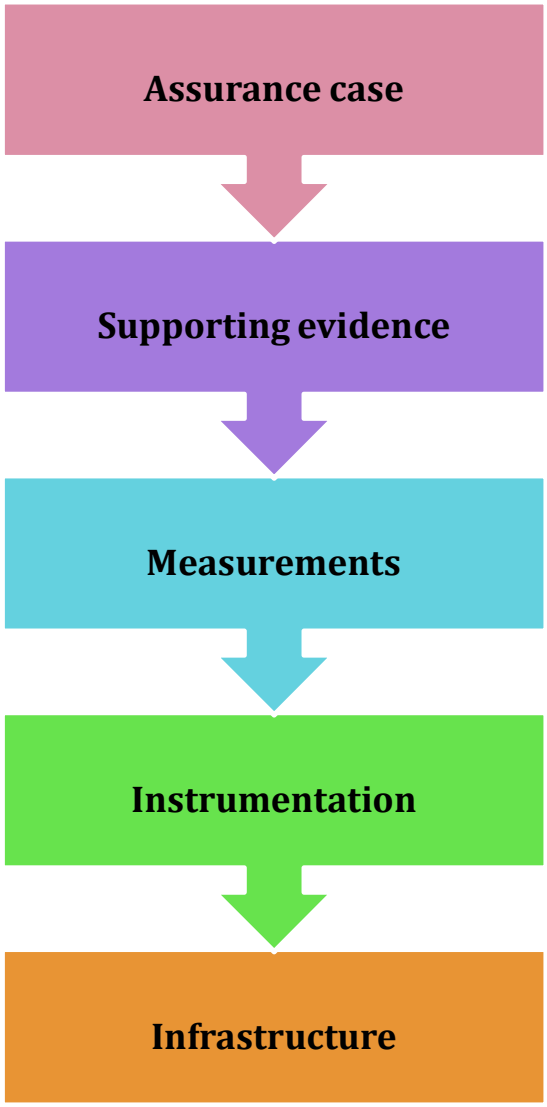
Happy side effect: The tools we need for this will also support diagnosis/debugging during development.

The assurance case life cycle



1. Make an **initial guess** at how the AI will team with the humans
2. Formulate **testable requirements** for autonomous functions
3. Codify **test oracles** for acceptable behaviors, including internal behaviors
4. Construct **assurance case outlines** – what arguments will convince? What evidence will they require?
5. Derive **instrumentation requirements** – what measurements will be needed to assess performance against the oracles and provide the needed evidence?
6. Collect evidence and **iterate**

Work backward to identify test infrastructure needs



1. Given the system assurance case, what evidence will be required?
2. What time series of measurements would produce that evidence?
3. What instrumentation is required to collect those measurements?
4. What infrastructure is needed to support that instrumentation?
    - Simulation testbeds?
    - Telemetry?
    - Training data?
    - Onboard recording?
    - Special R/F environment?

Some familiar tools still have their uses

Design of Experiments
(factor identification)

Observational Studies and Surveys
(especially for human-machine teaming)

Modeling and Simulation

What new tools might be useful?

Formal methods

Instrumenting cognition/explainable AI

Intelligent adversarial testing

Assurance case development tools

Examples: 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

Examples: Intelligent Adversarial Testing

The **Range Adversarial Planning Tool** (RAPT) developed at Johns Hopkins Applied Physics Laboratory automates adversarial testing of autonomous systems using simulations of the autonomy software and environment. RAPT builds a machine-learning model of the autonomy performance and then identifies regions of the configuration space with steep response gradients, indicating possible edge cases. RAPT then generates test designs that oversample the identified regions.

Similarly, the IBM **Adversarial Robustness Toolbox** (ART) supports verification of robustness and hardening for machine learning models.

Examples: Assurance Case Development Tools

Assurance Case Editor (**ACedit)**

Assurance Case Automation Toolset (**AdvoCATE**)

Evidence Confidence Assessor (**EviCA**)

Astah GSN (commercial product, see astah.net)

Each tool uses **Goal Structuring Notation** (GSN) as the graphical language for describing and manipulating arguments.

Reference: *Tool Support for Assurance Case Development*, Ewen Denney and Ganesh Pai, NASA Ames Research Laboratory

Bottom Line at the Bottom

Assurance cases for cognitive systems require more sophisticated arguments than “mere automation.”

Evidence to support those arguments requires the novel use of M&S and instrumentation inside the “black box.”

Tools exist to support the use of formal methods and automated development of assurance cases – but you have to use them from day 1, and take them seriously.

Backup

Case Study: Assurance Case Development

A Case Study in Assurance Case Development for Scientific Software

Mojdeh Sayari Nejad

MS (Computer Science) Thesis

McMaster University 2017

Assurance case for 3dfim+ software for analyzing functional MRI images of the brain

https://macsphere.mcmaster.ca/handle/11375/23075

Recurring Challenges for TEV&V of Autonomy

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 right kind 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.

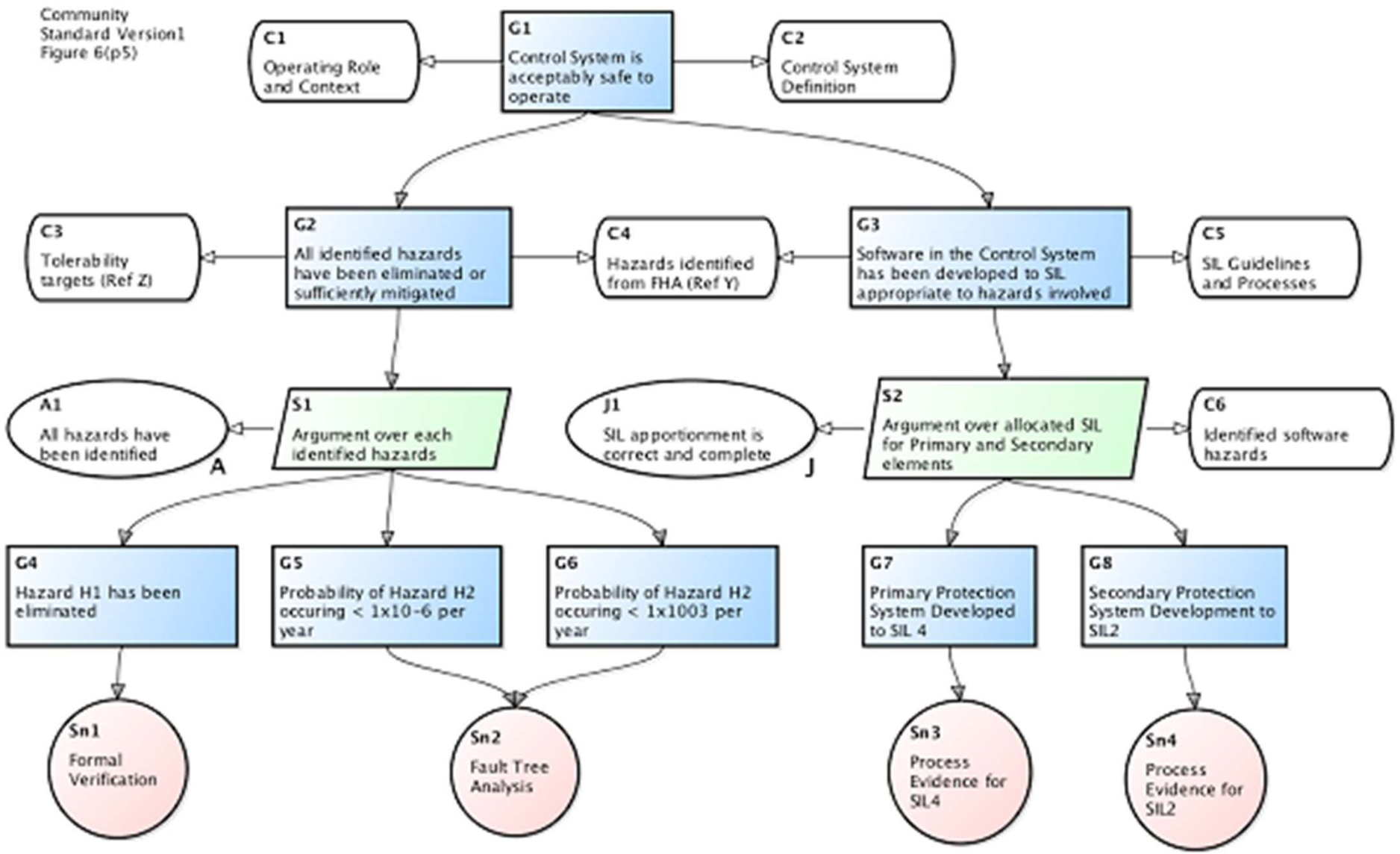
Recurring Challenges (continued)

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.

Example: Goal Structuring Notation (GSN)

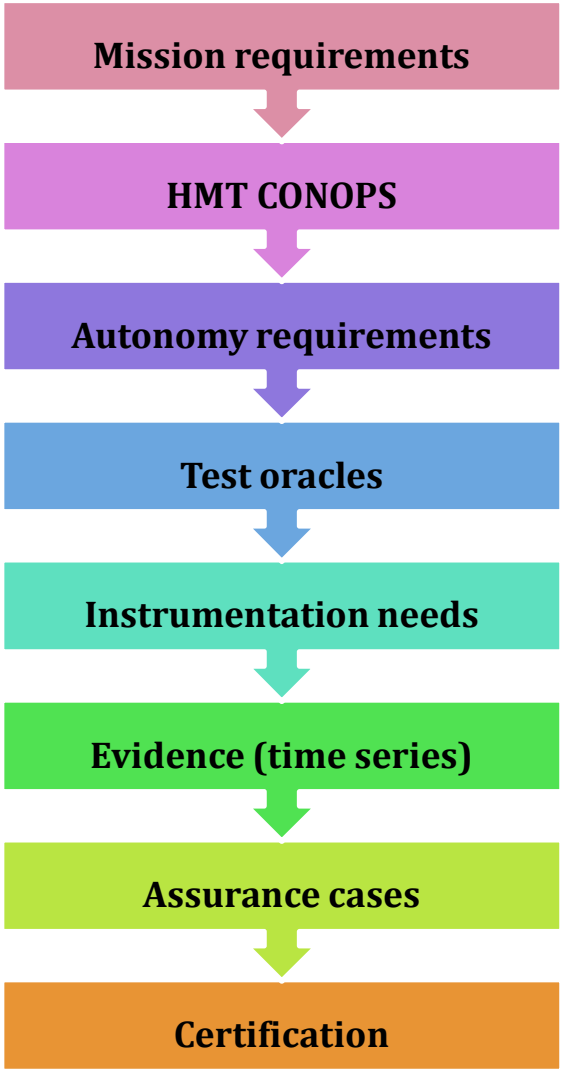


"Levels of autonomy" is a red herring

"It takes more sophisticated technology to keep the humans in the loop than it does to automate them out … On a commonly used scale of levels of autonomy, level one is fully manual control and level 10 is full autonomy … history and experience show that the most difficult, challenging and worthwhile problem is not full autonomy but the perfect five—a mix of human and machine and the optimal amount of automation to offer trusted, transparent collaboration, situated within human environments."

-- David Mindell, MIT

Autonomy TEV&V R&D Priorities



Specifying testable cognitive requirements

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

Formal verification methods

Instrumenting cognitive functions

- Aligned with oracles
- Support both assurance and trust

Virtual test environments

V&V of ML training and training data

Intelligent adversarial testing

Test oracle specification

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

Logic of assurance cases

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

Regression testing criteria

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
xx-06-2020	Final	May - June 2020

4. TITLE AND SUBTITLE
T&E of Cognitive EW: An Assurance Case Framework (Conference Presentation)

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

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

5d. PROJECT NUMBER
CARDPD

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-14269
H 20-0000256

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA

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

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

13. SUPPLEMENTARY NOTES
The material in this briefing consists solely of open source material and content previously approved for public release in IDA Document NS D-10523, "Attack Surfaces of Autonomy"; and IDA Document NS D-10922, "What Counts as Progress in the T&E of Autonomy?" an existing IDA publications that has been previously approval for public release.

14. ABSTRACT
Many international standards bodies have adopted the "assurance case" as a best practice for establishing the dependability (i.e., safety, security, reliability, effectiveness) of complex systems. This briefing provides a framework for test and evaluation (T&E) to support assurance cases for systems with autonomous capabilities, and outlines how that framework could apply to T&E of cognitive electronic warfare (EW) systems. It also discusses specific methods for developing evidence that could be used, and software tools for assurance case development and management.

15. SUBJECT TERMS
Test and evaluation; assurance case; autonomy; cognitive electronic warfare; dependability; perception; machine learning; test oracle; instrumentation

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	38	19b. TELEPHONE NUMBER *(Include area code)* (703) 575-1409

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18