INSTITUTE FOR DEFENSE ANALYSES

The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems

Brian A. Haugh, *Project Leader*
David A. Sparrow
David M. Tate

September 2018

IDA Paper
P-9292
Copy

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

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task AI-5-4458, "Artificial Intelligence Analyses," for Deputy Director, Information Systems and Cyber Technologies Office of the Under Secretary of Defense, Research and Engineering. 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

Daniel J. Porter, Joshua Alspector, Signe Redfield

For more information:

Brian A. Haugh, Project Leader
bhaugh@ida.org, 703-845-6678

Margaret E. Myers, Director, Information Technology and Systems Division
mmyers@ida.org, 703-578-2782

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-9292

The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems

Brian A. Haugh, *Project Leader*

David A. Sparrow

David M. Tate

Executive Summary

In support of a US/UK Technical Exchange Meeting on Artificial Intelligence & Autonomy Collaboration, the Institute for Defense Analyses (IDA) was asked to prepare a summary of academic work related to Test and Evaluation, Verification and Validation (TEV&V) of autonomous or artificial intelligence (AI) systems. This paper enumerates TEV&V challenges that have been identified by the commercial, academic, and government autonomy research communities, describes the focus of current academic research programs, and highlights areas where current research leaves unaddressed gaps in capability.

The academic research currently relevant to defense issues is concentrated in work on formal methods. For reasons ranging from development efficiency to operator trust, it will be essential to understand the inner workings of AI systems. "Explainable AI" has become a goal, and the US Defense Advanced Research Projects Agency (DARPA) has recently started a program by that name. "Cognitive instrumentation"—the ability to see into the inner working of the decision engines—will be essential, and work in this area is beginning as well. This instrumentation will also be needed to support run-time monitoring, which will now need to extend to monitoring the decision space as well as the physical envelope. Finally, adversarial testing is receiving increasing attention. This will be of special importance to the Department of Defense; rare but catastrophic failures are harder to avoid in this context than they are in commercial settings.

There is a brief discussion of the TEV&V tools that are being developed to facilitate testing of autonomous capabilities.

Contents

1. Introduction

Both the US Department of Defense and the UK Ministry of Defence have stated publicly that future defense capabilities are expected to depend heavily on autonomous systems—systems that make sophisticated judgments about the world, choose appropriate courses of action, and perhaps even adapt and learn over time. Developing and deploying such systems poses more than just a technical challenge in robotics and artificial intelligence (AI)—it also poses many challenges to the acquisition processes and workforces of the respective nations. From cost estimation to sustainment planning, every aspect of acquisition will be affected. In particular, test, evaluation, verification, and validation (TEV&V) of systems with autonomous capabilities may require not only novel methodologies and resources, but organizational and process changes as well.

This memorandum enumerates TEV&V challenges that have been identified by the commercial, academic, and government autonomy research communities; describes the focus of current academic research programs; and highlights areas where current research leaves unaddressed gaps in capability.

2. A Taxonomy of Challenges

Training materials currently in development by the US Department of Defense identify 10 recurring challenges for TEV&V that are specific to or exacerbated in autonomous systems.

1. **Instrumenting machine thinking**
   To diagnose the causes of incorrect behavior or inadequate performance, it will be necessary determine whether the problem lies in the Perception, Reasoning, or (course of action) Selection functions of the autonomous system—even after it has been established that the problem is not in the sensor hardware or signal processing. It will also be necessary to distinguish coding errors from inadequate algorithms or bad training 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 autonomous behaviors**
   In complex collaborative activities, it can be very difficult to determine what enables (or hinders) 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 understand how the system's various autonomous capabilities interact to enable (or hinder) mission execution. The requirements specification for autonomous behavior is often a problem here due to incomplete specifications based on an analogy with human behavior.

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 Selection. 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. TEV&V will necessarily include prototyping and experimentation to determine

what kind of internal model, using what kind of representation, is needed to achieve both performance and dependability.

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(s) will use that system and what the user interface should look like, and then work out the details of the Concept of Operations (CONOPS), tactics for effective employment, and training of future operators long after the basic design has been decided. For autonomous systems, where the system operates itself to some extent and interacts autonomously with humans, the details of CONOPS and tactics (and corresponding training) are part of the system design, at least on the machine side (and probably on the human side as well), 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, as well as 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 nuanced and multi-dimensional than simply asking "how much do I trust it?" In order to design, debug, and assure performance, TEV&V 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 and asymmetric hazard**
   Traditionally, TEV&V 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 into Operational Test and Evaluation, it is likely that the software will still contain major bugs and that the algorithms and training data being used might not be the final 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.
   Similarly, the most striking successes of AI, machine learning, and autonomous systems to date have occurred in contexts where the cost of error is low. For autonomous military systems, this will not generally be the case—designating incorrect targets, responding to spurious cyber attacks, misidentifying

individuals or objects, crashing unmanned vehicles into obstacles, and other potential failure modes of autonomous systems are all potentially very costly. Apart from driverless car efforts, there are few active research programs today for applications with highly asymmetric hazard functions.

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 unmanned aircraft system (UAS) to recognize other aircraft or physical proximity illusions that repeatedly triggered the UAS's collision avoidance subroutine might be much more effective than they would be 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. TEV&V of autonomous systems will need to be aware of this expanded attack surface.

8. **Emergent behavior**
   US 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 assured dependability for autonomous systems.

9. **Post-fielding changes**
   Systems that employ unsupervised learning or other adaptive control algorithms 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 TEV&V is concerned with the effectiveness and suitability of the system as it is today. Needing to be able to predict the effectiveness and suitability of the system as it might become is a new challenge.

10. **Verification and validation of training data**
    Machine learning—especially supervised or reinforcement learning—depends critically on the data used to train the AI. It is an axiom of computer scientists working in this area that "the intelligence is in the data, not the algorithm." 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, validation, and accreditation (VV&A) of the data used to train the AI that is similar to the need for VV&A of modeling and simulation.

3. Potential Approaches to Overcome These Challenges

Academic and government researchers have begun work on a number of techniques and methodologies that might help to overcome these core challenges. Current efforts fall into four primary categories: *formal methods*, *cognitive instrumentation*, *adversarial testing*, and *run-time monitoring*. We discuss each of these below, then consider the question of how well they cover the full set of capability gaps implied by the challenge list.

A. Formal Methods

1. Summary

Formal methods in software development allow developers to specify certain properties that the software should have, produce the software, and verify that it does have those properties without needing to confirm that empirically by testing for them. Properties to be specified might be things like:

> Property 1: the weapon cannot fire while turret is still rotating
>
> Property 2: the course-of-action selector can never get into an infinite loop

There are two approaches to formal methods for autonomous systems verification: (1) formal methods can be used after the fact as an analytic tool to verify some properties of existing software, or more importantly, (2) formal methods can be used as a design and development process that can assure much more about the behavior of the software to be developed.

Formal methods of the second kind are most commonly used in the development of complex safety-critical or security-critical systems or for expensive one-time development efforts (e.g., deep space probes). Applying formal methods to complex AI and autonomous systems is a natural extension of this.

2. Limitation

Although formal methods can be extremely useful, there are significant constraints on the current state of the art. These include the following:

> **Scalability**: There are currently fairly tight bounds on the size of development effort (or state space) that the techniques can be applied to.

Scope: Not all desired properties can be assured through formal methods, and there may be performance trades associated with achieving assurance.

Rigidity: Any change to a system developed using formal methods risks invalidating the assurance proofs, unless the formal methods are reapplied to the new specification.

Given that not all desired behaviors can be assured using formal methods, there are also open research questions concerning how to combine formal methods with empirical TEV&V techniques or run-time monitoring strategies.

B. Cognitive Instrumentation and Explainable AI

1. Summary

The key distinguishing feature of autonomous systems is that they make decisions, when interpreting their environments and selecting courses of action. Test and evaluation of autonomy will depend critically on the ability to assess the quality of this decision-making capability. In general, observed high-level system performance in limited test scenarios will not be sufficient to validate and verify autonomous decision-making in the ways necessary for successful development and deployment, especially for systems designed to team with humans. Instead, novel instrumentation approaches will be required to diagnose and characterize:

- Adequacy of architectures and algorithms (including machine learning).
- Adequacy and appropriateness of training data.
- Effectiveness of operational concepts for HMT.

2. Why Explanations Are Needed

Explanations of system behavior support at least four distinct goals for military systems:

1. **Diagnosis**: Knowing why the system is exhibiting undesired behavior is the first step toward fixing it.
2. **Prediction**: Being able to forecast how the system will behave in given circumstances is essential to effective employment of the system.
3. **Bounding**: Understanding the limits of dependable performance allows formulation of tactics/techniques/procedures for how the system is to be used and identification of where run-time monitoring of system state may be the only way to avoid undesirable behaviors.

4. **Trust**: If humans are teaming with autonomous machines, the overall performance of the human-machine system may depend on how well the humans feel they understand the reasons behind the machine's behaviors.

For traditional software employing procedural algorithms, the logic of the algorithm traditionally serves as the explanation. Describing the flow of control (e.g., using pseudocode or flowcharts) at various levels of description could "explain" the behavior of the system. As systems have grown in complexity, the ability of humans to understand the logic of the systems in a way that counts as an "explanation" has eroded.

Machine learning breaks this paradigm completely by replacing comprehensible procedural logic with a trained ability to generate outputs from inputs in a manner that looks much more like an intuition or hunch than like reasoning. The relationship between input and output is fundamentally opaque; there is no procedural logic to be traced. Producing surrogate procedural descriptions that summarize how the system is reaching its conclusions in ways that can function as explanations for purposes of diagnosis, prediction, bounding, and trust requires additional work and access to the internal states (and possibly the training algorithm) of the machine-learning module. (Note that different explanatory frameworks will generally be needed for each of these goals—the explanation appropriate for diagnosis may be very different from the explanation required for bounding or trust.)

The US Defense Advanced Research Projects Agency (DARPA) has recently launched a funded research program in "Explainable Artificial Intelligence" (XAI). The four goals mentioned above (diagnosis, prediction, bounding, and trust) are all represented in the Explainable AI program. "Explainability" in the XAI program involves multiple measures of effectiveness, depending on which of the goals is being pursued. These measures can include human satisfaction with the explanation, how helpful the explanation is in choosing how to interact with the autonomous system, how well the explanation supports diagnosis of unexpected behaviors, or how well the explanation predicts future system behavior.

3. Cognitive Instrumentation

The phrase *cognitive instrumentation* refers to any measurement of internal system states of the software modules that provide the autonomous capabilities of a system. There is an obvious analogy with physical measurement of internal system state (e.g., temperature, pressure, voltage, torque) during development, test, and evaluation of hardware systems. As with those physical analogues, the measurements may be only for TEV&V purposes, or they may be incorporated into the design of the system as run-time monitors on behaviors that cannot be adequately assured through other means.

One area of explainability research involves designing autonomous systems to self-report their reasons for their behaviors. This is a promising approach, but it does not avoid

the need for instrumenting the internal states of the system. At a minimum, verification and validation of the self-reporting system will involve the same diagnosis, prediction, and trust issues as the mission behavior of the system, with an associated need to look inside the black box and make sense of what is happening there. At some point, all system development requires knowing the truth about what is happening inside the system.

In the special case of HMT, where the machines have significant autonomous capabilities, cognitive instrumentation will include measurements of both the machines and the humans, as well as the interface between them. Measurement and characterization of trust by humans is already a rich area of academic research. Measurement and characterization of machine understanding of human intent is also receiving increasing attention. There are also active research programs exploring the design of HMT protocols. Some of these focus on the optimal allocation of work between humans and machines in various contexts. Others are concerned with optimal communications protocols—under what circumstances should humans or machines volunteer information (and which information), make suggestions, or ask for guidance? Without cognitive instrumentation, those optimization programs could easily devolve into guesswork and trial-and-error.

Cognitive instrumentation is a necessary condition for Explainable AI; any valid explanation of how the AI is thinking or why it is behaving a certain way must be based on accurate measurements of its internal states. However, the measurements themselves are not the explanation. A complete state description will generally be so complex that it is not understandable, whereas too small a subset of measurements will generally not be sufficient to support explanation. Additional effort will be required to identify a minimal sufficient set of measurements that can be used as inputs to explanatory models that can be understood in human terms for diagnosis, prediction, bounding, and developing appropriate operator trust. There are many tricky steps between "the weights in layer 17 of the neural network are as follows" to "the system seems to be recognizing cats by their ears and coloration." As noted above, different explanatory models will be needed for the different goals—an explanatory model supporting useful diagnosis for developers may be much more complex (and unwieldy) than a model that produces explanations of mission behavior for operators or commanders.

C. Adversarial Testing

As noted in Challenge 6, many defense applications will require very low probabilities of high-cost events occurring. For systems with extremely large and highly nonlinear state spaces, it will not be possible to provide that assurance statistically using traditional Design of Experiments techniques. Adversarial testing, especially when combined with machine learning and automated test design methods, provides a potentially more efficient means to identify and eliminate potential failure modes. In this approach, analytical techniques are used to identify test scenarios in which the system is most likely

to perform unacceptably and focuses testing in those portions of the state space to maximize diagnostic information and inform robust design.

As a specific example, one could apply reinforcement learning to the selection of environmental factors as input to the test cases in a modeling and simulation driven test bed that included the AI software running the autonomous systems. The environment could be treated as a "thinking adversary" in an asymmetric game. The environment would learn optimal strategies to defeat the autonomous system, even as the autonomous system learned improved strategies to counter the environment. This would both maximize the chance of finding key vulnerabilities and weaknesses and help discover ways to mitigate or avoid those vulnerabilities.

The vulnerabilities need not be isolated within the autonomous system itself—they might also arise from the proposed CONOPS, especially in the case of significant HMT. They could also be associated with bias or incompleteness in the training data used in supervised learning to develop the autonomous capabilities. Cognitive instrumentation would be needed to distinguish inadequate algorithms or models from inadequate CONOPS or flawed training.

Finally, we note that this approach requires domain expertise as well. The operating environment and mission must be well enough understood so that the apparent "weaknesses" are not associated with impossible circumstances. If the weakness occurs only during heavy rainfall at temperatures below –40° C, it will be a weakness we can live with. More subtly, the domain expertise will need to be sufficient to estimate probability of occurrence of the conditions that exposed the weakness. As mentioned before, utilization in the defense sector will require attention to the likelihood of seriously adverse outcomes.

D. Run-Time Monitoring

Given the difficulty of assuring that a system will not exhibit specific undesired behaviors, a natural thought is to instead monitor the system during operations and intervene when bad behavior is imminent. This approach is already common in engineering practice for safety-critical systems. It is mentioned explicitly in the recent US *Unmanned Systems Integrated Roadmap 2017-2042*:

> For the most demanding adaptive and non-deterministic systems, a new approach to traditional TEVV will be needed. For these types of highly complex autonomous systems, an alternate method leveraging a run-time architecture that can constrain the system to a set of allowable, predictable, and recoverable behaviors should be integrated early into the development process. Emergent behaviors from large-scale deployment of interacting autonomous systems poses a difficult challenge. The analysis and test burden would thereby, be shifted to a simpler, more deterministic run-time assurance mechanism. The effort for new approaches to TEVV endeavors to provide a structured argument, supported by evidence, justifying that a

> system is acceptably safe and secure not only through offline tests, but also through reliance on real-time monitoring, prediction, and fail-safe recovery.[1]

Although this mechanism might indeed be simpler than avoiding unpredictable behaviors in the first place, it is not without its own challenges. In general, any behavior whose dependability cannot be adequately assured through system design and training would need to be monitored, with a robust intervention standing by. This means not only intervening when the system is about to execute some undesired physical action (e.g., one that might risk harm to the system or to humans), but also intervening in any case where the system is making a bad decision or misinterpreting its environment. Detecting such cases and handling them gracefully will not always be easy. Research is required into architectures to support this concept, instrumentation needs and control algorithms to predict and avoid specific failure modes, systematic identification of conditions to be monitored for, robustness against attacks designed to invoke fail-safe behaviors, and so forth. It goes without saying that the fail-safe systems would themselves need to be verified and validated as well.

[1] *Unmanned Systems Integrated Roadmap 2017-2042*, US Department of Defense, 28 August 2018, p. 10

4. Resources and Tool Development

The exploding state space requires far more efficient and more intrusive testing of machine decision making than is currently the norm in defense software testing. The goal of the Science and Technology (S&T) programs will ultimately be to enable the development of tools. Some prototype tools have already been developed, such as the Range Adversarial Planning Tool (RAPT), a system that embeds the autonomous software in a virtual environment, then uses advanced optimization techniques to identify portions of the state space that exhibit either poor performance or high sensitivity to inputs. The outputs are then used to prioritize physical tests of the actual system; the results of those tests are used to update and improve the simulation models and to inform design changes.

Though RAPT is an example of an adversarial testing approach, we anticipate that many tools will be hybrid applications. For example, RAPT could be augmented with cognitive instrumentation, which would help to identify not only where in the performance space there are shortfalls, but why. Some work is underway using formal methods to design a base decision engine. This can be used as a starting point for performance optimization with machine-learning techniques. This in turn could be combined with run-time monitoring (including monitoring of the decision algorithms via cognitive instrumentation) to trigger a fail-safe recovery when things go awry. Alternatively, formal methods could be used to reduce the dimensionality of the test space, leading to more efficient coverage in either Modelling and Simulation (M&S) or open air testing.

Beyond specific tools, facilities able to mix live, virtual, and constructive (LVC) elements in the TEV&V of the AI elements driving autonomous systems will be essential elements in both design and TEV&V. The use of modeling and simulation in support of these efforts will have critical differences from the traditional uses of M&S. In particular, the M&S will have to support rapid exploration of a decision space, rather than high resolution modeling of a physical space. TEV&V of machines intended for active teaming with humans will probably require the creation of simple models of both the machine and the human.

As tools focused on TEV&V of autonomous systems and their AI drivers develop, and LVC facilities to support autonomous system development are built, S&T efforts that support their development or enable extension or integration of existing tools will be of particular value.

5. Data

The most impressive recent advances in AI have been driven by extremely data-intensive machine learning techniques, especially those involving neural network architectures. The increasing dominance of these techniques is such that "the intelligence is in the data, not the code" has become an axiom for many researchers. Although this is an oversimplification—the neural network architecture also matters—it is a reasonable characterization of the current philosophy of machine learning.

If the intelligence is in the data, then so too are the bugs. At best, the weights in a trained neural network accurately summarize the data used to train them. If the training data set is not perfectly representative of what actual operational data will look like, over- or under-represents certain cases, or contains unnoticed spurious correlations, those flaws will be perpetuated (and perhaps magnified) in the outputs of the trained system. Examples of this abound in the literature, including visual imagery analysis systems that "learned" to distinguish US from Russian military equipment based on whether the image was brightly lit or not, and job search systems that "learned" to recommend lower-paying jobs to women.

This suggests that verifying and validating the data used to train an AI system will be a critical part of TEV&V of the overall system. There is a useful analogy here with VV&A of modeling and simulation used during system development and testing. There are also important unresolved questions involving data labeling, reuse of training data across applications, incremental or supplemental training (as opposed to retraining from a clean slate) when additional data become available, production of synthetic training data, adversarial data, data security, proprietary rights to training data, and many other related issues. Research in these areas is just beginning; it is not clear that all of the important questions have yet been identified.

6. Summary

AI is widely expected to transform the defense enterprise. Defense will need to meet the challenge of how it can be utilized and accepted into service across the broad terrain of the defense enterprise. Defense-specific fundamental research is needed to underpin this utilization and acceptance. Although many techniques can be borrowed from the commercial sector, there are key recurring differences in the nature of the applications.

One of these is the basic difference between the costs of error in commercial and defense applications—particularly for modern weapons systems. For autonomous systems, especially those with advanced AI capabilities, the extensive decision space to be explored, coupled with the potentially catastrophic consequences of error, will naturally lead to some resistance to adoption. If TEV&V is not to be one of the barriers to AI entry into service, new and novel approaches to TEV&V will be needed.

A second major difference is in the nature of the data available in commercial and military applications. Because of the potential for serious consequences, training data for military AI systems will need a level of TEV&V much more stringent than in commercial settings where the costs of error are small. This represents a new or expanded role for TEV&V, similar to current VV&A of modeling and simulation.

Despite these important differences, it is prudent, if not essential, to capitalize on the academic work to date to the maximum extent possible. We have discussed several current academic research thrusts relevant to the challenges of TEV&V of autonomous systems:

- Formal methods for ensuring correctness and proving properties of software where possible.
- Cognitive instrumentation to diagnose flaws in Perception/Reasoning/Selection, evaluate the effectiveness of CONOPS, and provide evidence for assurance cases.
- Explainable AI, using information from cognitive instrumentation to establish assurance (among responsible authorities) and calibrated trust (among operators/teammates) of developed AI/machine learning/autonomous systems.
- Adversarial testing of decision engines to rapidly identify vulnerabilities and weaknesses.
- Run-time monitoring of key outcomes to be avoided.

These will prove essential in making safety and dependability assurance cases to users and certifiers. The possibility of seriously adverse outcomes requires an element of testing that preferentially seeks adverse outcomes and is able to assess the associated risks: probability and consequence. It is likely that successful fielding of autonomous systems will frequently involve deliberately trading away some mission capability in favor of higher robustness against worst-case outcomes and higher understandability of machine behaviors by humans.

Science and technology investment in these areas will be important. It will be of special value when it can lead to tools that either make TEV&V more efficient, or make critical information more available. This will address the challenges of the state space explosion and of the need to ensure that highly adverse outcomes will be sufficiently unlikely. In addition, ensuring the effectiveness and suitability of training data is more challenging and potentially much more important for defense applications than for commercial ones. A new science of training data management, involving not just TEV&V in specific applications but also reuse, configuration management, and synthetic data generation, may be required.

Selected References

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bass, E. J., and A. R. Pritchett. 2008. "Human-Automation Judge Learning: A Methodology for Examining Human Interaction with Information Analysis Automation. *IEEE Transactions on Systems, Man and Cybernetics A: Systems and Humans* 38:759-776.

Callow, Glenn. 2013. "Extending relational model transformations to better support the verification of increasingly autonomous systems". Loughborough University Institutional Repository. https://dspace.lboro.ac.uk/2134/13435

Christoffersen, K., and D. D. Woods. 2002. "How to Make Automated Systems Team Players." In *Advances in Human Performance and Cognitive Engineering Research* 2:1–12, edited by E. Salas. JAI Press, Kidlington, U. K.

Clarke, Edmund M. William Klieber, Miloˇs Nov´aˇcek, and Paolo Zuliani. 2012. "Model Checking and the State Explosion Problem." In *Tools for Practical Software Verification*, edited by Bertrand Meyer and Martin Nordio, 1–30. Berlin, Heidelberg: Springer-Verlag.

Daftry, Shreyansh, Sam Zeng, J. Andrew Bagnell, and Martial Hebert. 2016. *Introspective Perception: Learning to Predict Failures in Vision Systems*. arXiv:1607.08665v1 [cs.RO]

de Niz, Dio. 2017. "Certifiable Distributed Runtime Assurance." *Research Review 2017*. Carnegie Mellon University, Software Engineering Institute.

Department of Defense. 2013. "Autonomy Research Pilot Initiative Web Feature." Last modified June 14. https://www.acq.osd.mil/chieftechnologist/arpi.html.

Drouilly, Romain, Patrick Rives, and Benoit Morisset. 2015. "Semantic Representation for Navigation in Large-Scale Environments." In *2015 IEEE International Conference on Robotics and Automation* (*ICRA*), 1106–11.

Dua, Sumeet, and Xian Du. 2016. *Data Mining and Machine Learning in Cybersecurity*. New Boca Raton, FL: CRC Press.

GAO. February 2015. *High Risk Series: An Update*. GAO-15-290, Report to Congressional Committees. Washington, DC: Government Accountability Office.

Goix, Nicolas. 2016. "How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?" arXiv preprint arXiv:1607.01152. Presented at ICML2016 Anomaly Detection Workshop, New York.

Gunning, David. n.d. “Explainable Artificial Intelligence (xai).” Defense Advanced Research Projects Agency (DARPA). Accessed 2017.

Herlocker, J., J. A. Konstan, L. G. Terveen, and J. T. Riedl. 2004. “Evaluating Collaborative Filtering Recommender Systems.” *ACM Transactions on Information Systems* 22:5–53.

Higgins, Tim. May 14, 2018. “Tesla Considered Adding Eye Tracking and Steering-Wheel Sensors to Autopilot System.” *Wall Street Journal*. Accessed May 21 2018. https://www.wsj.com/articles/tesla-considered-adding-eye-tracking-and-steering-wheel-sensors-to-autopilot-system-1526302921.

IBM Research Ireland. https://github.com/IBM/adversarial-robustness-toolbox

Ilachinski, Andrew. 2017. *AI, Robots, and Swarms*. CNA Report DRM-2017-U-014796, January. https://www.cna.org/cna_files/pdf/DRM-2017-U-014796-Final.pdf

Kress-Gazit, Hadas, Morteza Lahinanian and Vasumathi Raman. “Synthesis for Robots: Guarantees and Feedback for Robot Behavior.” *Annual Review of Control, Robotics and Autonomous Systems*, 1:211-236.

Lee, I., S. Kannan, M. Kim, O. Sokolsky, and M. Viswanathan. "Runtime Assurance Based On Formal Specifications." (1999). *1999 International Conference on Parallel and Distributed Processing Techniques and Applications* PDPTA99, Volume 1, pp. 279-287. http://www.informatik.uni-trier.de/~ley/db/conf/pdpta/pdpta1999-1.html

Luckcuck, Matt, Marie Farrell, Louise Dennis, Clare Dixon, and Michael Fisher. 2018 “Formal Specification and Verification of Autonomous Robotic Systems: A Survey.” (Submitted on 29 Jun 2018.) https://arxiv.org/abs/1807.00048

Madhavan, P, and D. A. Wiegmann. 2007. “Similarities and Differences between Human-human and Human-Automation Trust: An Integrative Review.” *Theoretical Issues in Ergonomics Science* 8:277–301.

Monavon, Gregoire, Wojciech Samek, and Klaus-Robert Muller “Methods for Interpreting and Understanding Deep Neural Networks.” arXiv:1706.07979v1 [cs.LG] 24 Jun 2017

Morales, Javier, Maite Lopez-Sanchez, Juan A. Rodriguez Aguilar, Wamberto Vasconcelos, and Michael Wooldridge. February 2015. “Automated Synthesis of Compact Normative Systems.” *ACM Transactions on Autonomous and Adaptive Systems* 10, no. 1, Article 2. DOI:http://dx.doi.org/10.1145/0000000.0000000

Mullins, Galen E., Paul G. Stankiewicz, R. Chad Hawthorne, and Satyandra K. Gupta. 2018. “Adaptive Generation of Challenging Scenarios for Testing and Evaluation of Autonomous Vehicles.” *Journal of Systems and Software* 137 (March): 197–215. https://doi.org/10.1016/j.jss.2017.10.031.

Nixon, Mark S., and Alberto S. Aguado. 2012. *Feature Extraction & Image Processing for Computer Vision*. London and Oxford, UK: Academic Press.

North Atlantic Treaty Organization Science and Technology Organization website. n.d. Accessed May 11, 2018. https://www.sto.nato.int/pages/systems-concepts-and-integration-ft3.aspx.

Parker, Lynne. Private communications.

Redfield, Signe. Private communications.

Roske, Jr., Vincent P. 2016. "Perspectives on the Test and Evaluation of Autonomous Systems" (IDA document D-5733). Alexandria, VA: Institute for Defense Analyses.

Rouf, Christopher et al. March 2017. "Formal Methods for Verification and Testing of Autonomous Systems". Presented at NDIA Test and Evaluation Conference.

Saffiotti, Alessandro. 1997. "The Uses of Fuzzy Logic in Autonomous Robot Navigation." *Soft Computing* 1, no. 4: 180–97.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2017. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." arXiv preprint arXiv:1708.08296.

Sarter, N. B., D. D. Woods, and C. Billings. 1997. "Automation Surprises." In *Handbook of Human Factors and Ergonomics* 2:19–35, edited by G. Salvendy. New York: Wiley.

Scheidt, David, et al. "Safe Testing of Autonomous Systems Performance*" I/ITSEC 2015.*

Schmidt, Aurora, et al. July 2016. "Complementary Formal Techniques for Verification and Validation of Complex Autonomous Systems" Presented at Safe and Secure Systems and Software Symposium (S5).

Shepardson, David. May 7, 2018. Uber Sets Safety Review; Media Report Says Software Cited in Fatal Crash. *Reuters*. Accessed May 21, 2018. https://www.reuters.com/article/us-uber-selfdriving/uber-hires-former-ntsb-chair-to-advise-on-safety-culture-after-fatal-crash-idUSKBN1I81Z4.

Szegedy et al. 2013. Google. https://arxiv.org/abs/1312.

Tak´acs, Arpad, et al., "Automotive Safety in the Development Pipeline of Highly Automated Vehicles." Submitted to *IEEE SMC*.

Tate, David M., Rebecca A. Grier, Christopher A. Martin, Franklin L. Moses, and David A. Sparrow. 2016. "A Framework for Evidence-Based Licensure of Adaptive Autonomous Systems" (IDA Paper P-5325). Alexandria, VA: Institute for Defense Analyses,

Westin, C., C. Borst, and B. Hillburn. 2016. "Automation Transparency and Personalized Decision Support: Air Traffic Controller Interaction with a Resolution Advisory System." *Proceedings of the International Federation of Automatic Control* 49:201–6.

Wright, J. L, J. Y. C. Chen, M. J. Barnes, and P. A. Hancock. 2017. "Agent Reasoning Transparency: The Influence of Information Level on Automation-Induced

Complacency." ARL-TR-8044. Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Young, Reed. Private communications.

Zantedeschi, Valentina, Maria-Irina Nicolae, and Ambrish Rawat. 2017. "Efficient Defenses Against Adversarial Attacks" (Submitted on 21 Jul 2017 (v1), last revised 30 Aug 2017 (this version, v2)). https://arxiv.org/pdf/1707.06728.pdf

Zhang, Xuezhou, Xiaojin Zhu, and Stephen Wright. 2018. "Training Set Debugging Using Trusted Items." In *The Thirty-Second AAAI Conference on Artificial Intelligence* (*AAAI*).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
14-09-18	Final	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems	HQ0034-14-D-0001
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBERS

6. AUTHOR(S)	5d. PROJECT NUMBER
David A. Sparrow, David M. Tate	AI-5-4458
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES	8. PERFORMING ORGANIZATION REPORT NUMBER
Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882	P-9292

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR'S / MONITOR'S ACRONYM
Richard W. Linderman Deputy Director, Information Systems and Cyber Technologies Office of the Under Secretary of Defense, Research and Engineering 6000 Defense Pentagon, Washington, DC 20301	ISCT USD/R&E
	11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Brian A. Haugh

14. ABSTRACT

In support of a US/UK Technical Exchange Meeting on Artificial Intelligence & Autonomy Collaboration, IDA was asked to prepare a summary of academic work related to Test and Evaluation, Verification and Validation (TEV&V) of autonomous or AI systems. This paper enumerates TEV&V challenges that have been identified by the commercial, academic, and government autonomy research communities, describes the focus of current academic research programs, and highlights areas where current research leaves unaddressed gaps in capability.

15. SUBJECT TERMS

Autonomy, Artificial Intelligence, Test and Evaluation, Verification and Validation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Richard W. Linderman
Unclassified	Unclassified	Unclassified	Unlimited	26	19b. TELEPHONE NUMBER (Include Area Code) 571-372-6734