INSTITUTE FOR DEFENSE ANALYSES

Assessing the Quality of Decision-making by Autonomous Systems

David A. Sparrow
David M. Tate
John C. Biddle
Nicholas J. Kaminski
Poornima Madhavan

June 2018
IDA Paper P-9116

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305-3086

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

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-14-D-0001, project AX-2-4383, "Test and Evaluation of Autonomous Systems," for the Office of the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(T&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

The authors would like to thank Andrew X. Richardson and Jeffrey A. Snyder of the IDA Systems and Analyses Center for their technical review of this report.

For More Information

David A. Sparrow, Project Leader
dsparrow@ida.org, 703-578-2992

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-9116

Assessing the Quality of Decision-making by Autonomous Systems

David A. Sparrow
David M. Tate
John C. Biddle
Nicholas J. Kaminski
Poornima Madhavan

Executive Summary

IDA was tasked to develop methodologies for assessing the quality of decision-making by autonomous systems as part of the test and evaluation, verification and validation (TEV&V) process for autonomous systems. This paper identifies the key challenges and necessary innovations associated with testing and evaluating the quality of decision-making by machines with autonomous capabilities, either on their own or when operating in human-machine teams.

Modern computer processing power and inexpensive memory have enabled new algorithms and computational approaches that led to a state-space explosion in autonomous systems. The number of possible combinations of current inputs, memory contents, mission contexts, and possible courses of action is so large and varied that it is impossible to ever test even a representative sample of these combinations. Data from multiple modalities are processed to build and maintain a dynamic world model that describes both the external environment and the internal state of the machine. Systems extrapolate possible courses of action, rather than choosing from a predefined list, then prioritize and select an action while accounting for current conditions, other entities, commander's intent, and system status. The resulting state space is not merely too large to test exhaustively—it is also too complex and nonlinear for traditional statistical methods (e.g., design of experiments) to be able to guarantee adequate "coverage" of the state space. Autonomous systems that share responsibilities (perception, reasoning, and course of action selection) with humans show even faster state-space explosion and present additional development challenges arising from the need to include the human-machine teaming concept of operations (CONOPS) as part of the system design. For such systems, much more experimentation and discovery will be needed during development than is common (or desirable) in a traditional program.

Autonomous capabilities may also involve greater potential safety and vulnerability hazards than traditional systems. Shifting some cognitive tasks from humans to machines increases the "failure surface" of the system in ways that are not yet well-understood. For instance:

1. Failure modes traditionally avoided through reliance on human training and common sense will require explicit assurance through test and evaluation (T&E), possibly requiring new T&E methods and resources.

2. Autonomous machines that share both design and training are likely to exhibit identical failure modes, raising the possibility of coherent failures (possibly simultaneous) across an operating force.
3. Machine decision-making, in certain circumstances, may prove too rapid or too remote for adequate human oversight.

Not all autonomous capabilities, such as some standard practices for testing systems that make decisions by simple thresholding or that fuse information from a small number of sensors, will pose new challenges to TEV&V. What characterizes these systems, and makes traditional testing effective, is the low dimensionality of the state space. Commercial automobiles already feature autonomous capabilities of this kind, such as anti-lock braking systems, smart cruise control, and hands-free parallel parking. In the future, as more sensors are added and the algorithms integrate information from more sensors and control more of the car's actions (e.g., steering, acceleration, braking), the dimensionality of the state space will increase until at some point exhaustive testing will no longer be feasible.

One key to addressing the state-space explosion will be the ability to search the state-space efficiently for problematic areas. This may not be sufficient, but work has begun in this area. As noted above, statistical sampling of the environment/mission set will not work, because the relevant sampling distributions are likely to be unknown for autonomous systems—even in the absence of an adversary. System decision-making behavior (perception, reasoning, and selection) can in principle be diagnosed and explained—but this will require novel instrumentation of the internal cognitive processes and new characterizations of what proper cognition looks like from the inside.

Modeling and simulation (M&S), and live, virtual and constructive testbeds will be essential tools for the analysis and assessment of autonomous capabilities. These tools will need to explore the decision envelope of autonomous capabilities, rather than the physical envelope with which we are familiar. Further, there will need to be transparency of decision-making in both the human and the machine elements of the human-machine team. This transparency will require specialized instrumentation and probably extensive data logging. The tools and the associated transparency will be needed to shape exploration of the state space, as well as to assess quality of decision-making. Some of these tools will be new or applied in new ways to the development and TEV&V of autonomous capabilities.

These M&S support tools will be an essential part of the test resources needed for TEV&V of autonomous capabilities. M&S to support exploration of a decision space, artificial intelligence to enhance efficiency and effectiveness of test planning and design, and instrumentation of both human and machine decision-making all remain open challenges. The instrumentation will at times be restricted by the space, weight, and power constraints common to DoD systems. It will also need to accommodate ongoing regression

testing and expanded range hazards, in recognition of the potential for greater hazards with autonomous systems.

The challenges imposed on TEV&V by the state-space explosion, and on development and TEV&V by any human-machine systems with substantial teaming, do not necessarily arise in programs that seek to add only minimal autonomous capability to existing manned or unmanned systems, analogous to anti-lock brakes or smart cruise control in cars. As a result, there is at least the prospect that the necessary tools and infrastructure could keep up with the T&E demands of increasing autonomous capabilities in some of the near-term future systems.

Contents

1. Introduction 1
   - A. Autonomous Capabilities and Artificial Intelligence 1
   - B. Overarching T&E Challenges of Autonomy 1
     1. Exploding State Space 1
     2. New Elements of Design 3
     3. Challenges of Autonomy and the DoD Acquisition Process 3
2. Assessing Autonomous Decision-making 5
   - A. The Importance of Diagnosis 5
     1. The Nature of Performance Shortfalls or Undesired Behavior 6
     2. Diagnosis in the Case of Human-Machine Teaming 7
   - B. Instrumenting Decision-making 8
     1. Distinguishing Poor Reasoning from Flawed Perception 9
     2. Distinguishing Bad Choices from Poor Reasoning 11
     3. Distinguishing Incomplete World–Models from Incorrect World-Models 11
     4. Distinguishing Bad Algorithms from Bad Training Data 12
     5. Unsupervised Learning 13
     6. Teaming CONOPS 14
   - C. Transparency, Explanation, and Trust 15
     1. Making Autonomous Decision-making Transparent to Humans 16
     2. Enabling Appropriate Trust by Human Team Members 17
     3. Experimentation for Human-Machine Teaming 19
3. Summary and Discussion 23
   - A. Research 23
   - B. Development 24
   - C. Infrastructure 24
   - D. Final Thoughts 25

Appendix A. The Impact of Autonomy throughout the Acquisition Life Cycle A-1
Appendix B. References B-1
Abbreviations C-1

1. Introduction

A. Autonomous Capabilities and Artificial Intelligence

The DoD is pursuing autonomous capabilities in support of a wide variety of missions. The 2018 National Defense Strategy specifically calls out artificial intelligence (AI) and autonomy as necessary enablers of future U.S. military capability. To date many of these systems have provided straightforward enhancements of human performance. Commercial automobiles already feature autonomous capabilities of this kind, such as anti-lock braking systems, smart cruise control, and hands-free parallel parking. The anti-lock braking system decides whether to pump the brakes on the basis of fairly simple thresholding of an easily understood sensor. Capabilities of this sort, which represent evolutionary refinements over their non-autonomous predecessors, generally pose no new challenges for development, testing or fielding.

Systems based on the straightforward examination of thresholds do not pose additional development or test and evaluation (T&E) challenges. The perception, reasoning, and selection steps are all straightforward and one-dimensional. At the other end of the spectrum, we now have the ability to design systems that take in multidimensional inputs, construct and maintain complex internal world models, develop alternative courses of action, and select among those possible actions in ways that cannot be understood in terms of a linear sequence of logical steps. Such systems may require new approaches to development, testing, and fielding.

Both the "Third Offset" of the previous administration and the recent 2018 National Defense Strategy specifically call for AI, autonomy, and "big data" analytics to differentiate U.S. military capabilities from those of potential adversaries. As part of this focus, there are programs in the pipeline that will rely on active teaming between a human and an AI-enabled autonomous partner, providing significant mutual support. This real-time interaction between humans and AI introduces timing sensitivities and psychological factors that add orders of magnitude to the complexity of ensuring safe, secure, and dependable system-of-systems performance.

B. Overarching T&E Challenges of Autonomy

1. Exploding State Space

It may be useful to think of the autonomous capabilities embedded in a machine in terms of how they implement the OODA loop (Observe-Orient-Decide-Act) (Hammond

2001; Roske 2016). Both humans and machines with autonomous capabilities perceive, reason, and select courses of action. [1] We say "select" rather than "decide" to emphasize that in the machine case, there are implicit and explicit decisions being made during the "perception" and "reasoning" stages as well. (In contrast, the evolved human ability to observe a scene instantly and without conscious reflection prompts us to say that the human simply "saw" what was there.) These perception- and reasoning-related decisions must be designed into the machine, for example by explicit instructions or via a trained neural net. For example, the system must decide whether a given sensor input indicates the presence of an entity, decide what kind of entity, and decide how many entities.

If the machine fails to respond appropriately, we do not know a priori whether the problem is in the perception of the entities, the reasoning about their status, or the selection of an appropriate course of action. (It could also be a problem with the sensor itself, and not a failure of the decision-making at all.) For machines, we need additional information to be able to determine where in the perceive-reason-select sequence the decision-making went wrong. The complexity of how the system's AI processes its inputs and selects actions is what makes autonomy possible. The algorithms used typically involve high orders of recursion, feedback loops, and parallel processing of information, combined with extensive use of training data. It is generally impossible to understand what the software is doing in terms of line-by-line execution logic, and the possible combinations of inputs and execution paths is intractably large. This is referred to as the "state-space explosion"—any attempt to enumerate the possible distinct states the system could be in, or the possible execution paths of the software, is doomed to failure (Clark et al. 2012). Statistical sampling of the possible combinations of state space (machine status/external environment/mission/training data…) will not solve this problem, because the relevant sampling distributions are likely to be unknown for autonomous systems—even in the absence of an adversary. For adaptive systems—those that learn or modify their function after fielding—even predicting what states might be reachable in the future is difficult. Confidence in the autonomous capability will depend on building confidence in the entire perception, reasoning, and selecting sequence. We will need to identify techniques that enable testers to know that they have explored the decision state space adequately, even though it cannot be explored exhaustively or with established statistical techniques.

While developing these techniques will be challenging, there are precedents in both the defense and safety communities. The approach is to build a persuasive "dependability" or "assurance" argument using all the available information as evidence (Tate et. al. 2016). For example, in the defense world, systems with energetic materials are fielded after analyses have established very low limits (typically less than 1 chance 1 million or 1 chance

[1] We note that for some AI approaches, and for humans in many circumstances, there may not always be clean divisions between perception, reasoning, and selection. We will continue to use this language for illustrative purposes.

in 10 million) on the probability of a catastrophic failure. It is infeasible to test to statistical significance such low probabilities. Similarly, in the flight-safety and medical-device communities there are procedures for establishing safety when exhaustive testing is not a possibility (NATO 2018).

2. New Elements of Design

Traditionally, DoD systems are designed to achieve certain performance measures, perhaps with some attention to envisioned operator-machine interactions. Then, after design and prototyping are complete, the human operators are trained in using the system. For AI-enabled systems with autonomous capabilities in the areas of perception, reasoning, or selecting courses of action, this approach is unlikely to work (Ilachinski 2017, viii.) Overall performance of the human-machine system will typically be sensitive to the details of the concept of operations (CONOPS) for how and when humans and machines will interact. There are a panoply of trust issues in getting humans to appropriately use the perception, reasoning, or course-of-action selecting capabilities of a machine teammate. In cases where any of these activities or responsibilities are shared, it remains an open question how best to dynamically assign these responsibilities. CONOPS will now be in part a feature of design. The design will need to explicitly address the type and level of operator trust that will maximize operational effectiveness (Parasuraman, Sheridan and Wickens 2000).

There is currently no theoretical basis to guide the design of human-*autonomy* interfaces and CONOPS (Ilachinski 2017, xvi), but there is a considerable body of work on human-machine interfaces (e.g., Billings 1997; Bass and Pritchett 2008). Further, there is beginning to be research into human-machine interface (HMI) and related CONOPS for operation of unmanned or remotely controlled systems (Rice, Keim and Chhabra 2015); however, we have found little research on HMI related to systems with substantial autonomous capabilities. In the absence of a theoretical basis, early experimentation and testing will be required as part of design and development. The experimentation will need to include actual human operators or close surrogates. Diagnosis and improvement of team performance will require visibility into the decision-making processes of both the humans and the machines. The instrumentation required for this will be new, possibly system-specific, and will differ substantially between human and machine within the joint human-machine system.

3. Challenges of Autonomy and the DoD Acquisition Process

We note here that these challenges manifest themselves throughout the acquisition process. Current Policy on Autonomous Weapons Systems (DoDD 3000.09) even mandates activities before formal program initiation and after fielding. How the challenges manifest themselves throughout the acquisition life cycle are described in Appendix A.

2. Assessing Autonomous Decision-making

A. The Importance of Diagnosis

As noted in Section 1.B.1, state-space explosion and highly nonlinear responses to changes in inputs will make it impossible to test perception, reasoning, and selection functions exhaustively or by statistical sampling. This means that it will be impossible to verify and validate the dependability of system decision-making using only mission-level performance measures and "black box" testing. Assessing the quality of decision-making will require visibility into the inner workings of the various decision engines providing the autonomous capabilities.[2] During development, system design and engineering will depend on being able to allocate blame for undesired behaviors or performance shortfalls to know what problem needs to be fixed. Diagnosis—being able to explain why the system is behaving the way it does—will thus be central both to being able to make the system work at all and to achieving confidence that its performance will be dependable.

The importance of diagnosis is well illustrated by recent highly publicized accidents involving vehicles with autonomous and semi-autonomous capabilities. For example, in two separate incidents, a Tesla vehicle operating in the semi-autonomous "auto-pilot" mode rear-ended a stationary fire-truck at 60 mph (Statt 2018; Stewart 2018) . In another recent incident, an autonomous vehicle developed by Uber fatally struck a pedestrian who was crossing the roadway outside a crosswalk at night (Lee 2018a). In the latter case, the vehicle was under test and a safety driver was present, but the driver was not able to intervene to prevent the accident. These incidents, which are currently under investigation by the National Traffic Safety Board, are examples of catastrophic failures that may occur in testing and fielding of autonomous systems. Determining the root cause(s) of such incidents is crucial but challenging. Were these incidents the result of hardware failures, software defects, or deliberate design choices? If the system was working as designed, are there design improvements that could reduce the likelihood of such events, or are these incidents reflective of unavoidable trade-offs that must be tolerated? At the time of this writing, the National Transportation Safety Board (NTSB) investigations have not yet been completed, so we do not have enough details to provide definitive answers. Some details in news reports to date (Shepardson 2018; Nemo 2018; Stewart 2018; Lee 2018b) provide

[2] We note here that "white box" or "clear box" testing also requires visibility into the internal structure of the component or system. In white box testing, the visibility is intended to ensure more exhaustive coverage, which we believe will be infeasible for advanced systems with autonomous capabilities. What we believe will be needed is visibility that supports an assessment when exhaustive coverage is infeasible (see istqb.org and softwaretestingfundamentals.com).

useful examples for our wider discussion of diagnosis in autonomous decision making. We will refer to these incidents below to illustrate specific aspects of the diagnosis problem.

1. The Nature of Performance Shortfalls or Undesired Behavior

Even if exhaustive testing is impractical, we expect extensive testing using a variety of modeling and simulation (M&S) tools to provide important information. Sometimes the result will be that everything appears to be working correctly. If that is not the case, an important function of diagnosis is to enable classification of shortcomings into:

- Situations that must be endured, because the problem cannot be fixed.
- Trade-offs—problems that can be lessened by making something else worse.
- Situations where feasible changes in design or implementation could improve performance across the board.

Consider a simple example of a Doppler radar system to be used to detect ground targets of interest. For fast-moving targets, very high performance is expected. Doppler processing can easily distinguish movers from the stationary background. If the system is having trouble identifying targets, it is not functioning as designed—the problem is in the implementation. This should be a problem that can be fixed without performance trade-offs.

For slow-moving targets, the situation is more complicated. Wind or leakage from objects with large returns can induce false alarms even in a correctly functioning system. Raising the Doppler threshold for detection would reduce the false alarms, but at the cost of reducing the number of true detections. There is no pure solution to this problem—rather, an assessment must be made about which combination of missed detections and false alarms is operationally preferred. Diagnosis can inform this choice by highlighting technical limitations and human factors issues associated with a given point on the trade-off curve, but the underlying problem always remains.

For stationary targets, there is an additional complexity: there may be background objects indistinguishable from the target. This is the case even for high-resolution "imaging" radars. Given the sensor and size, weight, and power (SWaP) constraints, there may be no design that will perform as desired when the target is imaged against certain backgrounds. Recognizing when this is the case is important for efficient development of a useful system.

The questions for diagnosis then are:

- Did the system perform as expected?
- If not, is the environment one in which it is possible for it to perform well?

- If the system is expected to perform well, does the performance need to be traded off or can shortcomings be fixed?

If the performance is below expectations then a deeper level of diagnosis is required to localize the shortfall in one or more autonomous capabilities, or their interactions, or the interactions between humans and autonomy. However, even the question of which regime we are in will be difficult for many autonomous systems, due to the lack of a predictive theory of which problems are easy (like fast-moving targets) and which problems are hard (like stationary targets).

The recent accidents involving Uber and Tesla vehicles illustrate the distinction between performance that can be traded off and performance that can be fixed. The Tesla system, for example, relies on radar that has a high false-alarm rate for stationary objects while the vehicle is moving. Hence, to avoid frequent and unnecessary (and potentially dangerous) stops during "typical" operation, the autopilot system is designed to *ignore* radar returns from stationary objects, leaving the human driver responsible for avoiding any actual stationary obstacles.[3] Similarly for Uber (and likely other autonomous vehicles), the alarm rate due to genuine but benign objects in the road (e.g., small rocks, trash bags, paper, etc.) leads to a trade-off between caution and speed/comfort. Uber has reportedly prioritized comfort in this trade-off, and some suggest that the fatal accident was a result of this design choice (Lee 2018b; Shepardson 2018). Hence these systems likely *behaved as intended*. If this is the case, reducing fatalities will require not merely improved perception of potential hazards, but also a modified human-machine CONOPS, matching the false-positive and false-negative rates to a role that humans can perform reliably. Uber's intent to develop a fully autonomous system will depend on being able to reduce both error rates simultaneously to acceptable thresholds, which may or may not be possible at present. Until it is, the optimal design will not (in general) be the design that gives the smallest possible role to the human driver.

2. Diagnosis in the Case of Human-Machine Teaming

As we have just seen, there are additional failure modes in the case of human-machine teaming. In particular, there are failures that need to be attributed to the interaction between human and machine, and to the allocation of responsibilities between human and machine, rather than to either one individually.

Extend the radar example in the previous section to add a target cuing system to the Doppler radar detector. Suppose we observe that the human system operator is failing to notice a certain type of threat. There are many conceivable reasons for this:

- The radar is not detecting that threat.

[3] Ignoring stationary or slow moving targets is typical for Doppler radar systems.

- The perception function is not correctly classifying the target
- The system is reasoning that cuing is not appropriate for this target.
- The system is selecting the wrong course of action for making the human aware of the target.
- The human is failing to notice the cuing information provided by the system.

The problem might also be some combination of the above. Determining which of these are working as intended and which are not may require extensive instrumentation of the internal states of the autonomous capabilities. Only then can the analysis proceed to the question of whether the problem is caused by errors in the code, incorrect choice of algorithm, inappropriate training data, poor human-machine interface design, inadequate (human) training, or some combination of those factors.

The many possible approaches to human-machine teaming will be constrained by the design. In particular, the possible command-and-control relationships will be determined by the design, which in most cases limits possible CONOPS. The case with which we are most familiar is the human as *operator* of the machine. There is a strict hierarchy. If the human and the machine are *teammates*, both the possible failure modes and the diagnosis become more complex. It may be impossible to "localize" failures in either the machine or the human. For example, there may be multiple teaming courses of action that would be successful, but the machine might pursue one while the human pursues another, leading to failure that cannot easily be attributed to either. We will consider these questions in more detail in Section B, where we discuss how to measure them.

B. Instrumenting Decision-making

The term *instrumentation* suggests specific, quantitative measurements such as velocity, acceleration, or response time. Much of traditional developmental testing is based on identifying the appropriate physical measurements needed to support system development and evaluation. Operational testing, in contrast, requires determinations of *effectiveness and suitability*—inherently qualitative judgments. There are also important qualitative developmental test and evaluation (DT&E) determinations, such as readiness for initial operational test and evaluation (IOT&E). In the case of autonomous systems, these qualitative determinations will also include correctness and adequacy of the training data used (Zhang, Zhou, and Wright 2018), the quality of the human-machine teaming CONOPS (Ilachinsky 2017, xvi), and the quality of the learning exhibited by machines to be fielded with unsupervised learning capabilities (Goix 2016).

In Section A we discussed the importance of being able to understand why an autonomous system is making the decisions it makes. This is a key enabler both of successful system development and of eventually establishing that a system is effective

and suitable for certain purposes. In this section we discuss several specific kinds of diagnostic information that will be needed to support successful development, timely certification, and effective operations when humans and machines work together.

1. Distinguishing Poor Reasoning from Flawed Perception

Beginning with the earliest stage of information processing—perception—it is necessary to distinguish breakdowns that occur at the perception stage from those that occur at the reasoning stage (Saffioti 1997). Returning to our Doppler radar example, we need to be able to tell whether the problem is with how the radar returns are being interpreted or with the conclusions drawn from those interpreted returns.

Human beings sometimes have trouble seeing the distinctions between sensory input, interpretation of that input, and extrapolation (reasoning) from those interpretations because much of that process is subconscious in human perceptions. It seems perfectly natural to say "I see a brown dog," but in fact the ideas of "brown" and "dog" arrive very late in the processing of visual input (Agrawal et al. 2014). The brain "sees" a pattern of nerve excitations arriving from the retina, along with parallel inputs that will be interpreted as color information. The brain processes this information into a constructed photo-like image, which is presented to the conscious mind. That image is further interpreted as a 3-D visual field, adding information about relative sizes and distances. Semantic tags like "dog" might be added at this point as part of perception, for familiar kinds of dog. For unfamiliar kinds of dog, additional reasoning might be required (it has fur, four legs, and a muzzle, and a human is walking it on a leash).

For autonomous systems, every stage from raw sensor input (in terms of photon counts or voltages) to semantic tagging of inferred entities (e.g., "there is an object about 12 meters away at bearing 117°, and it is a dog") must be designed and implemented using a combination of signal processing, logic, and artificial intelligence algorithms. Perception is *hard*, and fundamental challenges remain even after five decades of research and development (Nixon and Aguado 2012). It only seems easy and natural to us because we all rely on highly evolved neurobiological systems that do the hard work for us without conscious effort on our part.

To distinguish flawed perception from downstream errors in reasoning, we would need to have access to the outputs of the perception function in a human-interpretable format that could be compared against the designers' specification of what "correct" perception should look like, given the inputs to the sensors. This is more complex than it sounds, for several reasons. First, the internal world-model of the system will typically not be anything like the internal world-model maintained by a human being (see, e.g., Drouilly,

Rives, and Morisset 2015).[4] Basically, the world model is the database that can be accessed and used by the various decision engines. This database will have been explicitly designed into the machine and will usually contain a number of implicit, and hence impossible to modify, assumptions. Translating world-model states and state-changes into terms that human testers can compare against "ground truth" may be challenging all by itself. Second, what constitutes a "correct" (or even adequate) world model is highly mission-dependent. There is no operational value in world-model contents that are never actually used for perception, reasoning, or selection—or that only add noise to those decisions. Finally, it will also generally not be true that the information that is most useful to humans when doing a given mission is the same as the information that is most important to an autonomous system doing the same mission (Drouilly, Rives, and Morisset 2015).

Instrumenting perception, then, requires both measuring the sensor outputs that are being presented to the perception module and measuring how the machine's world-model is changing as a result. Diagnosis of whether perception is working as intended will further require knowing what the machine's internal representation of ground truth *should* look like, which possible features of an internal world-model are important for correct reasoning and selection, and whether all this processing is happening fast enough to support real-time operations.

Consider again the fatal accident involving the Uber self-driving vehicle. In that case, the autonomous vehicle failed to stop when a pedestrian crossed its path unexpectedly. Early indications suggest that the vehicle *did* detect the pedestrian, but did not choose to initiate evasive action (Shepardson 2018; Lee 2018b). The question, then, is why not? One possibility is that the pedestrian was misidentified as a benign object in the road (e.g., a plastic bag), indicating incorrect perception. Another possibility is the system may have misidentified where the pedestrian was with respect to the roadway, also indicating incorrect perception. Either of these cases would involve corrective action involving the perception module. Alternatively, the system may have extrapolated that the pedestrian would be out of the way by the time the vehicle reached that point of the roadway, which would be a case of incorrect reasoning. And finally, as noted above, the vehicle may have accurately identified the hazard, but left corrective action to the driver due to the high false-positive rate of such identifications. These possibilities are purely speculative at this stage and are only discussed for illustrative purposes. Having knowledge of the vehicle's world model before, during, and after the incident will be necessary to fully tease out where the failure lies.

---

[4] By "world-model" we refer to the database of information available to the perception, reasoning, and selection functions. The possible data values that populate this database and the feedback loops by which the outputs of perception, reasoning, and selection can modify those contents are design-time choices. These choices will always reflect implicit assumptions and are impossible to modify at run-time.

2. Distinguishing Bad Choices from Poor Reasoning

Just as we need to be able to distinguish incorrect perception from downstream errors in decision-making, we also need to be able to distinguish downstream errors in course-of-action selection from poor reasoning. A bad choice is one that would be inconsistent with mission goals and priorities, despite the system's current understanding of the world being complete and correct.

We saw above that distinguishing poor reasoning from flawed perception requires being able to compare internal system world-model states against ground truth. In the same way, distinguishing poor selection from poor reasoning also requires being able to compare the system's understanding of the world (as arrived at through reasoning) against the "correct" representation of the world for purposes of selecting courses of action. The purpose of the outputs of perception is to support good reasoning; the purpose of the outputs of reasoning is to support good choices. There is no a priori reason to suspect that the same kinds of simplifications and/or errors are equally important in both cases. This leads to the conclusion that the relevant measures of effectiveness and measures of performance for individual AI modules within an autonomous system are both architecture-dependent and mission-dependent.

In terms of diagnosis, this will often lead to a multistage problem. First we must determine what kinds of error or data summarization in a world-model are problematic for the downstream decision modules. Then we can test for whether a given module is producing the kind of outputs needed downstream. If it is not, we must further determine whether the problem is that the module has not been coded correctly or whether it is the working as designed, but is the wrong kind of module. If it is working as designed, is the problem in the choice of algorithm, the choice of training data, or the hardware it is implemented on? Can that be remedied, or do we need a new design for this part of the architecture? These questions induce a feedback loop in the design process that could potentially involve major changes to the system architecture.

3. Distinguishing Incomplete World–Models from Incorrect World-Models

Like the distinctions made above, an incomplete (i.e., inadequate) world-model is often confused with an incorrect world-model. An incorrect world-model is inconsistent with ground truth. An incomplete world-model is one where the world-state information maintained by the system is not adequate to support good operational decision-making, no matter how good the perception, reasoning, and selection algorithms being employed, even when the contents of the model are fully consistent with ground truth.

Consider the example of an autonomous vehicle deciding how to brake. Real-time traction feedback from the drive wheels might or might not be sufficient information for effective braking in all circumstances. Significantly better performance in icy conditions might be possible if the world-model included temperature information, or recent weather

history, or a database of road surface materials. If absence of such information is preventing acceptable braking performance in some conditions, that is an example of an incomplete world model.

World models can be incomplete either because they fail to represent necessary information of a particular type (as in the example just given), or because they represent that information at the wrong level of granularity (e.g., knowing that it has rained within the last week is not as valuable as knowing that it is has been raining for half an hour). Distinguishing those two cases may require experimentation. Identifying missing classes of world-model data early in development will be particularly important, since the remedy might involve additional sensors or communications in addition to improved data-fusion algorithms.

4. Distinguishing Bad Algorithms from Bad Training Data

Many current approaches to autonomy rely heavily on the use of *multimodal supervised learning*, a type of machine learning in which a system is trained to produce desired outputs for a given input through the use of labeled multimodal training data.[5] The data labels, which establish ground truth for each training instance, are used to provide corrective feedback when the algorithm produces incorrect output and positive reinforcement when it produces correct output. Multimodal learning has proved to be a powerful tool in areas like image classification, speech and handwriting recognition, and speech synthesis (Baltrušaitis, Ahuja, and Morency 2017).

The capabilities of any supervised learning system depend on the characteristics of the training data set. To begin with, the system cannot learn to produce outputs that were not labeled in the training data—an image classifier cannot guess whether a given image contains a cat unless the training data instances were labeled as "contains a cat" or "no cat." More subtly, the system can only learn to respond appropriately to the range of cases presented in the training set. If all the training pictures labeled as "contains a cat" showed either house cats or tigers, the system probably will not be very good at categorizing pictures containing cougars or jaguars. If the system is meant to be able to recognize cougars as cats, this is a problem that is neither an error in the design nor a bug in the code.

Still more subtly, systems intended to recognize rare events are especially difficult to train. A common approach is to use a "balanced" training set in which positive and negative instances are roughly equal in number, but to apply a weighted penalty function in the

[5] The term *multimodal* here refers to inputs of more than one type (e.g., a combination of visual images, sounds, radar and/or lidar signals, stored references, etc.). For machine-learning purposes, video inputs are themselves multimodal, involving a combination of color, contrast, intensity, and time-series information. Similarly, human speech carries information not only through phonetics, but also pitch, timbre, pace, para-verbal sounds, etc. Interpreting video imagery or human speech is thus a multimodal problem all by itself.

training updates, so that false positives are penalized more heavily than false negatives. This introduces bias into the system response, in exchange for increased confidence in positive responses. At the same time, because positive instances are rare, it may be difficult to find enough distinct labeled positive instances to construct an adequate balanced training set. The usual approach in this case is to oversample the available positive instances, which increases the risk of overfitting to the peculiarities of those instances.

If a supervised learning system exhibits poor false-positive or false-negative behavior in practice, it might be difficult to determine whether the problem is due to the algorithm or due to the training data used. Cross-validation can help, but is of limited value for rare events because of the oversampling of the positive cases, making all test sets look alike. Experimentation with different algorithms and different sample and cross-validation schemes may be necessary. There is also ongoing research in new "explainable artificial intelligence" techniques (see, e.g., Samek, Wiegand, and Müller 2017) that can (in some cases) reverse-engineer the parameters of the trained supervised learning module to determine how (in human-understandable terms) the system is making its output decisions. These techniques require detailed access to the internal topology and numerical weights of the algorithm being used. The Defense Advanced Research Projects Agency is currently pursuing a substantial effort in this area (Gunning n.d.).

5. Unsupervised Learning

A number of commercial applications of AI exhibit learning without the use of labeled input data. Called *unsupervised learning*, it has been used primarily in the areas of classification and anomaly detection. Unsupervised learning can be used before fielding (e.g., to learn to categorize unstructured data), or it can be used to continuously improve performance during operations. DoD has expressed interest in fielding military systems with the ability to learn while deployed, particularly in the area of counter-cyberattack systems that may need to adapt to changing threats on time scales too fast for human command and control (Dua and Du 2016, 100ff.).

In practice, one approach to T&E of systems that continue to learn during deployment might be to hedge our bets by allowing the learning to take place, but only permit the systems to act on the new beliefs after an examination of lessons learned. Regardless of whether the fielded systems are expected to be permitted to act on unsupervised learning, the quality of the learning algorithms will need to be assessed. In most cases, there will be an unsupervised learning phase during development. During that phase, it will be important to develop an understanding of the quality of learning and the behavior of the learning algorithm, which will apply to post-fielding learning as well. This raises the question of how to instrument learning and how to interpret the measurements taken. What does correct learning look like, at the level of algorithm parameters and machine-learning outputs? What warning signs should testers be looking for?

In terms of diagnosis, the evaluations to be supported by the measurements include the following:

- Has the learning algorithm been coded correctly (debugging)?
- Is the learning algorithm behaving as predicted?
- Does the learned behavior improve performance in the context that prompted the learning?
- Does the learned behavior decrease performance in other contexts?
  - If so, are we in a constrained trade space, or is it possible to preserve the improved performance while avoiding the decreased performance?
  - Do we need to retrain the CONOPS, or change the CONOPS, or both?

In general, assessing the performance of unsupervised learning is difficult. Even defining measures of performance that are meaningful and operationally useful is a challenge (Goix 2016). In practice, it may prove to be easier to build unsupervised learning systems than to assess how well they are working.

6. Teaming CONOPS

Assessing the teaming CONOPS will require a lot of transparency. Visibility will be needed into the decision-making of all team members. The "team" may be understood to be anything from a single human-machine pair, to a collection of collaborating autonomous machines, to an arbitrary mix of autonomous human and machine agents collaborating to accomplish a mission. In all cases, the CONOPS will need to be evaluated on the basis of how well the elements of the team perform together to accomplish the mission. (Kalyanam et al. 2016)

Ultimately, the human must be trained to the CONOPS engineered into the machine. Mismatches in expectation between human teammates and machine designers, or CONOPS that are not well-suited to human capabilities, can lead to serious operational shortfalls, as evidenced by the recent accidents involving Uber and Tesla vehicles. In the Tesla incidents, vehicles operating in the semi-autonomous autopilot mode rear-ended stationary fire-trucks at high speed (>50 mph) (Statt 2018; Higgins 2018). Based on warnings found in the Tesla manual, it is likely that detecting stationary objects while the vehicle is driving at highway speeds is difficult for the radar system used by Tesla, and hence Tesla vehicles rely on the human drivers to be alert at all times to intervene in such cases (Stewart 2018). In both incidents, however, the drivers were distracted during operation and did not have their hands on the wheel. Similar incidents with Tesla vehicles suggest that drivers may become overconfident in Tesla's semi-autonomous capabilities with time and, as a result, are not as engaged with driving as intended. The CONOPS

assumed by the designers are perhaps unrealistic; being alert for long periods of time with no actual duties is not something humans are good at (Davies and Parasuraman 1982).

The Uber vehicle that was involved in the fatal crash was designed to operate autonomously and was under test (Lee 2018a; Shepardson 2018). Hence, the intended Uber CONOPS does *not* in principle involve human teaming for driving. The driver in this case was present as an external safety monitor to mitigate the impact of potential failures during testing. But video released from the vehicle during the incident suggests that the driver was distracted and could not perform corrective action when the vehicle ultimately failed to recognize the pedestrian in the roadway. This can be seen as a failure in test safety planning because unrealistic expectations are placed on safety drivers. Again, the NTSB investigations are not complete, so to date, these are simply possible explanations. This incident does suggest that even for "fully" autonomous systems, human teaming CONOPS still need to be considered and optimized to ensure safe operations during testing.

We note that the fixes to the Uber and Tesla systems will require different solutions and probably different approaches. This is a consequence of differences in the target CONOPS. In general, during CONOPS development, it will be vital to be able to trace specific operational performance outcomes back to details of the CONOPS, even as the humans involved are themselves learning and adapting. The instrumentation needed to understand how CONOPS are affecting outcomes in pure machine-machine collaborations is already difficult. Section C addresses the even more complex problem of instrumenting for human-machine CONOPS evaluation, including instrumenting human cognitive behaviors.

C. Transparency, Explanation, and Trust

This section addresses how to assess the quality of decision-making in a collaborative context. One of the major challenges of humans collaborating with autonomous machines is that a machine capable of performing without human input would largely result in the human being out of the loop during routine performance. Specifically, out-of-the-loop unfamiliarity comprises three conditions: (1) the human is unaware of the autonomous system's state or the logic driving its operations (i.e., perception is degraded); (2) the human realizes too late when something goes wrong with autonomous system functions (i.e., reasoning is disrupted); and (3) the human is either too slow or completely unable to intervene due to skill degradation over time (i.e., decision-making and action execution are weakened). There is a general belief that these problems could be lessened if the human is better kept in the loop by receiving continuous explanations of autonomous systems functions (i.e., making the systems more transparent) or by adjusting parameters that influence human trust in the autonomous system. Trust and transparency are not mutually exclusive variables; system transparency plays a significant role in shaping trust in the

system just as trust plays a vital role in operator interaction with both transparent and opaque systems.

1. Making Autonomous Decision-making Transparent to Humans

Automation transparency has been defined as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Wright et al. 2017). This might not realistically imply full awareness of all these elements on the part of the human; the emphasis is generally on the "operator's comprehension" or information that is required by the human to maintain just sufficient awareness of system functions. Transparency, which has been treated as a measure of the automation's openness in information communicated to the operator through the interface, encompasses what the automation is currently doing, which information is being used, how it is being processed, and when it is provided (Westin, Borst and Hillburn 2016). From a design standpoint, decisions have to be made regarding how much and in what ways information should be provided about the criteria, uncertainty, and rationale underlying automation's judgments and problem-solving (Bass and Pritchett 2008). Research has indicated that acceptance and trust in automation can suffer because of the system's opacity (Christoffersen and Woods 2002; Sarter et al. 1997) or mismatches in underlying strategy because the human would solve this problem differently (Westin, Borst and Hillburn 2016). Past and most current models of human-automation interaction have assumed a human in charge of automation in a supervisory control position, which requires a high level of awareness of system function on the part of the human. But in the emerging domain of autonomous systems and human-machine teaming, the need for machine transparency and human awareness of machine functions is even greater.

The increase of automated technologies stems from the view that operators should do as little as possible, since they constitute a major source of variation and unpredictability in system performance. But the process of increasing automation often results in humans being assigned to tasks that the automation designers were incapable of automating due to their technical complexities. When combined with lack of transparency about automation functions, human operators are forced to interact with systems that are difficult to understand and use, which reinforces the rate of human error in collaborative operations with automated systems, rather than reducing it (Helldin et al. 2014). Automation surprises resulting from the automation not performing as expected or acting in a way not anticipated, have been associated with several "out-of-the-loop" human performance issues (Sarter, Woods, and Billings 1997). Specifically, CONOPS that influence machine design may, in turn, drive poor decisions in humans. During the initial stages of trust development, humans often rely on additional sources to substantiate information provided by the system, but the need for verification diminishes as trust in the system increases. This is why

transparency of automated system decisions or relevant supplemental information is necessary, at least in the initial stages of interaction between a human and a machine.

Several frameworks have been proposed to build greater transparency into the designs of autonomous systems. For example, Billings (1997) proposed a set of "human-centered automation" guidelines that highlight the need for including the human operator in the execution of automated tasks, providing appropriate information distribution, and implementing automated functions that are easy to learn and use. For example, the mixed-initiative approach (Tecuci, Boicu, and Cox 2007) stresses the importance of dialogue between the human and the automation and of requiring human input during the problem-solving processes, whereas the team-player approach looks upon the automated system as a member of the team that needs to be taken into account when coordinating the tasks allocated to either the human or the automation. Despite the differences in these frameworks, the common underlying message is that designers must keep humans in the loop to avoid well known problems such as degraded trust and inappropriate automation utilization. Human workload could also be kept at an acceptable level to maintain situational awareness (Helldin et al. 2014).

One common method for increasing automation transparency is to provide explanations underlying the automation's behavior. In the context of e-commerce and semantic web services, diagnostics applications in health care, and museums and cultural institutions, these explanations typically provide an argument for why the user should accept a recommendation (e.g., by comparing it with previous choices), noting what users with similar preferences have chosen or pointing out why a certain item is believed to match the user's characteristics.

Although well-designed explanations can foster trust and lead to better use of automation, poorly designed explanations can be counterproductive by obstructing understanding, resulting in degraded decision-making (Herlocker et al. 2004). Furthermore, increasing transparency by providing more information can exacerbate mental workload if the amount of information exceeds what the operator is capable of processing at any given time (Marois and Ivanoff 2005). From an interface-design perspective, high transparency can lead to cluttered displays (Moacdieh and Sarter 2015). Therefore, optimal transparency should ideally present vital functional information while simultaneously reducing operator information-processing demands. In any system, the design process is highly dependent on domain knowledge, and iterative testing and evaluation are the best way forward to determine optimal parameters for automation transparency (Westin, Borst, and Hillburn 2016).

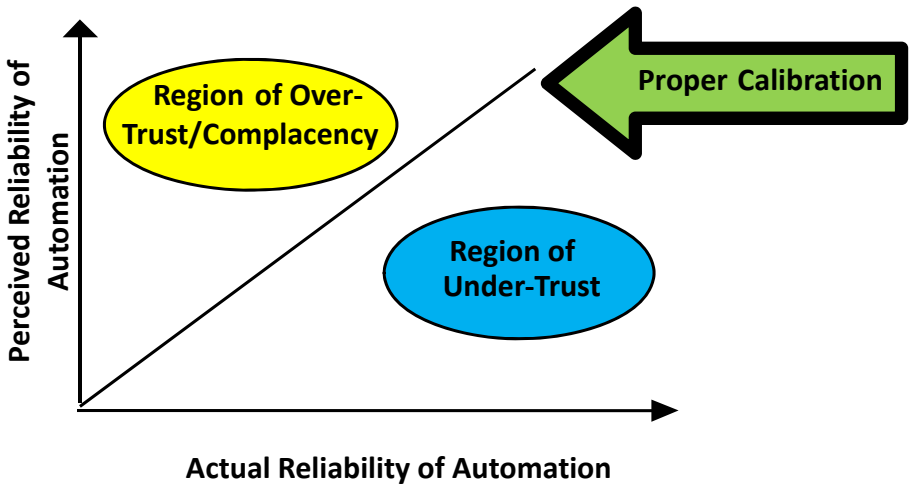
2. Enabling Appropriate Trust by Human Team Members

The prevalence of autonomous systems is expected to lead to a significant redistribution of operational responsibility between human operators and automated

systems. With increasing autonomy, the role of the human will metamorphose from that of a primary controller to that of an active teammate sharing control with automation. Specifically, autonomous systems with both human and machine components need to be modeled as *partners* rather than as tools. These partners would ideally support or assist each other in performing functions that might be difficult or even impossible for either component to perform alone.

One method that interface designers are increasingly using to improve the effectiveness of partnerships is to integrate anthropomorphic attributes (McBride and Morgan 2010). These attributes create more "natural" interactions intended to elicit user trust and increase system acceptance (Marsh and Meech 2000). Ideally, autonomous machines should be designed to interact or behave in a manner similar to a human, imitating human language structures where applicable while also possessing unique knowledge and functional algorithms that may be inaccessible to the human teammate (Madhavan and Wiegmann 2007). No matter how robust the design, however, it is likely that such autonomous system software will fall short of expectations at some point. Such shortfalls are most likely to occur when humans misunderstand the capabilities of the machine (and vice versa), or when the entire system has become the subject of hacking or some other form of technological sabotage. Such situations will lead to a loss of trust on the part of the human.

Trust refers to the expectation of, or confidence in, another. It is based on the probability that one party attaches to cooperative or favorable behavior by other parties (Barber 1983, Muir 1987). For optimal trust in teammates within an autonomous system, the human must be aware of and understand several characteristics of the autonomous system. According to Sheridan (1988), humans use seven system design characteristics to assign trust: reliability, robustness, familiarity, understandability, usefulness, explication of intention, and dependability. Furthermore, it is not sufficient for human operators just to learn to trust a machine; the trust must be *calibrated* to the actual characteristics of the automated system. Calibration is a term used to describe the process by which automated system partners learn to adjust their behavior based upon the specific characteristics (e.g., reliability and performance) of the system. When trust is miscalibrated, the perceived and actual performance of the system are not in proper alignment with one another (McGuirl and Sarter 2006). As can be seen in Figure 1, mistrust (or over-trust) is defined as human trust exceeding the automation's capabilities, leading to over-reliance on the automation, where the human deems periodic verification and validation of information being provided by the system as unnecessary. Distrust (or under-trust) occurs when operators underestimate the reliability of the automation and fail to rely on it as must as they should. Distrust prevents realization of the full benefits of the automated tool; mistrust prevents maximization of the human's strengths.



Source: Adapted from Wickens, Gempler and Morphew (2000).

Figure 1. Model of Trust Calibration.

Although many factors influence trust, effective training methods for trust in automation have received little attention. The Technology Acceptance Model incorporates ease of use along with two other trust characteristics, perceived usefulness and behavioral intention, to determine whether or not a user will accept an automated system (Mathieson, Peacock and Chin 2001). The system design characteristics believed to influence user trust also include system integrity, level of security (Jian, Bisantz, and Drury 2000), and the level of automation (Parasuraman et al., 2000). Two variables have been repeatedly demonstrated to build and maintain human trust in machines: (1) consistency, the sustained good performance (of the system) over time (Kantowitz, Hanowski, and Kantowitz 1997), and (2) predictability, the repeated appearance of the same (machine) error that the human can predict and acclimate to (Muir and Moray 1996). Overall, training to enable calibrated trust in autonomous systems should focus on the logic driving system functions, the principles underlying interface (i.e., visible to humans) design, and an analysis of operational conditions.

To establish optimal indices for both transparency and trust, experimental test beds must be constructed based on the principles of experimental design for behavioral research. This includes creating simulations and scenarios that can be iteratively tested with "sample" autonomous machines and human participants. The parameters for such experimental design and some examples in the domain of human-autonomy interaction are discussed in the next section.

3. Experimentation for Human-Machine Teaming

Maes (1994) presented a model of an informal "testing" approach to study trust and human-machine teaming. In this model, as the user spends more time with a machine, the

user becomes more able to predict the actions of the machine, and the degree of trust increases. This supports the theory that trust is developed over time through mutually satisfying interactions between two parties. However, if trust is breached during the course of the experiment, its immediate effects will be revealed in (1) shifts in objective task performance and (2) verbal reports of trust in the system.

The human-autonomous agent interaction problem space is relatively new; in the last few years, a few autonomy research programs, such as the U.S. Department of Defense Autonomy Research Pilot Initiative (ARPI; Department of Defense 2013), have started to investigate some of the key human-autonomy teaming issues that are necessary for mixed-initiative teams to perform effectively. ARPI encompasses some good examples of how experiments can be designed to test the effectiveness of human-autonomy teaming in a variety of defense-relevant contexts. One notable example under the ARPI, the Autonomous Squad Member (ASM) project, encompasses a suite of research efforts framed around a human's interaction with a small ground robotic team member in a simulated dismounted infantry environment (Chen et al. 2018). The ASM is a robotic agent carrying supplies while autonomously moving toward a rally point. The ASM's reasoning process uses information from the environment, outcomes of its past actions, current resource levels, and its understanding of the current state of its human teammates to inform its actions (Gillespie et al. 2015). The autonomous agent's behavior embodies transparency in that it helps its human teammates to maintain situational awareness by conveying information about its plans, perceptions, reasoning, decision-making, and projected outcomes through its interface. The ASM's communication (to the human) also includes information about its perceived interpretations of the human's intent. This is an initial step in capturing bidirectional transparency between human and machine in an experimental setting, one in which the human would receive valuable information not only about machine states but also about the machine's mental model of the human teammate's behavior.

In a typical human factors experiment, the participant would monitor a simulated environment for threats and would perform a predefined "mission" with different display configurations (i.e., interfaces that communicate with the autonomous machine), with each interface reflecting a different level of information complexity. For example, one set of participants would interact with an interface that would only provide information relevant to the current states and goals of the autonomous system; a second interface condition would provide information on the machine's reasoning process and projected outcomes; a third interface condition would provide information on the machine's interpretations of human behavior, etc. Various combinations of the three interfaces are possible, with the highest level of complexity (and potentially workload) associated with an interface that provides information combining all of the above.

Along the lines of information complexity, transparency is another variable that can be experimentally manipulated by changing the amount and specificity of information provided to the human at specific points in the mission. For example, Chen and Barnes (2012) conducted an experiment to study human interaction with RoboLeader, an autonomous planning agent that acted as a mediator between an operator and a team of subordinate robots. In a series of scenarios, participants engaged in multiple tasks while guiding a convoy of robotic vehicles through a simulated environment with the assistance of RoboLeader. Communication transparency was manipulated via the content of RoboLeader's reports, which varied the amount of reasoning conveyed to the human teammate. The no-transparency condition consisted of RoboLeader's simply notifying the participant when a route change was recommended. In the medium-transparency condition, RoboLeader notified the participant when a route change was recommended and included the reason for the suggested change (e.g., "dense fog observed"). The high-transparency condition was the same as in the medium condition, but also included when the information was received that RoboLeader based its recommendation on (e.g., "dense fog observed," 1 hour). Results revealed that joint human-agent performance was most efficient in the medium-transparency condition because sufficient information was provided to make informed decisions (as opposed to the "no transparency" condition) without leading to information overload and complacency (or, over-trust of and over-reliance on the autonomous agent) that was observed in the high-transparency condition.

In experimental paradigms to test human-autonomous agent interaction, several variables are measured on the human side of the equation: trust (in the autonomous agent), reliance (willingness to depend on the autonomous agent), complacency potential (tendency to over-depend on the autonomous agent and failure to intervene or overturn automation errors), and situation awareness and workload (during the interactive experience). On the "machine side" of the equation, the most important variables to measure are the accuracy of the autonomous agent's perceptions of human behavior, its ability to translate this interpretation into decisions and to convey these decisions on an easily understandable interface, transparency of machine decision-making to the human, and appropriateness of the timing of machine interactions with humans. To measure these parameters and establish optimal operational indices for mixed-initiative systems, iterative testing and evaluation must be conducted in experimental settings similar to the examples discussed in this section. To date, the research literature addresses the relevant quantities underlying collaboration and how to measure them. However, we have not yet progressed to a theory that allows us to assess quality of decision-making in general, as opposed to decision-making linked to a specific scenario. Despite progress in identifying relevant factors and how to measure them, this remains an open area for research.

3. Summary and Discussion

We have identified two elements of autonomous systems that will require new approaches to test and evaluation: the state-space explosion powered by advances in computer hardware and the requirement for human teaming with machines with autonomous capabilities. These challenges have impacts throughout development, test, and evaluation; materiel-release; and operations, as detailed in the appendix. Robust human-machine teaming will probably also require experimentation that links CONOPS and machine design. Section 2 of this paper described how diagnosis, instrumentation, and transparency of the human and machine processes would be instrumental in successful fieldling of autonomous capabilities. Here we provide initial thoughts on the research, development, and infrastructure needed to achieve this.

A. Research

Research to mitigate the impact of the exploding state space is needed to support the diagnosis function. Techniques are needed to more efficiently search the state space for problem areas. Some work in this area is underway by focusing testing on regions of the state space where outputs change rapidly with small changes in inputs (e.g., JHU APL RAPT; Mullins et al. 2018). In addition, there is work at SEI in developing analytics for monitoring operational software behavior by recording and assessing the relations of inputs and outputs over time. Progress in these areas will be essential in developing confidence for those systems that cannot be tested exhaustively (de Niz 2017). As an added benefit, progress will also improve our ability to test systems that can be tested exhaustively.

One of the challenges will be avoiding rare, catastrophic outcomes. For a system to be accepted it is generally necessary to argue that as possible outcomes get worse they also become increasingly rare and further, that the bad outcomes are so rare that the risks are justifiable. This applies even to systems with no autonomous capabilities, and the techniques outlined above are initial steps in this direction. These approaches will be even more important for testing autonomous capabilities.

A particular challenge for such systems is the possibility of emergent behavior, that is, behavior radically different from anything anticipated or intended by the designers. There is no general underlying theory that would allow us to forecast what regions of the input space might trigger emergent behavior or how extensive (and hence easy to find in testing) these regions are. There is ongoing work in formal methods to address this and other challenges.

Finally, we note that research into human behavior when interacting with machines with significant autonomous capabilities remains an open and active area, as discussed above. There are no characterizations of the nature and timing of information exchanges between humans and machines needed to enable effective collaboration and engender appropriate trust. Experimentation will be needed to address the general questions of optimal information flow. Further experimentation in the context of specific systems will be needed as well. Characterization of a machine's "world view" in terms comprehensible to a human is another area of unfinished research.

B. Development

Development of novel instrumentation approaches that provide visibility into the decision processes of the machines—and into the cognitive processes of the humans—will be critical. Diagnosis will be key in development and in testing to ensure dependability. Tools that can collect and archive relevant information about both human and machine will be needed.

In addition, there will be suites of M&S tools with novel needs for resolution and fidelity. M&S tools to characterize the decision envelope as distinct from the physical envelope will be needed. These tools can be developed to use low-resolution or low-fidelity models of the environment, but will need precise representations of sensor input format expected by the perception module. A low-fidelity weather model can be used to drive a perception engine and allow one to assess the decisions being made, but the formats have to match for it to work.

Tools of this type allow one to examine a point in the decision space. Tools built to explore the space efficiently (built on approaches such as RAPT mentioned above) will also be needed. Efficient exploration may often require much faster than real-time M&S support. Building these tools with just the minimum resolution for an assessment will be a key element in the efficiency.

C. Infrastructure

The essential autonomy-driven infrastructure will be software test beds with extensive data-archiving capability. The test beds will be needed to house the M&S developed to drive the scenarios in which the perception, reasoning, and selection decision engines operate. In addition, extensive data archiving and retrieval capacity will be needed for two reasons. First, the experimentation and feedback loops required early in development can probably be executed more efficiently if the data on earlier designs is maintained. Second, in cases where exhaustive testing is not possible, the development record will be part of the argument made for dependability at the time of materiel release decisions. We will want the M&S-driven test and evaluation to be as efficient (and therefore as exhaustive) as possible. This will extend the circumstances when exhaustive testing is feasible and

increase confidence in the systems when it is not. Live, virtual, and constructive facilities will be needed to address questions of human machine teaming. How CONOPS are designed into the machine and how the CONOPS are tested will require participation of operators or operator surrogates—either other humans or models. Early experimentation may require both AI-driven simulation of human behavior and human emulation of AI behavior to support this exploration of possible teaming concepts. Live, virtual, and constructive facilities will also be needed for diagnosis of the human-machine teaming. It will be necessary to immerse the humans (and the machine) into a virtual reality that provides transparency into their decision-making.

D. Final Thoughts

The challenges imposed on test and evaluation, verification and validation (TEV&V) by the state-space explosion, and on development and TEV&V by any human-machine systems with active teaming, do not necessarily arise in programs that seek to add only minimal autonomous capability to existing manned or unmanned systems. The research and tools discussed here can make these developments more efficient, and perhaps expand the boundaries of what capabilities can be developed and tested with traditional approaches. As a result, there is at least some prospect that the necessary tools and infrastructure could keep up with the T&E demands of increasing autonomous capabilities in future systems.

Appendix A.
The Impact of Autonomy throughout the Acquisition Life Cycle

The development and acquisition of autonomy-based capabilities will introduce challenges to test and evaluation (T&E) throughout the system life cycle. Here, we examine the impact of challenges and recommendations discussed above on each phase of the system life cycle. Figure A-1 displays the standard system life cycle as defined in DoD Instruction 5000.02, Section 5.c.(3). We will use this life cycle to frame the alterations that autonomy introduces to the typical T&E process.

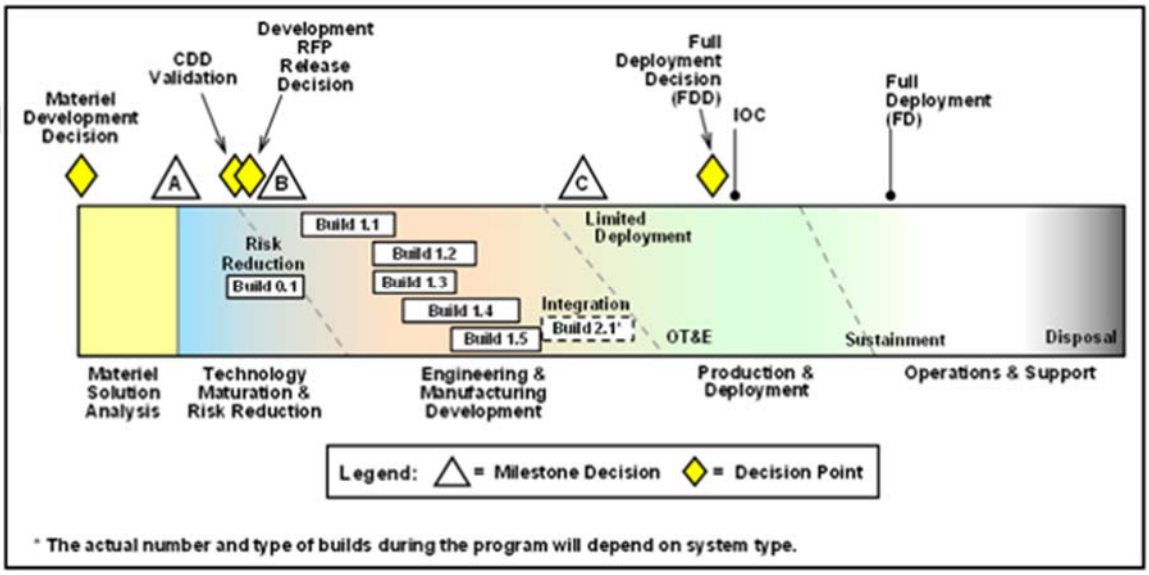


Figure A-1. Operation of the Defense Acquisition System from DoD Instruction 5000.02, Section 5.c.(3) January 2015, USD(AT&L)

Before Formal Development

Autonomy introduces practical challenges even before entry into the standard system life cycle, in the form of an initial assessment to determine the autonomous capabilities with the greatest impact, which is necessary for subsequent evaluation efforts. The inherent complexity of autonomous system acquisition will require early planning of activities and assessments to enable the overall process to ultimately fit together in a timely manner to provide sufficient information for appropriate system diagnosis. This theme will continue throughout the acquisition process, with the general effect of shifting activities to an earlier point in the process. To this end, any autonomous capabilities must be characterized in

order to identify appropriate strategies for instrumentation of the autonomous system and subsequently tailor the remainder of the process in anticipation of the challenges associated with the T&E of autonomous systems before formal development. The statement of DoD Directive 5000.01 with regard to tailoring program strategies is especially important during the development of autonomous capabilities, due in large part to the T&E challenges associated with developing a sufficiently rich body of quantitative data to accurately assess autonomous capabilities.[1] In fact, the GAO (2015) has linked the failure of several large-scale IT development efforts to the lack of appropriate program tailoring. Given the complex nature of autonomous capability development and the IT basis for autonomous capabilities, program tailoring based on an initial characterization of the involved autonomy will be fundamental to the success of autonomous capability development.

Beyond the determination of instrumentation strategies, an initial characterization of autonomous capabilities must also determine the applicability of DoD Directive 3000.09, "Autonomy in Weapon Systems." This directive establishes DoD policy and assigns responsibilities for the development and use of autonomous and semi-autonomous functions in weapons systems, including both manned and unmanned platforms. In so doing, this document establishes guidelines designed to minimize the probability and consequences of failures of autonomous capabilities. The directive has a specific focus on weapon systems that apply force, whether lethal or nonlethal, kinetic or non-kinetic. Figure A-2 displays a flowchart that may be used to determine if DoDD 3000.09 applies.

[1] DoDD 5000.01, paragraph 4.3.1 states, "There is no one best way to structure an acquisition program to accomplish the objective of the Defense Acquisition System. MDAs and PMs shall tailor program strategies and oversight, including documentation of program information, acquisition phases, the timing and scope of decision reviews, and decision levels, to fit the particular conditions of that program, consistent with applicable laws and regulations and the time-sensitivity of the capability need."

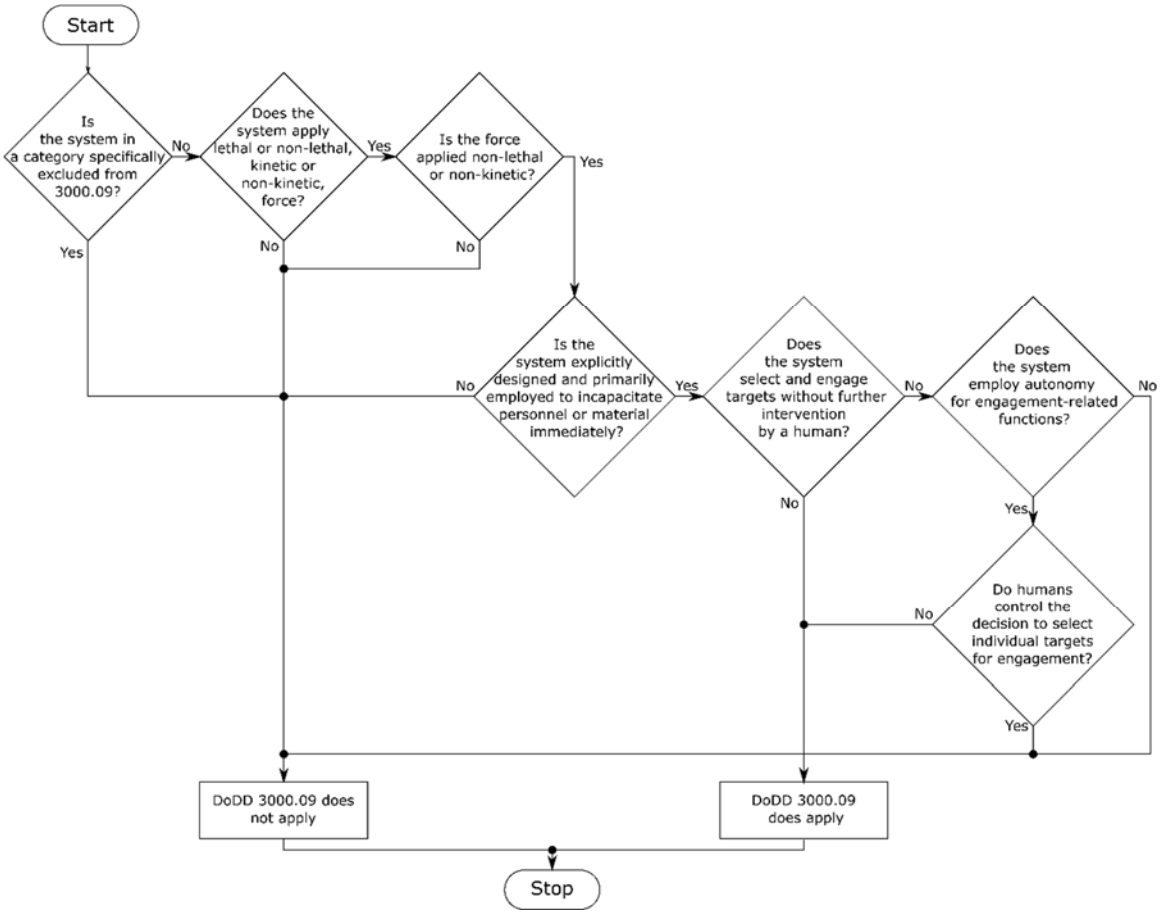


Figure A-2. Flowchart for Determining Whether DoDD 3000.09 Applies

When DoDD 3000.09 does apply, additional efforts are necessary to ensure that the system design and development approach display several qualities (enumerated in DoDD 3000.09; see below) before the formal development of the autonomous capability may proceed. Ensuring these qualities will require developing a body of quantitative evidence to demonstrate the existence of specific attributes of the autonomous capability. This implies a shift of T&E tasks necessary to support an initial assessment of autonomous weapon systems, to even before entry into the formal development process. Moreover, autonomous weapon system development requires laying the groundwork for both verification and validation (V&V) and T&E even before the first step of formal development as a result of DoDD 3000.09. That is, Enclosure 3 of DoDD 3000.09 specifically requires that:

> Before a decision to enter into formal development, the USD(P), USD(AT&L), and CJCS shall ensure:

1. The system design incorporates the necessary capabilities to allow commanders and operators to exercise appropriate levels of human judgement in the use of force.

2. The system is designed to complete engagements in a timeframe consistent with commander and operator intentions and, if unable to do so, to terminate engagements or seek additional human operator input before continuing the engagement.
3. The system design includes safeties, anti-tamper mechanisms, and information assurance in accordance with Reference (a) [DoD Instruction 8500.01], addresses and minimizes the probability or consequences of failures that could lead to unintended engagements or to loss of control of the system.
4. Plans are in place for V&V and T&E to establish system reliability, effectiveness, and suitability under realistic conditions, including possible adversary actions, to a sufficient standard consistent with the potential consequences of an unintended engagement or loss of control of the system.
5. A preliminary legal review of the weapons system have been completed, in coordination with the General Counsel of the Department of Defense (GC, DoD) and in accordance with References (b) [DoD Directive 5000.01] and (c) [DoD Instruction 5000.02], DoD Directive 2311.01E (Reference (f)), and where applicable Reference (d) [DoD Directive 3000.03E].

DoDD 3000.09 is primarily concerned with minimizing failures that could lead to unintended engagements or loss of control of autonomous weapons systems. To that end, the directive mandates the creation of a plan for "rigorous hardware and software V&V and realistic developmental and operational T&E, including analysis of unanticipated emergent behavior" (DoDD 3000.09 Enclosure 2.a). Note this plan must also include the definition of metrics and procedures to "identify any new operating states and changes in the state transition matrix" (DoDD 3000.09 Enclosure 2.b.(1)). to support subsequent regression testing whenever the autonomous capabilities are updated. Therefore, this plan is intended to provide a path to developing trust in autonomous capabilities, with specific focus on demonstrating that autonomy will not degrade the safety of those that employ it.

While DoDD 3000.09 does not apply to all systems that have significant levels of autonomous capabilities, the approach to laying the groundwork for T&E early is likely to be necessary for all such systems. As noted above, the complexity inherent in autonomous capabilities leads to a state-space explosion that makes exhaustive evaluation of an autonomous capable infeasible. Coping with this reality demands building up sufficient data to support an understanding of the autonomous operation, which itself demands evaluation of the autonomous capability that continues throughout system development and is based on the particular capabilities involved. Early assessment of the involved capabilities and development approach is therefore necessary to support the development process. Furthermore, autonomous systems that include significant human-machine interactions demand early identification of appropriate strategies for developing and

assessing CONOPS as well as initial planning for any necessary experimentation. Thus, even in the situation that the formality of DoDD 3000.09 does not apply, the need for framing T&E activities before formal development based on identified characteristics of the system is present for any autonomous capability.

Materiel Solution Analysis

During the Materiel Solution Analysis (MSA) phase, the T&E groundwork laid before the start of formal development must be refined to support ongoing system development. In particular, the Initial Capability Document, or other documentation (e.g., DoDAF or AoA guidance), should be reviewed for further details regarding any complex autonomy. Several capabilities that may appear in the Initial Capability Document will suggest the potential for challenges that will require refining the T&E approach (e.g., handling big data is likely to imply machine-learning and data-management challenges). Identifying any challenges early will allow addressing them by developing appropriate mechanisms for system instrumentation and tailoring the acquisition process to the specific realities of the system being developed.

Identifying challenges will require a careful consideration of the system design, with particular focus on key performance parameters (KPPs), key support areas, and their supporting rationale. For example, the Net Ready KPP applied to autonomous capabilities implicitly requires human-machine cooperation analogous to machine interoperability. Because testing shared human-machine missions and decision-making remains an immature field, several additional challenges will likely arise throughout the remainder of the acquisition process. As exemplified here, missions that are new or conducted in new ways, such as involving artificial agents, may signal implicit challenges that should be anticipated early.

Beyond simply refining the approach to T&E, measurements of the system must begin during the MSA phase. Measurements must start early to support diagnosis of complex autonomous systems and facilitate identification of separable aspects of the systems to simplify the operation of subsequent efforts. In addition, data collection must start early, even though the data may be used late in development or in support of fielding decisions; the need to establish trends in autonomous system behavior demands that data regarding system operation be collected throughout all phases of development. Such archived data will be essential for assessing the quality of decision-making in those circumstances when state-space explosion makes exhaustive testing infeasible. Note that the required data-collection approach and necessary supporting infrastructure will depend on the situation. Hazard information is certainly necessary for autonomous systems, although the nature of the approach for collecting this information depends on the applicability of DoD 3000.09 or the expectation of software safety releases. Alternatively, subsequent T&E of human-machine teaming systems will require early characterization data to support further system

refinement. Migration of responsibilities from the operator to the software introduces additional coupling of development, test, experimentation, and possibly evolving tactics, techniques, and procedures. Each of these elements must be separated, to the extent possible, on the bases of data that characterize the nature of the interaction between human and machine to support a tractable evaluation of the total system. Thus, early establishment of measurement systems is critical to the ongoing development of autonomous systems.

A refined understanding of the autonomous capabilities and early measurement data must also be applied to plan later approaches to system evaluation. Note that autonomous capabilities may change how missions are accomplished, or even what missions can be accomplished. Thus, planning for evaluation of new features of mission accomplishment will cut across all other challenges and begin during the MSA phase. Since systems with autonomous capabilities are agents as well as tools used by operators, planning during the MSA phase must identify the method to evaluate the agents' actions. This planning must also determine an approach to developing appropriately calibrated trust of operators. The identification of both the evaluation approach and the calibration of trust require early measurement data. Note that beyond simply supporting evaluation of intended system behaviors, planning must also enumerate those things that the system must not do. Such an enumeration requires the identification of risks (probability and consequence) of any undesirable actions in a more general fashion than is likely available from initial documentation. Risk identification should include both risks of system malfeasance and risks associated with unintended emergent behavior. Planning the evaluation in the MSA phase will allow subsequent T&E operation.

Technology Maturation and Risk Reduction

Experimentation and prototyping is significant for the success of autonomous system development, and thus the competitive prototyping of the Technology Maturation and Risk Reduction (TMRR) phase is critical for the subsequent development of the autonomous capabilities. To enable useful experimentation, appropriate facilities and mechanisms must be identified according to the characteristics of the autonomous capability. These facilities and mechanisms must appropriately address any safety considerations and provide the appropriate instrumentation for T&E. An approach for conducting the necessary experimentation will therefore need to be defined sufficiently early in the acquisition process to enable assessment of the required facilities and mechanisms.

The operation of T&E during TMRR largely relies on the software development producing testable and representative portions of the system at regular intervals. Planned T&E activities must therefore include assessment of the software-development approach employed for autonomous system development and tailor the T&E schedule to the particular development schedule. In addition, the T&E planning should anticipate ongoing testing and assessment of cyber vulnerabilities as the system evolves. Furthermore, because

the T&E planning should identify the points at which training data must be available, the T&E during the TMRR must be based on the software-development schedule for the autonomous capabilities.

In addition to the standard software-development considerations, the T&E of autonomous systems must also consider any human-machine interactions early in the process as well. Particular emphasis is needed for cognitive elements formerly provided by trained humans that now need to be designed and matured in the machine. These elements have the effect of shifting many system considerations left when autonomous system acquisitions are compared with that of more traditional systems, since these elements are solidified during design/build time and not in subsequent personal training. In support of such testing, identifying human operators to support iterative testing later in the acquisition process must occur during the planning of the TMRR phase to allow sufficient time for the associated training and logistics necessary to bring human resources to bear. Note that human operators must be identified both for interacting with the machine portions of the autonomous systems as well as to serve as component analogues for any live, virtual, and constructive simulation activities. Thus, the TMRR must address the practical realities of involving human participants in the acquisition process rather than allowing these activities to occur during a later phase.

Planning during the TMRR must also support later fielding decisions. Specifically, the data and analyses necessary to support full-rate production (FRP) and full-material release must be identified early so that these data may be appropriately collected and analyzed. Note that this is especially true for data related to human-machine teaming as well as data related to range hazards, since both these cases demand establishing system trends rather than examining performance relative to a threshold.

The TMRR phase ends with Milestone B, implying that testing has shown the following qualities for software systems, such as those that underpin autonomous capabilities:[2]

- Algorithms exist and will provide the needed capability.
- SWaP constraints will not preclude the effective use of the algorithm.
- Early measurements on latency, power, and cooling are appropriate.
- The data necessary to train algorithms exist.

Establishing these capabilities via T&E during TMRR will demand a few specialized activities that depend on the nature of the autonomous system involved. For example, data may be necessary to train algorithms both to check for the performance and to examine

[2] Major Defense Acquisition Programs: Certification Required before Milestone B or Key Decision Point B Approval, 10 USC Section 2366b (2011).

robustness; such data should be provided during the course of the TMRR or their generation should be specified. Systems that involve human-machine teaming will require a more iterative assessment as the operators and designers eventually determine the most effective means of human-machine teaming interaction.

Finally, a DT&E sufficiency assessment addressing plans, schedules, resources, risks of concurrency, and entrance criteria for the production phase must be conducted (mandated by DoD Instruction 5000.02). In addition to the concepts already highlighted herein, the entrance criteria should specifically include a demonstration of adequate transparency, attention to hazards, and the presence of data to support quantification of risk.

Programs That Enter at Milestone B

Programs that enter the acquisition cycle at Milestone B should still largely have addressed the above concerns, albeit in a potentially less formal manner. Of particular note, a determination of the applicability of DoDD 3000.09 must already be made for programs that enter at Milestone B. In the case that DoDD 3000.09 applies, these programs must demonstrate safety assurance through commander control before entering the formal acquisition process. Beyond the concerns of DoDD 3000.09, the requirements of 10 USC Section 2366b for technology demonstration and DT&E sufficiency assessments still apply for all systems. Risks of concurrency, in particular, will warrant special attention. The formal structure phases discussed above should ensure that input data for training, plans for human-machine teaming testing, and data archiving to support ultimate fielding are all in place; these aspects of the T&E process must still exist in the absence of a formal process. In addition, for systems with substantive human-machine teaming, extensive experimentation to support CONOPS development will need to be completed for a positive sufficiency assessment. Generally speaking, the full set of activities outlined above for the MSA and TMRR phases will need to be completed.

Engineering and Manufacturing Development

During the completion of autonomous system development, the four notions of safety, reliability, interoperability, and cybersecurity must all be treated on an equal footing with performance. T&E must therefore have sufficient capability and capacity to archive measurements and assessment results across five notions to support later full-material release. Note that the assessments of each concept are logistically interrelated (e.g., a safety evaluation and supporting data will be needed for open-air testing) and a clear plan that addresses each relevant aspect of the autonomous system is necessary at the onset of Engineering and Manufacturing Development (EMD). Furthermore, examination of software safety, performance, reliability, interoperability, and cybersecurity will require extensive system integration as well as an integrated approach to regression testing. This

is driven by software coupling across domains through multiple interfaces and will require the critical system software to be tested and embedded in the system sufficiently early to allow for subsequent T&E. Throughout the integration process, cybersecurity T&E should address both attack vectors and at-risk functionality, archiving a historical record of identified vulnerabilities and responses. The completion of autonomous system development therefore presents a multidimensional program that requires integration of both the system and the testing approach.

Those systems that involve human-machine teaming will continue to present additional challenges to the development process. Significantly, the involvement of human operators, and potentially M&S surrogates for human operators, will be critical to the final system design. In large part, this involvement should support experimentation to determine appropriate CONOPS for the overall system of human and machine. In planning for EMD, any additional logistics or training necessary for the involvement of human operators must be considered.

The EMD phase ends with Milestone C, which requires an assessment of the sufficiency of the developmental T&E completed and the operational T&E planned. This assessment is likely to necessitate a complete summary of DT&E activities, including both completed DT&E as well as plans and resources for remaining DT&E. This summary effort will need to identify any risks exacerbated by the deferred DT&E, which must consider the additional complexities and assessment interdependencies implied by the development of an autonomous system. In addition, the Milestone C assessment will need to characterize the level of demonstrated readiness for IOT&E across several challenges specific to autonomous systems. For example, IOT&E must address hazards that might arise from post-fielding adaptability of the system. In addition, planning for IOT&E must consider the availability of validated and verified, quality input data for the operational testing of data-dependent systems. Moreover, there must be adequate attention to human-machine interfaces in training of warfighters that will participate in operational testing.

Milestone C decisions are often primarily about risk. As a result, the evaluation must include a sufficient treatment of what the system must not do or must do only under very rare circumstances. Supporting this will likely require an evaluation of the history of regression testing results as a means to provide insight into the maturity and stability of the software. This historical view of the autonomous capability is especially important in developing an understanding of system trends in light of the infeasibility of exhaustive testing approaches. Utilizing system-evaluation results along with collected supporting data enables the connection of flaws in perception or decisions to resultant impacts on performance and reliability in a substantive manner. Note that reliability of perception or decision-making software does not directly map to either the failure rate (i.e., mean time between failures for electronics) or probability of failure (i.e., percent failure for munitions) of the overall capability system; that is, “Failure Definition and Scoring Criteria” are

unlikely to capture relevant features of autonomous capability such as the reliability of individual phases of the autonomous operation. Instead, custom, subsystem-specific definitions of software reliability for autonomous capabilities (e.g., definitions that focus on the particular activities of perception) are likely necessary.

Production and Deployment

Milestone C will no longer be the end of development or developmental test for autonomous systems. Software development continues throughout system operation, as new bugs are found and patched or features are added. In addition, regression testing never ends and is instead an ongoing, continuous effort. While such continuation of regression is only mandated for weapon systems (DoDD 3000.09), such an effort is necessary to ensure appropriate system behavior in changeable environments, missions, or CONOPS. As missions continue to evolve, ongoing development and developmental test efforts will be necessary. The test and evaluation master plan must address an OT&E plan for the future, especially if the system is self-adaptive.

Planning for OT&E of autonomous capabilities must address several unique challenges. As noted above, the CONOPS will be a feature of the design for human-machine teaming systems. Recall that experimentation and testing are necessary to support the design and development of CONOPS due to the lack of an available theoretical basis. Human operators will need to be available to support the continuous software development associated with the experimental refinement of CONOPS. In addition, the complexity of autonomous systems, especially when operating in groups or when teamed with human operators, implies that OT&E will eventually reveal unexpected features of the underlying algorithms. OT&E must then rely on system transparency to turn such features into opportunities or fixes, as appropriate. Finally, while testing will continue throughout the system operation, fully continuous testing may not be necessary—spot checking may be sufficient. The planning effort must evaluate each of these aspects and develop an approach appropriate for the system at hand.

Operations and Support

The perception and decision-making elements of fielded autonomous capabilities are expected to continue to advance. This will certainly be the case for any self-adaptive or self-organizing system. In addition, missions will evolve, and ongoing use of autonomous capabilities are likely to yield new standards for CONOPS. As capabilities and missions evolve, so too will expectations for the system safety, performance, reliability, interoperability, and cybersecurity, even if these expectations are not reflected in formal requirements or specifications. T&E will therefore be required to describe and potentially support adaptation of autonomous capabilities throughout their overall lifespan. This directly affects software maintenance, which will extend beyond finding bugs and

incorporating deferred capabilities to include the addition of unplanned capabilities. As a result, system-regression testing will be ongoing and likely rely on T&E or M&S capabilities initially established during the system development. Thus, T&E should expect to continue to advance during the sustainment of autonomous systems.

Summary of Life-Cycle Impacts

Autonomous capabilities of machines will pose new and novel challenges for T&E throughout the system life cycle. T&E will have to shift many activities to address the inherent complexity of autonomous capabilities and minimize risks of concurrency. The requirement to design machines to do what humans were trained to do will require earlier operator involvement and increased attention to the human-machine interactions throughout the life cycle. T&E must be designed throughout to mitigate the distinctive challenges associated with autonomous capabilities; this will include beginning T&E earlier and archiving more data than would be done for systems with no autonomous capabilities.

Appendix B.
References

Agrawal, Pulkit, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. 2014. "Pixels to Voxels: Modeling Visual Representation in the Human Brain." arXiv preprint arXiv:1407.5104.

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

Barber, B. 1983. *The Logic and Limits of Trust*. New Brunswick, New Jersey: Rutgers University Press.

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.

Billings, C. 1997. *Aviation Automation: The Search for a Human-Centered Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chen, J. Y. C, and M. J. Barnes. 2012. "Supervisory Control of Multiple Robots in Dynamic Task Environments." *Ergonomics* 55: 1043–58.

Chen, J. Y. C., S. G. Lakhmani., K. Stowers, A. A. Selkowitz, J. L. Wright, and M. Barnes. 2018. "Situation Awareness-Based Agent Transparency and Human-Autonomy Teaming Effectiveness." *Theoretical Issues in Ergonomics Science* 19:259–82.

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.

Davies, D, and R. Parasuraman. 1982. *The Psychology of Vigilance*. London, U.K; Academic Press.

de Niz, Dio. 2017. "Certifiable Distributed Runtime Assurance." *Research Review 2017*. Poster Paper. 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. 2015. *High Risk Series: An Update,* GAO-15-290, A Report to Congressional Committees. Washington, DC: Government Accountability Office, February.

Gillespie, K., M. Molineaux., M. W. Floyd., S. S. Wattam, and D. W. Aha. 2015. "Goal Reasoning for an Autonomous Squad Member." Technical Report GT-IRIM-CR-2015-001, 52–67. Atlanta, GA: Georgia Institute of Technology.

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, 2016.

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

Hammond, Grant T. 2001. *The Mind of War: John Boyd and American Security.* Washington, D.C. Smithsonian Institution Press.

Helldin, T., U. Ohlander., G. Falkman, and M. Riveiro. 2014. "Transparency of Automated Combat Classification." *Engineering Psychology and Cognitive Ergonomics* 22–33.

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

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

Jian, J., A. M. Bisantz, and C. G. Drury. 2010. "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics* 4:53–71.

Kalyanam, Krishnamoorthy, Meir Pachter, Michael Patzek, Clayton Rothwell, and Swaroop Darbha. 2016. "Optimal Human–Machine Teaming for a Sequential Inspection Operation." *IEEE Transactions on Human-Machine Systems* 46, no. 4: 557–68.

Kantowitz, B. H., R. J. Hanowski, and S. C Kantowitz. 1997. "Driver Acceptance of Unreliable Traffic Information in Familiar and Unfamiliar Settings." *Human Factors* 39:164–76.

Lee, Timothy B. 2018a. “Report: Software Bug Led to Death in Uber’s Self-driving Crash.” *ArsTechnica,* May 7. Accessed May 21, 2018. https://arstechnica.com/tech-policy/2018/05/report-software-bug-led-to-death-in-ubers-self-driving-crash/.

Lee, Timothy B. 2018b. “Uber Self-driving Car Hits and Kills Pedestrian [Updated].” *ArsTechnica,* March 19. Accessed May 21, 2018. https://arstechnica.com/cars/2018/03/uber-self-driving-car-hits-and-kills-pedestrian/.

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.

Maes, P. 1994. “Agents that Reduce Work and Information Overload.” *Communications of the ACM* 37:30–40.

Maoris, R, and J. Ivanoff. 2005. “Capacity Limits of Information Processing in the Brain.” *Trends in Cognitive Science* 9:296–305.

Marsh, S, and J. Meech. 2000. “Trust in Design.” *Conference on Human Factors in Computing*. The Hague, Netherlands.

Mathieson, K., E. Peacock, and W. W. Chin. 2001. “Extending the Technology Acceptance Model: The Influence of Perceived User Resources.” *The DATA BASE for Advances in Information Systems* 32:86–112.

McBride, M, and S. Morgan. 2010. “Trust Calibration for Automated Decision Aids.” *Institute for Homeland Security Solutions: Research Brief* 1–11.

McGuirl, J. M, and Sarter, N. B. 2006. “Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic Systems Confidence Information.” *Human Factors* 48:656–65.

Meyer, Bertrand, and Martin Nordio, eds. 2011. *LASER, International Summer School 2011, Elba Island, Italy, Revised Tutorial Lectures.* Berlin, Heidelberg: Springer-Verlag.

Moacdieh, N, and N. Sarter. 2015. “Display Clutter: A Review of Definitions and Measurement Techniques.” *Human Factors* 57:61–100.

Muir, B, and N. Moray. 1996. “Trust in Automation. Part II: Experimental Studies of Trust and Human Intervention in a Process Control Simulation.” *Ergonomics* 39:429–60.

Muir, B. 1987. “Trust between Humans and Machines, and the Design of Decision Aids.” *International Journal of Man-Machine Studies* 27:527–39.

Mullins, Galen E., Paul G. Stankiewicz, R. Chad Hawthorne, 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.

Nemo, Leslie. 2018. “Tesla Crash Shows Drivers Are Confused by ‘Autonomous’ vs. ‘Autopilot.’” *Futurism,* May 18. Accessed May 21, 2018. https://futurism.com/tesla-crash-confused-autonomous-autopilot/.

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.

Parasuraman, R., T. B. Sheridan, and C. D. Wickens. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans* 30:286–97.

Rice, Thomas M., Erik A. Keim, and Tom Chhabra. 2015 "Unmanned Tactical Autonomous Control and Collaboration Concept of Operations." Master's thesis, Naval Postgraduate School. https://calhoun.nps.edu/handle/10945/47319.

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

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. Wiley, New York.

Shepardson, David. 2018. Uber Sets Safety Review; Media Report Says Software Cited in Fatal Crash. *Reuters,* May 7. 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.

Sheridan, T. B. 1988. "Trustworthiness of Command and Control Systems." *Proceedings of the IFAC Man-Machine Systems* 427–31.

Statt, Nick. 2018. "Tesla Crash Involving Autopilot Prompts Federal Investigation." *The Verge*, May 16. Accessed 21 May 2018. https://www.theverge.com/2018/5/16/17363158/nhtsa-tesla-autopilot-crash-investigation.

Stewart, Jack. 2018. "Why Tesla's Autopilot Can't See a Stopped Firetruck." *Wired,* January 25. Accessed May 21, 2018. https://www.wired.com/story/tesla-autopilot-why-crash-radar/.

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,

Tecuci, G., M. Boicu, and M. T. Cox. 2007. "Seven Aspects of Mixed-Initiative Reasoning." *AI Magazine* 28:11–18.

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.

Wickens, C. D., K. Gempler, and M. E. Morphew. 2000. “Workload and Reliability of Predictor Displays in Aircraft Traffic Avoidance.” *Transportation Human Factors* 2:99–126.

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.

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

Abbreviations

AI	artificial intelligence
ARPI	Autonomy Research Pilot Initiative
ASM	Autonomous Squad Member
CONOPS	concept of operations
DARPA	Defense Advanced Research Projects Agency
DT&E	developmental test and evaluation
EMD	Engineering and Manufacturing Development
GAO	Government Accountability Office
IOT&E	initial operational test and evaluation
KPP	key performance parameter
M&S	modeling and simulation
MSA	Materiel Solution Analysis
NTSB	National Transportation Safety Board
OODA	observe-orient-decide-act
SWaP	size, weight, and power
T&E	test and evaluation
TEV&V	test and evaluation, verification and validation
TMRR	Technology Maturation and Risk Reduction
V&V	verification and validation

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
June 2018	Draft Final	Apr 2018 – Jun 2018

4. TITLE AND SUBTITLE

Assessing the Quality of Decision-making by Autonomous Systems

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

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Sparrow, David A.
Tate, David M.
Biddle, John C.
Kaminski, Nicholas J.
Madhavan, Poornima

5d. PROJECT NUMBER
AX-2-4383

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
Systems and Analyses Center
730 East Glebe Road
Alexandria, VA 22305-3086

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Paper P-9116

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OSD (DASD(T&E))
3090 Defense Pentagon, Room 5A1076
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

OSD (DASD(T&E)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The key distinguishing feature of autonomous systems is that they make decisions, both in interpreting their environments and in 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 this paper, we argue that observed system performance will not be sufficient to evaluate autonomous decision-making in the ways necessary for successful deployment, especially for systems designed to team with humans. Instead, novel instrumentation approaches will be required to support diagnosis and assessment of the algorithms, training data, and operational concepts supporting teaming. An appendix describes how these challenges will manifest themselves throughout the acquisition life-cycle.

15. SUBJECT TERMS

Artificial Intelligence; Autonomy; CONOPS; Experimentation; Test and Evaluation; V&V

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Patrick Clancy
Uncl.	Uncl.	Uncl.	SAR	48	19b. TELEPHONE NUMBER *(include area code)* 571-372-4145

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