APPLIED RESEARCH LABORATORY FOR
INTELLIGENCE
AND SECURITY

7005 52nd Avenue, College Park, Maryland 20742
The University Affiliated Research Center
for the Human Domain

"Should you Rely on That AI?" A Report From the January 2021 Workshop to Explore the Multi-Domain Challenge of AI Operational Testing

Nine Key Recommendations for Our Two Most Critical Challenges, and One Outstanding Gap

Executive Summary

On January 28, 2021, the University of Maryland, Applied Research Laboratory for Intelligence and Security (ARLIS) held a workshop to explore the multi-domain challenge of operational testing of Artificial Intelligence (AI). The workshop's objective was to bring together leading thinkers in operational test and evaluation, to identify the most critical challenges in the test and evaluation of AI and Autonomy and explore potential solutions to them. Workshop participants contended that the greatest challenge for development of AI and autonomous systems was a lack of consideration for 'the human element of a mission-critical system.' Repeatedly, participants referred to the human operator as 'the most important asset in the human machine system', realizing that for some, the goal of AI is to replace human effort. Workshop panelists stressed that removing the human from the loop would reduce the efficacy of the system. Instead, participants pointed out that when it comes to the development of AI systems, consideration of the human is often an afterthought, or worse, seen as a factor that limits performance of the machine. This paper builds on those discussions to further discussion on operational testing for Artificial Intelligence, Autonomy, and Augmentation (AAA)[1] technologies and tools, by identifying two key challenges, nine recommendations for them, and one outstanding gap that requires further investigation.

Today's national security threats are increasing in number, complexity, and subtlety, with our nation's adversaries relying less on kinetic warfare and more on cyber and information operations. Adapting to this changing landscape will require the speed and agility that can only be realized within a new paradigm of human-machine teaming. For this reason, the United States Department of Defense (DoD) is investing heavily in the design and development of AAA technology to support the warfighter.

Maximizing the efficacy of human-machine teams requires adapting how AAA technologies are tested and evaluated, certified, developed and acquired for operational use. This includes moving

[1] Coats, D. and S. Gordon. 2019. "*The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines.*" Office of the Director of National Intelligence. January 16. https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf

from a test mindset wherein systems are tested against functional requirements, to a readiness mindset, wherein the human and machine are evaluated as a system, and the system's performance is assessed within the operational context. Critically, evolution of TEVV will require full lifecycle development and new analysis standards that meet operational needs where they are, when they occur.

If the US is to have competitive advantage in the development, adoption, and deployment of AAA technologies, then system procurement, design, Test, Evaluation, Validation and Verification (TEVV), and sustainment must be considered as sociotechnical issues, wherein the human is as critical a component as the machine or algorithms. At the acquisition phase, this means defining hard-to-articulate, non-functional requirements that are key to mission success, e.g., usability, flexibility, and resiliency. Digital tools must be complementary, by design, to teams and developed to support operator workflows by scaffolding interpretation of mission-critical information and decision-making. By design, tools can and should take advantage of the diversity and complexity of human behavior and creativity. Tools should be evaluated in their utility to support operator cognition continuously throughout design and development allowing them to evolve with socio technical and system architectures throughout their lifecycles. These gaps are addressable through both cultural and technical innovation--moving from a test mindset to a readiness mindset, meeting dynamic operator needs rather than requiring operators to meet the needs of autonomous components.

Our workshop resulted in a set of recommendations to enable developers and test engineers to answer the three most critical user-related questions:

- **Is this tool useful?** Does it solve the right problems at the right time for the right people?
- **Is it usable?** Does it alleviate or create cognitive friction by way of operators using it?
- **Is it being used appropriately?** Are operators willing to leverage the power of AI? Are they failing to utilize what is available? Or, equally problematic, are operators relying on it in ways that are inappropriate?

Workshop Summary

In January 2021, the University of Maryland's Applied Research Laboratory for Intelligence and Security (ARLIS), a university affiliated research center, held a workshop to gather industry and academic partners to foster meaningful discussion on how to best improve operational testing methods to determine mission effectiveness complex machine systems. This document serves to highlight the key insights and further conversation to answering these hard questions, which are at the core of our mission, in hopes of charting a path forward. The workshop explored the potential methods to enable a full lifetime testing approach to AI and autonomous systems. Participants were industry professionals and academics, and was moderated by Dr. Brian Pierce, Prof. Adam Porter, and Dr. Craig Lawrence.

"Once you get the edge on T&E, V&V, you have the edge on AI"

(Quote from Greg Zacharaias, PANEL 1: Role of Simulation, Test, Training, Qualifications, Assurance Cases in Operational Testing)

What do we mean by Operational Test and Evaluation for AAA?

In the context of United States DoD systems, Operational Test and Evaluation (OT&E) is a fielded test activity, under realistic, mission-relevant conditions, of any item or component of a weapons system, equipment, or munitions for the purposes of determining its operational effectiveness and operational suitability for combat.[2] This paper will refer to them simply as Test, Evaluation, Verification, and Validation or TEVV.

To articulate what is needed for TEVV for AAA technologies, we must first define what we mean by these technologies. Artificial Intelligence is a set of techniques and methods that can enable a machine to reason about data or context. Machine learning is a subset of AI methods that looks for patterns in data. As such, the patterns that are identified, which may be present in the data set, may not be relevant to the question being asked or may reflect bias of the data set or of weights of the algorithm. Similar coded biases may arise across other forms of artificial intelligence.

This paper is focused on autonomous systems and AI; while the concept of Augmentation is part of AAA, it is not a focus of this paper; however, we will use the acronym AAA to refer to the combined concepts of AI and autonomous systems for simplicity. Augmentation is any technology that enhances human performance, usually either cognitively or physically. We define an autonomous system as a machine that can act with independence and initiative. Autonomous systems do not necessarily require artificial intelligence, nor are all artificial intelligence systems enabled with the capability of autonomy (initiative or independence). Clough (2012) argues that “many stupid things are quite autonomous (bacteria) and many very smart things are not.”[3]

Specific to AAA technology, we wish to assess how well the human and machine system together perform a task and develop subtasks to reach mission goals. We want to know whether these technologies solve operator needs, and to describe the utility of the system and quantify its impact to their performance in terms of time on task or improved likelihood of mission success. Operational T&E must evolve to assist decision makers to understand whether a given technology is worth investing in, and what sort of features or improvements might make a system more useful or usable. This evolution reduces the risk that expensive development will go to waste because operators will not adopt the technology. Finally, OT&E helps to quantify the costs as well as the benefits of technology enhancements, enhancing the cost/benefit analysis.

Challenge: Evolving Test, Evaluation, Verification and Validation (TEVV)

Methods that worked for standard systems won’t work for advanced AAA technologies. Let us begin by considering TEVV. The adoption of AI into mission critical systems poses unique and often high-stakes challenges for safety and security; therefore, TEVV has occupied a critical role

[2] Title 10, United States Code (U.S.C.) § 139 Defines Operational Test and Evaluation as:

> **(i)** the field test, under realistic combat conditions, of any item of (or key component of) weapons, equipment, or munitions for the purpose of determining the effectiveness and suitability of the weapons, equipment, or munitions for use in combat by typical military users; and
> **(ii)** the evaluation of the results of such a test.

[3] Clough, B. T. (2002). *Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway*. Air Force Research Lab Wright-Patterson AFB OH.

in system development in part because it is a standard practice for deployment of any system. Test, Evaluation, Verification and Validation is straightforward for systems whose output and context are predefined and predictable. However, standard TEVV methods do not appropriately transfer to non-deterministic systems operating with dynamic information in unstructured environments. Standard approaches to TEVV are especially limited when applied to machine learning or AI, given their internal complexities and huge state spaces. How is it possible to test what cannot yet be explained or fully observed?

Further, standard TEVV practices leave a gap between functional requirements and the relationships between those measures and mission success. Frequently these methods are not capable of properly determining the potential risks posed by AAA systems, or to anticipate when design outcomes may have unintended and dangerous consequences. TEVV practices also tend to focus entirely on the technology, discounting the interaction between the human and technology, which is often only considered in terms of 'usability' and even then, as an afterthought. Finally, standard TEVV methods do not address questions that arise from the ambiguity of control and responsibility that AI and autonomous technologies can create.

Development of TEVV methods appropriate for assessing the sociotechnical system consisting of autonomous systems, AI, and the human will require dedicated effort. Because it is impossible to fully test all cases, ARLIS proposes moving from a "big data" mindset to a "smart data" mindset, which could help mitigate the realities of near-infinite state spaces and finite testing time. A "smart data" mindset promotes gathering test data strategically and looking for anomalies. Specifically, collecting information that indicates changes in the environment can offer greater insight and practicability, enabling testing authorities to make stronger and broader conclusions with less data collection. In a 'smart data' mindset, simulation becomes key, as it allows for the articulation of the state space, and a bounding of both common and edge cases before implementation and deployment. Definition of these boundary conditions aids in defining high fidelity use contexts in which a system might be deployed, to characterize, bound, and determine the probability of normal and off-normal situations where accidents are more probable.

It is necessary however to be aware that the operational context changes; technology will continue to advance, motivated by need and modernization. For this reason, continuous testing and development is required to learn from and adapt to what is learned in continuous testing and dynamic environments. It is also necessary to point out that more critical and complex tasks and contexts, especially those that pose risk to human life, require more rigorous testing. Below we provide an overview of recommendations to evolve TEVV to enable the inclusion of these '-ilities' (which are covered in recommendation 2) and support the performance assessment of sociotechnical systems. Below we make a set of concrete recommendations for changes to TEVV practices.

Recommendation 1: A System of Systems Approach to TEVV

Test, evaluation, verification, and validation must be predicated on the concept of the human as one aspect of the human-machine system, which itself is made up of several, equally important systems. A system, by definition, is made of multiple components, which as we stated earlier, includes the human user. To neglect consideration of the human in this system is to insufficiently evaluate the performance of the entire system. Testing methods that focus on measuring the speed and performance of only machine components will fall short if the role, capabilities, needs, and skills of the human operator are not included in this calculus: the human is an essential part of the system and contributor to mission success. Measures of performance should consider individual performance of the components, how well they integrate with each other and how these features and their implementation integrate with the human user.

The federal government has recently initiated a push to incorporate AAA capabilities to increase the efficiency of government agencies. As a result, AI is being explored for tasks and systems for which AI has never been integrated. At best, this can lead to awkward system development; at worst it can lead to misapplication of AI to problems or poor solutions. A system of systems approach to TEVV will also help to address the problem of integrating AI into systems that were developed without it. By taking a more holistic, ecosystem-level approach to TEVV, integration can be examined not merely based on the single technology, but the system-of-systems can be evaluated as a whole. Using this method, a gap assessment can be performed, and the most appropriate techniques can be leveraged for the problems that are most difficult for human operators.

Recommendation 2: Measure Not What Is Easy to Measure, Rather, What Should Be Measured.

Algorithmic speed and precision are important measurements, but they are not always (or often) the most relevant to the mission success and system performance. For AAA systems, the most important metrics are often the hardest to quantify. What "operational impact" means should include non-functional requirements, which are much harder to measure than speed and precision.[4] Measures selected for TEVV should reflect performance features that drive mission success and technology adoption. Commonly referred to as the "-ilities," there exists a set of system qualities which are separate from specific mission or functional objectives, that tend to focus on "how well" a system performs in terms of an ability rather than specific functional performance.[5] We acknowledge that these are not typically agreed upon or prioritized by all stakeholders and have yet to be clearly defined. However, we argue that "operational readiness" must be redefined to include these measures, and TEVV especially must be re-imagined including these measures, such as operator utility. Examples of these '-ilitiies' include characteristics such as flexibility, maintainability, survivability, utility, and usability, which drive not only technology adoption and trust, but ultimately human-machine system performance and mission success. Inclusion of these '-ilities' will shift the definition of the human-machine system performance from simple measures, such as the speed of tools and precision and accuracy of computer vision algorithms toward a measure of a system's appropriateness for the tasks at hand, its fitment within a workflow, and whether or not operators find it to be useful, usable and able to be trusted.

Metrics for testing of complex AAA systems must define utility and performance in terms of human needs and probability of mission success, rather than what is easy to measure or typically measured. Operational testing must use relevant, reliable, replicable, and repeatable metrics to measure mission success with operational users in the operational environment. Evaluation based on operational needs means that we must understand the operator, their processes, and their knowledge, skills, and abilities; this understanding needs to be based on knowledge of human cognition and memory, not only as individuals, but also in teams. For each new application of autonomy or AI, the first phase of operational testing must be to define these as requirements at the beginning of the acquisition process. This requires developing an understanding of an operator

[4] Flournoy, M., Haines, A., & Chefitz, G. (2020)
[5] Douglas, Andre. (2021)

not only in terms of their workflow within a team, but individually and at a cognitive level. From here, tests and methods can be developed to allow performers or testing authorities to demonstrate measures of compliance.

Recommendation 3: Lifecycle TEVV and User Research

Historically, deployment of AI expected that learning in AI algorithms would be disabled before deployment, reducing variability, and eliminating the potential that the AI will learn new unexpected and potentially incorrect behaviors during deployment (although this may be changing). To fully enable the potential of AI, techniques to allow systems will be able to learn from operator inputs, preferences, and changing information, which will allow AI to become true team members. For any system to have and to convey full situational awareness, lifelong learning is a requirement especially when temporal features are involved. For this reason, machine learning systems that are continuously improving cannot be said to be complete as soon as it is deployed, with its algorithms frozen. This desire for lifelong learning requires lifelong testing. Such a system will continue to evolve, and so must be capable of internal exploration, relying not only on obtaining new information through active sensory collection, but also deriving new information from speculation and "what-iffing", creating projections of not only future events, but the consequences of actions. Lifelong learning and the paired lifelong test cycle also involve writing operational testability into systems at the highest level, utilizing a comprehensive suite of test cases and developing an understanding of the relationships between them.[6]

If learning is continuous throughout deployment for an AI or autonomous system, then the ownership of development must shift accordingly. Typically, the contractors that develop military systems are responsible for their performance only until the systems are fielded. At that point, it is the military department or different contractors who are responsible for maintaining them. While this makes sense for static deliverables such as submarines that are "complete" at a discrete point in time, a system that is continuously evolving has no such "final delivery". The management of complex and uncertain learning systems will require a lifetime development and testing approach, starting in the requirements gathering and acquisition process phases, and including monitoring and testing throughout development, deployment, and sustainment phases post-deployment.

Current system development and procurement methods for the DoD are not yet agile, but aspire to be, one recent and notable exception being the Air Force's agile software development effort, Kessel Run. The hallmark of agile development is the ability for development teams to adapt to new information during the design process or after initial deployment. In this way, continuous operator feedback can be conducted during the design process. New technologies can be utilized, and changes to the dynamic environment can be addressed as they are discovered. This creates resiliency within our processes, and it saves money and time in costly re-designs down the line.

Recommendation 4: Match the Context of Test to Context of Use

Research and testing must be done in such a way as to develop an understanding of the operational context. It also involves writing operational testability into systems at the highest level, which requires utilizing a comprehensive suite of test cases and developing an understanding of the relationships between them. These test cases must address all critical aspects of the operational

[6] DARPA programs, such as Assured Autonomy and Lifelong Learning Machines (L2M), seek to address the problem of operational testability at the beginning of the acquisitions cycle.

context to include workflow analysis and an understanding of individuals and teams their interactions and interdependencies, and areas of responsibility. Development of test cases and match to operational contexts should leverage cognitive science, to develop an understanding of how users, novices and trained experts make decisions and balance risk and what biases may be present that impact decision making and technology adoption.

Matching the context of the test to the context of operational use can be accomplished several ways. Operational testing and research can be performed by using technology in cooperative field exercises or wargaming efforts. Leveraging passive data collection methods within applications as they are fielded is one way to gather information of a system in real-time, which is especially attractive as it does not disturb or disrupt the operator. Finally, simulation can be leveraged to gain insight into the interactions between operators and these technologies.

Recommendation 5: Leverage Simulation to Model Users in Operational Context

Simulation is frequently used in system development to assess multiple solutions quickly and inexpensively within the problem space. Standard TEVV methodologies rely on data from performance with operators. We argue that simulation can be an effective and important tool for testing, expanding on our current base of formal verification. However, significantly more research and work are needed here.

Development of simulations of operators or “digital twins" has potential to reduce the cost and time involved in TEVV. A “digital twin” may be an ergonomic simulation of the physical characteristics of a human, However, advances in cognitive architectures hold promise for enabling simulation and modeling of human decision making. This type of “digital twin” could allow researchers to model operator behaviors in observed and suspected areas of difficulty or complex, critical systems, without risk to human operators. It has the potential to allow for an early and meaningful exploration of usability and utility, to quantify the return on investment that a proposed solution might offer, even when operators are in short supply. Simulation of human performance and behavior could permit the exploration of edge cases to find the limits of operational feasibility and appropriate constraints for system use, sometimes called assurance cases. It allows for the ability to study emerging behaviors and explore possibilities more easily and less disruptively than with the actual system. In this way, simulation can allow for evaluators to test a system and offer insight into the operational impact of system changes before they reach the field and allows for exploration of options until a desired behavior is achieved. Simulation should never completely replace operator testing. Yet use of simulation could enable developers to leverage unobtrusive sensors and measures in fielded systems to allow for real time and continuous feedback from operators, which can then be fed back into models for continuous improvement.

Challenge: Trust and Adoption for AAA

Failure of operator adoption for such costly military systems, given the tremendous cost of development of AI, can be catastrophic, or simply wasteful of time, energy, or other resources. DARPA has made multiple attempts to use "AI" to automate planning of air operations (e.g., JAGUAR, RSPACE), but the automated planning tools failed to be adopted by the actual program of record: the AOC Weapons System. Anecdotally, among the chief concerns was that the operators did not understand or "trust" plans that were generated by the AI. Currently, a critical focus of the Air Combat Evolutions (ACE) program is an objective measurement of human trust in the AI; a base assumption of the program is that trust in the AI is a key element of technology adoption and distrust leads to technology abandonment.

We define trust as a willingness to be vulnerable to the unpredicted actions of another entity.[7] With respect to AI and autonomy, it means that the machine element of the system is enabled to take initiative and act independently of the human. This may be because the action taken by the machine is too complicated for the human to understand or simply faster than the human can act. Trust is among the greatest drivers in adoption of AAA technologies. Trust is theorized to have many drivers, such as predisposition to trust, experience, learning and memory, and affective and cultural elements. Over- and under-reliance on AI can be derived from a misperception of AI system performance, its vulnerability to attack, AI misunderstanding operator cues, or a user's assumptions of the system's developers and their motivations. Indeed, there are two aspects that need to be considered: trust and trustworthiness. Trust indicates whether a user is willing to be vulnerable to the actions of the system (if given a choice), and trustworthiness is related to the extent the system is reliable and predictable. Miscalibrated trust can manifest as under-trust, where operators fail to utilize technology, or as over-trust, wherein operators rely on the system when it may be within its failure bounds. Both are equally problematic. Ethical considerations for AI also factor into this trust calculus and require deep thought about liability and agency for the use of these systems as well as consideration of ethics and liability in the collection, storage, and use of information including personal identifying information (PII). Below we list a set of recommendations which have implications for the development of trust between human and machine systems, and ethical considerations in the deployment of these advanced systems.

Recommendation 6: A Process to Determine the Right Roles for AI

Realizing the full benefits of human machine teaming involves determining the right allocation of tasks by determining which tasks are best suited for humans vs. machines, and where each might be co-contributors to problem solving. Opportunity exists wherein AAA technologies can reduce an operator's cognitive load by integrating multiple streams of intelligence by sorting data and captured materials, creating links between pieces of information, and offering prediction and other decision support methods. AI also has the potential to reduce physical risk, by classifying objects at standoff and close range, as well as improving logistics and equipment maintenance processes. AI may improve readiness by assisting with recruiting and HR by identifying early what makes a successful operator, pilot, or submariner.

Operator work can be tedious at times. For instance, traditional practices for intelligence gathering include slow, methodical bottom-driven targeting. Adoption and use of technology have already changed the speed of the fight, requiring individuals to do more and to have a greater impact while combatting new threats that include culture, information, social media, and cyber-attacks. Knowing that we need to rely on new tactics and advancements, we must ask ourselves how AI might aid operators, and what capabilities would enable a machine to act as a collaborative teammate. (Attribute to Darsie Rogers). Consideration of the appropriate role for AI must go beyond the standard "machine does what the machine is good at, and human does what the human is good at." The ideal, then, is for AI to augment human cognition, improving decision quality,

[7] Lee & See, 2004; Colquitt, Scott, & LePine, 2007; Marble & Greenberg, in press.

reducing decision time, and curating information as it comes in, but allowing humans to always retain agency and responsibility.

An important step when determining the right role for AAA is determining the acceptable level of risk for a given application, and whether the introduction of such technology is appropriate given that constraint. Determination of roles should further consider not just what the human can be trained to do, but how adoption of AI or autonomy changes how the task can be performed. The constraints on how the task can be performed without AI or autonomy may not apply with these systems. Further, developers must consider how inclusion of AI and autonomous systems shifts workload within the task. Often, when autonomous systems are incorporated into a task, the human user's workload increases as they are not supervising task performance, correcting the action of the autonomous systems, and performing additional tasks that they are now assumed to have capacity to perform. Users and developers need to see that AI is not a magic wand, which cannot in and of itself be expected to "fix" a broken system or process.

Recommendation 7: Set Expectations Appropriately

Effective communication and consistency in setting expectations with sponsors, developers, and users alike is critical for trust development. It is important to clearly convey what AI is capable of and in what contexts. It may be more important that the user understand where a given capability still has room for improvement or how trustworthy the system may be for the current use context. Miscalibration of trust leads to abuse, misuse, and disuse. Developing a shared understanding of the strengths and limitations of AAA technologies will help reduce the potential for miscalibration of trust, which can lead to technologies being abandoned.

One of the most important expectations to set with sponsors and users is that AI is not a magic wand and cannot in and of itself be expected to "fix" a broken system or process. Clearly elucidating system boundaries and capabilities increases the resilience of the human-machine system to mistakes and accidents, which are, on some level, inevitable. At present, these systems are inherently brittle, sometimes with solutions that work within a lab environment or even a single context failing to translate to theatres, situations, goals, or landscapes. Even if the limits of a system are conveyed and clearly understood, it is likely that a user will knowingly or unknowingly push the technology past its safe performance limits. Properly setting expectations communicates that accidents are expected, and it includes protocols necessary for recovery if the system is pushed past its operational boundaries or is simply not behaving properly.

Recommendation 8: User Centered Design

Developing systems that can and should be trusted require user and stakeholder engagement from the beginning. An understanding of the mission context and goals, operator workflows, existing systems, and domain knowledge are underappreciated contextual requirements for developing AI that meets operator needs.

A seamless user interface is critical for development of a usable AI. Such an interface may include searchable information, transparency in decision provenance, and insight into the outcomes of decisions made by the system. A sense of familiarity with the system can be created by various means, to include modeling interfaces after existing operator protocols, using common operator language in the interface, designing specifically against operator goals, and projecting the outcomes of decisions to the operator before actions are committed. Creating interactions that create a sense of predictability, reliability, and consideration for the end-user all work together to engender trust.

Recommendation 9: Certification and Assurance Cases

Certification may contribute to the perceived development of trust and public perception. Certification necessitates the development of common frameworks and standards for evaluation, as well as priorities for those that fund and develop these technologies. There exist numerous challenges for certifying systems that are or contain adaptive computing environments. Current procurement and certification methods are incongruent with the challenges to certifying AI. For instance, certification for AAA will not have the same considerations as certification for an ordnance, however. While some lessons learned from software engineering for obtaining authority to operate can be applied to AAA systems.

Important questions loom over the idea of certifying such systems. Certification for static systems, such as an ordnance, involves knowing what the specific outcome of the test should be. If the results of the test indicate a deviation from the expected or desired behavior, then the system fails the test. Learning approaches that adapt and are dynamic are impossible to certify using these conventions and methods. Certifying completely dynamic systems would have to be limited to the contexts in which the system was tested and verified, because it may be impossible to precompute all possible system configurations.

Certification also typically verifies that a system cannot move into unstable, incorrect, or unsafe configurations during operation. This is increasingly challenging for black-box systems One potential solution could be to constrain the scope of AI to alleviate this issue. A common argument is not to deploy systems that are in 'learning mode' so that their behavior does not evolve past the point of deployment. It can be imagined that at some point, a system that learns and adapts has surpassed the original certification or changed to the point where it is no longer valid. However, this limitation is one that evolution of TEVV methods may overcome. For this reason, it is especially important to reduce risk for high stakes use cases for implementing AI. Instances where lives are on the line such to include combat or HADR response scenarios will require a greater level of testing than implementations for HR, for instance.

Workshop participants recognized the gap between current TEVV practices and those that would be required for AAA systems, articulating methods that could be employed to solve these issues. Other barriers to entry for AAA technologies that were addressed include the generation of trust in AAA technologies. An especially critical and unsolved gap remains in the development of ethical standards and practices for emergent technologies and their use.

Outstanding Gap: Ethics for Development and Deployment of AAA Systems.

Increasing machine initiative, independence, and intelligence in the human-machine team leads to questions of ethics as well as questions of liability. Machines can and will act faster than a human can. If that action leads to unintended consequences, it is not clear who would be responsible, the human teamed with the machine or the developer who created the behavior, the industry or entity that deployed the machine in that context, or even the machine itself. As machine independence, initiative, and intelligence increase, it becomes more likely that a machine may lead performance of all or some part of a task. This type of machine leadership or even shared responsibility raises questions of liability. More importantly, it emphasizes the need for discussion and significant research on how to enable machines to reason about the impact of changes in context on human priorities and human mores. There is a transitive property of the perception of ethics, wherein a person will ascribe to the machine the intentions they ascribe to the technology developer or sponsor. Since public perception of the appropriateness of the actions of the technology or even the developers and sponsors of it can impact trust, public perception cannot be ignored.

Further ethical concerns include data collection and access to data. Many agencies cannot or do not wish to share data, due to privacy concerns. These considerations have the potential to limit the ability of these agencies to leverage the potential of AAA technologies, unless solutions (such as shareable algorithms) are explored. The opaqueness of algorithms can lead to bias in algorithms, which in turn can have significant impact on the suitability, even the fairness, of the output of these systems.[8] This 'coded bias' can be a result of the algorithm or of the data used to train the machine. Transparency of complex algorithms, wherein a human supervisor can see and understand the basis of the output, may not be possible simply because the data set itself is too complex for the human to process. Other solutions to ensure fairness and equity of the outputs of machine algorithms must be explored. A multifaceted approach that leverages lifecycle testing, comprehension requirements, and independent oversight has great potential.

Finally, solutions to the questions of ethical behavior of AAA must address issues of scale: What may work for a small application may not be possible for widespread systems. We argue that ethics for AAA systems must be 'pushed left', that is, considered at the outset of system development, not at the time of deployment, and different aspects of ethics visited and revisited at every step of development. While we admit that ethical considerations comprise a critical and present need for deployment and development, the instantiation of methods to enable ethical response of AAA requires significant further work. We argue that development of the readiness mindset, which embraces lifelong testing, has potential for insight and solution to some of these issues.

Conclusion: Moving from a "Test Mindset" to a "Readiness Mindset"

This report articulates workshop recommendations for changes in TEVV, design, and acquisition processes that are required for systems and tools that leverage AAA technologies. We argue for a readiness mindset that focuses on developing and measuring what is meaningful in the operational setting. It also means establishing the context of use, determining the risk of application within this context, measuring the relevant operator metrics, and presenting it in a way that is understandable, meaningful, and actionable to the operator. Similarly, we have discussed how such a readiness mindset may give insight into developing system characteristics that support and enable human trust of AAA. Finally, we covered critical gaps and questions surrounding ethics for AAA systems and proposed a set of issues to be addressed.

We thank our participants, and we hope to invite them back for a follow-on workshop to discuss ethical standards and development of AAA Technologies.

[8] DARPA's Explainable AI (XAI) program aims to produce more explainable machine learning methods to enable operators to understand, trust, and manage these systems.

About the Applied Research Laboratory for Intelligence and Security (ARLIS)

Applied Research Laboratory for Intelligence and Security (ARLIS) is a UARC based at the University of Maryland College Park and established in 2018 under the auspices of the OUSD(I&S). ARLIS is intended as a long-term strategic asset for research and development in artificial intelligence, information engineering, acquisition security, and social systems. One of only 14 designated United States Department of Defense (DoD) UARCs in the nation, ARLIS conducts both classified and unclassified research spanning from basic to applied system development and works to serve the U.S. Government as an independent and objective trusted agent.

ARLIS is driving the development of a new paradigm of operational testing for AAA technologies. We at ARLIS recognize that our people are our greatest competitive strength in National Security. We argue that to increase the utility, effectiveness, resilience, and to decrease costs associated with development, autonomous machines and AI should be developed by, with, and for collaboration with the users. With this mindset, ARLIS demonstrates its role as the “UARC for the Human Domain.” Critically, the human domain is all tasks that involve people as producers, consumers, or supervisors. As a trusted government partner in Test and Evaluation, Verification and Validation (TEVV) for AAA, ARLIS does practical and translational research to help groundbreaking technologies in complex, critical systems and to ensure adoption, and full realization of the potential of AAA.

Bibliography

Clough, B. T. (2002). *Metrics, Schmetrics! How the heck do you determine a UAV's autonomy anyway*. Air Force Research Lab Wright-Patterson AFB OH

Coats, D. and Gordon, S. (2019). *The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines.* Office of the Director of National Intelligence. January 16. https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, Trustworthiness, and Trust Propensity: A Meta-analytic Test of Their Unique Relationships with Risk Taking and Job Performance. *Journal of Applied Psychology*, 92(4), 909.

Douglas, A. (2021). *Designing Complex Adaptive Systems Using a Stakeholder-Driven and Goal-Focused Framework.* George Washington University, ProQuest Dissertations Publishing, pp. 1–191.

Flournoy, M., Haines, A., & Chefitz, G. (2020). Building Trust Through Testing. https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

Marble, J. L., Greenberg, A. M., Bonny, J.W., Kain, S. M., Scott, B. J., Hughes, I. M. & Luongo, Mary E. (in press). Chapter 13. Platforms for Assessing Relationships: Trust with Near Ecologically-valid Risk, and Team Interaction. To appear in Lawless, W.F., Llinas, J., Sofge, D.A. & Mittu, R. *Engineering artificially intelligent systems*. LNCS Springer.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human factors,* 46(1), 50-80.

Department of Defense. (2020, September 9). *DoD Directive 5000.01, The Defense Acquisition System*.

Appendix A. "Should You Rely on that AI" Panelists

PANEL 1: ROLE OF SIMULATION, TEST, TRAINING, QUALIFICATIONS, ASSURANCE CASES IN OPERATIONAL TESTING	
Moderated by: **Dr. Brian Pierce**	Former Director (and Deputy Director), Information Innovation Office, former Deputy Director, Strategic Technology Office, Defense Advanced Research Projects Agency (DARPA/I2O, DARPA/STO); Visiting Research Scientist, Applied Research Laboratory for Intelligence & Security
Prof. John Dickerson	Assistant Professor, Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland; Chief Scientist, ArthurAI
Dr. Sandeep Neema	Program Manager, Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/I2O)
Lt General (ret) Darsie Rogers	Former Deputy Director, Defense Threat Reduction Agency and Commander, Special Operations Command for CENTCOM; Professor of the Practice, Applied Research Laboratory for Intelligence & Security, University of Maryland
Prof. Hava Siegelmann	Professor, Computer Science, Neuroscience and Behavior Program, University of Massachusetts; Former Program Manager, Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/I2O)
Dr. Greg Zacharias	Chief Scientist, Operational Test and Evaluation, Office of the Secretary of Defense
PANEL 2: MOVING TO A FULL-LIFETIME TESTING APPROACH	
Moderated by: ***Prof. Adam Porter***	Scientific Director and Executive Director Fraunhofer USA Center MidAtlantic; Professor, Computer Science, UM Institute for Advanced Computing Studies and the Institute for Systems Research, University of Maryland Affiliate, Applied Research Laboratory for Intelligence & Security
Prof. Jeffrey Herrmann	Professor, Mechanical Engineering, Institute for Systems Research, Center for Risk and Reliability, and Maryland Robotics Center, University of Maryland
Dr. Mikael Lindvall	Technology director, Fraunhofer USA center Mid-Atlantic
Prof. Douglas Schmidt	Cornelius Vanderbilt Professor of Engineering (Computer Science), Associate Provost of Research, and Data Science Institute Co-Director, Vanderbilt University
PANEL 3: A NEW LOOK AT POLICY, STANDARDS, AND REQUIREMENTS SPECIFICATION	
Moderated by: ***Dr. Craig Lawrence***	Director of Systems Research, Applied Research Laboratory for Intelligence and Security and Visiting Research Scientist, Institute for Systems Research, Clark School of Engineering, University of Maryland
Dr. Chad Bieber	Director, Test and Evaluation. Project Maven, Johns Hopkins University Applied Physics Laboratory
Lt General (ret) Edward "Ed" Cardon	Former Commander of the Second United States Army/United States Army Cyber Command; Former Director of the United States Army Office of Business Transformation Professor of the Practice, Applied Research Laboratory for Intelligence & Security, University of Maryland
Prof. Michael Horowitz	Richard Perry Professor of Political Science, Director, Perry World House, University of Pennsylvania
Dr. Jane Pinelis	Chief, Test and Evaluation of AI/ML, Joint Artificial Intelligence Center
Prof. Ben Shneiderman	Professor Emeritus, Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS); Founding Director, Human-Computer Interaction Lab; Affiliate, Institute for Systems Research and College of Information Studies, University of Maryland