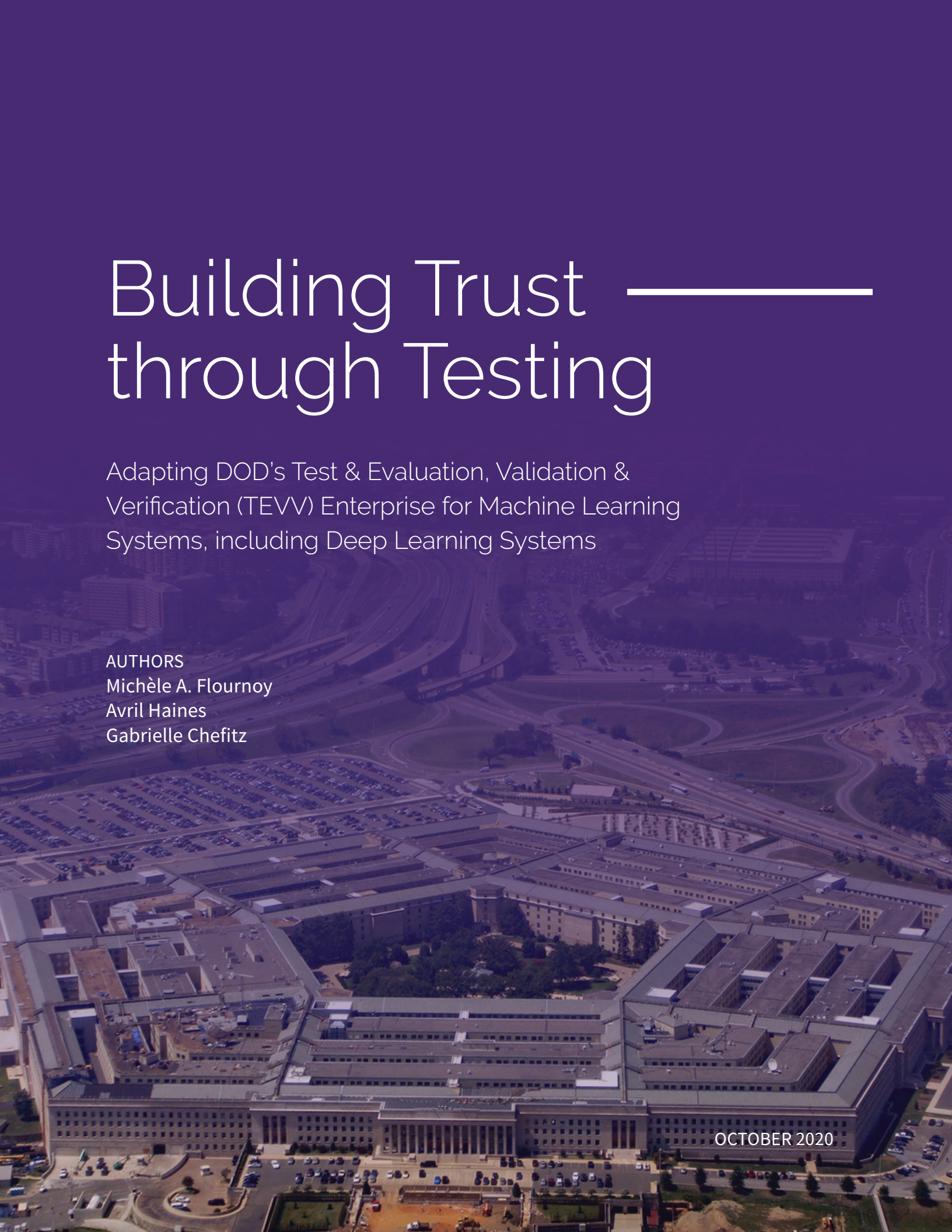
Building Trust through Testing

Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems

AUTHORS
Michèle A. Flournoy
Avril Haines
Gabrielle Chefitz



OCTOBER 2020

PRINT AND ELECTRONIC DISTRIBUTION RIGHTS

Cover image: mariordo59/Wikimedia. https://commons.wikimedia.org/wiki/File:Aerial_view_of_the_Pentagon,_Arlington,_VA_(38285035892).jpg

Building Trust through Testing

Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems

AUTHORS
Michèle A. Flournoy
Avril Haines
Gabrielle Chefitz

OCTOBER 2020

The United States is at an inflection point in an age of mounting transnational threats, unprecedented global interdependence, and resurgent great power competition. This moment is taking place in the context of a technological revolution that exacerbates the challenges we face while simultaneously offering potential solutions, providing breakthroughs in climate, medicine, communications, transportation, intelligence, and many other fields. Many of these breakthroughs will come through the exploitation of artificial intelligence (AI) and its related technologies—chief among them machine learning (ML). These advances will likely shape the economic and military balance of power among nations and the future of work, wealth, and inequality within them.

Innovations in ML have the potential to transform fundamentally how the U.S. military fights, and how the Department of Defense (DOD) operates. Machine learning applications can increase the speed and quality of human decision-making on the battlefield, enable human-machine teaming to maximize performance and minimize the risk to soldiers, and greatly improve the accuracy and speed of analysis that relies on very large data sets. ML can also strengthen the United States' ability to defend its networks against cyberattacks at machine speeds and has the power to automate critical components of labor-intensive enterprise functions, such as predictive maintenance and personnel management.

Advances in AI and machine learning are not the sole province of the United States, however. Indeed, U.S. global leadership in AI remains in doubt in the face of an aggressive Chinese challenge in the field. Numerous DOD and academic reports reflect on the need to invest more in AI

research and development, train and recruit a skilled workforce, and promote an international environment supportive of American AI innovation—all while promoting safety, security, privacy, and ethical development and use. However, far too little attention is placed on the issue of trust, and especially testing, evaluation, verification, and validation (TEVV) of these systems. Building a robust testing and evaluation ecosystem is a critical component of harnessing this technology responsibly, reliably, and urgently. Failure to do so will mean falling behind.

This report will first highlight the technological and organizational barriers to adapting DOD's existing TEVV ecosystem for AI-enabled systems, with a particular emphasis on ML and its associated techniques of deep learning (DL), which we predict will be critical to future deterrence and warfighting while presenting unique challenges in terms of explainability, governability, traceability, and trust. Second, this report will offer concrete, actionable recommendations to DOD leadership, working with the intelligence community, the State Department, Congress, industry, and academia on how to advance the TEVV system for ML/DL by reforming processes, policy, and organizational structures, while investing in research, infrastructure, and personnel. These recommendations are based on the authors' decades of experience working in the U.S. government on national security and dozens of interviews with experts from government, industry, and academia working on ML/DL and test and evaluation.

New Technologies Require New Testing Approaches

The Defense Department needs to reform its existing testing and verification system—its methods, processes, infrastructure, and workforce—in order to help decision-makers and operators understand and manage the risks of developing, producing, operating, and sustaining AI-enabled systems. Several DOD reports and policy documents identify TEVV as a barrier to AI adoption and call for increased research into new methodologies, including the Pentagon's AI Ethics Principles[1] and AI Strategy,[2] which states, "we will invest in the research and development of AI systems that are resilient, robust, reliable, and secure; we will continue to fund research into techniques that produce more explainable AI; and we will pioneer approaches for AI test, evaluation, verification, and validation."

However, DOD has yet to translate this stated goal into a real plan of action. Advancing the Defense Department's TEVV enterprise for ML/DL systems is critical for several reasons.

First, developing an effective TEVV approach that is sufficiently predictive of performance is critical to building the trust in these systems necessary to deploy and leverage these capabilities at scale. The United States has already seen this dynamic with nuclear power, for example, where lost trust in the technology has prevented policymakers from harnessing nuclear power for clean energy.

The Pentagon cannot let TEVV become a barrier to fielding AI-enabled systems in an operationally relevant time frame, but must do so in a manner that engenders trust in such systems and is consistent with U.S. values and principles. The ultimate goal of any TEVV system

should be to build trust—with a commander who is responsible for deploying a system and an operator who will decide whether to delegate a task to such system—by providing relevant, easily understandable data to inform decision-making.

Fielding AI systems before our competitors may not matter if DOD systems are brittle and break in an operational environment, are easily manipulated, or operators consequently lose faith in them. Military operations present a challenging environment. The Defense Department needs ML/DL systems that are robust and secure. They need to be able to function in a range of environmental conditions, against adversaries who are adaptive and clever, and in a manner that engenders trust by the warfighter.

Second, the context in which DOD operates means these technologies are prone to adversary attack and system failure, with very real consequences. Machine learning systems have an increased potential for failure modes relative to other systems, such as bias due to a distribution shift in data, as well as novel vulnerabilities to attacks ranging from data poisoning to adversarial attacks. One could easily imagine an image classifier that accidentally classifies a civilian school bus as a tank or an adversary exfiltrating a model processing sensitive intelligence, surveillance, and reconnaissance or communications data. Image classification algorithms developed for one environment (e.g., the desert) could turn out to work incorrectly in another environment (e.g., cities).

Third, with an effective TEVV system, the United States can reduce barriers to innovation and facilitate U.S. leadership in ML/DL technologies. As most of the innovation in ML/DL will come from the private sector, unless the U.S. government is able to effectively draw on private sector work in this arena, it will not be able to leverage the best cutting-edge technology. Research on new TEVV methods and organizational reforms to adapt the current system is simply not keeping pace with private sector development. Without urgent reforms and prioritized investment in new research and infrastructure, the Defense Department will lose its chance to shape industry's approach to ML/DL development in a manner consistent with DOD standards for safety, reliability, and accountability. It will lose the opportunity to take advantage of new private sector developments, while allowing other nations without such standards to adopt the latest innovations. It is critical that the U.S. government not only shape its own U.S. industry standards but also promote compatible global standards and norms.

Fourth, adversary advancements will likely increase pressure to field AI-enabled systems faster, even if testing and assurance are lacking. China has elevated AI to be a major national priority across sectors and is already exporting armed drones with varying degrees of autonomy.[3] Russia is also pursuing R&D on AI for military

purposes[4] and fields AI-enabled robotic systems in Syria with little regard for ethical considerations.[5] However, it shouldn't be a race against our competitors to field AI systems at any cost. It's a race to field robust, lawful, and secure AI systems that can be trusted to perform as intended.

Finally, high standards for robustness, assurance, interpretability, and governability can ultimately be a tremendous source of strategic advantage, incentivizing industry to harden systems to adversary attack.

Taken together, these risks and opportunities suggest that devising an effective, efficient, and ethical TEVV process is critical for maintaining the U.S. military and economic competitive edge, as well as deploying reliable and trustworthy ML/DL systems.

REPORT SCOPE

Many of the recommendations in this report may apply to several critical emerging technologies beyond ML/DL. Nonetheless, this report will focus on addressing challenges associated with developing the necessary operational and organizational infrastructure to advance TEVV for ML/DL, with a special focus on DL.

For the purposes of this report, we will define AI, ML, and DL using DOD terminology. The Department's 2019 AI Strategy defines AI as "the ability of machines to perform tasks that normally require human intelligence."[6] The Defense Innovation Board defines machine learning as the capability of machines to learn from data without being explicitly programmed.[7] The data used to train machines comes in three general types: supervised (uses example data that has been labeled by human "supervisors"), unsupervised (uses data but doesn't require labels for the data), and reinforcement learning (has autonomous AI agents that gather their own data and improve based on their trial and error interaction with the environment).[8] Finally, deep learning is a special form of ML that deploys neural networks with many layers of connected neurons in sequence. Neural networks are a specific category of algorithms very loosely inspired by biological neurons in the brain.

Deep learning is a potentially powerful tool with important operational and organizational applications that is expected to be increasingly deployed in DOD systems. However, it has two challenging features. First, its output often lacks explainability and traceability. Second, it is vulnerable to adversarial attacks. These defining features mean DL will not be appropriate for all problem sets, and may not be trusted in certain operational settings. Likely, most real-world AI systems will be hybrid models that increasingly use deep learning as a component of a larger system. While it is still early days in the development and testing of even basic ML applications, the processes and procedures established today will set the standard for more complex, future applications, including those that employ deep learning,

which is why it is so critical that the Department develop a framework for addressing these challenges now.

Second, this report will focus predominately on Department of Defense efforts, which we assess currently lag behind those of the intelligence community in developing and fielding these systems. Nonetheless, we acknowledge that many deep learning applications will be essential for the intelligence community, and that the Pentagon already leverages tools developed within the IC. Successfully researching, developing, testing, and fielding these systems will require closer coordination between DOD and the IC.

Finally, this landscape will continue to evolve rapidly as the technology matures. This report identifies a number of challenges and opportunities, but we acknowledge that many more will come to light as the research progresses and more ML/DL systems become operational.

The Challenge: Technological Features and Bureaucratic Barriers

The current rigid, sequential development and testing process for major defense acquisition programs—such as hardware-intensive systems like ships, airplanes, or tanks—is not well suited for adaptive emerging technologies like ML/DL. The current technology acquisition process takes a linear, waterfall approach to development and testing. Companies must pass through a series of acquisition phases and milestone decision points—moving from prototyping/technology maturation to manufacturing and development to production and deployment. At the outset of a program, a test and evaluation master plan is developed, which describes T&E activities over a program's life cycle—including developmental test and evaluation (DT&E), operational test and evaluation (OT&E), and potentially live-fire test and evaluation (LFT&E) at different phases—and identifies evaluation criteria for the testers.

This approach is not well suited for ML/DL, which requires a more agile, iterative development and testing approach. With ML/DL systems, development is never really finished, so neither is testing. Further, ML/DL system performance is difficult to characterize and bind, and the brittleness of such systems means they will require regular system updates and testing. Exhaustive up-front testing does not make sense for these types of non-determinative systems. Therefore, the Defense Department must embrace the commercial best practice of Development, Security, and Operations (DevSecOps), a collection of processes, principles, and technologies that enables an integrated and automated approach to development and testing.[9]

Below we will first outline a few of the key technological features of ML/DL that make TEVV so challenging and require the Defense Department to reimagine its approach. Then we will discuss the organizational and institutional barriers to adapting DOD's TEVV approach.

TECHNOLOGICAL FEATURES

ML/DL systems are not robust and it is difficult to characterize system performance.

A fundamental challenge of ML/DL is its brittleness; it has trouble functioning correctly if the inputs or environmental conditions change. However, testing these systems in all possible scenarios and with all ranges of inputs is simply not feasible. It is nearly impossible to predict all the ways a system could break, or an adversary could manipulate or spoof it, which is partly why these systems may be especially vulnerable to adversarial attacks.

As these systems become more sophisticated, as is the case with deep learning, their output becomes even less transparent, making it harder to determine the conditions under which they might fail and what steps could correct system behavior. Even when operating under the best conditions (within the same distribution of inputs or environmental conditions present during training), DL models generally don't work to the reliability standards needed for safety-critical systems,[10] and therefore may not be appropriate for these applications, at least given the current state of the technology.

ML/DL systems are also particularly vulnerable to operational edge cases (both unintended and intended), which are cases that occur beyond the bounds of a system's operational envelope or normal operating parameters. Because it is very difficult to characterize the actual performance envelop of these systems, it will be important to prioritize stress testing system performance with boundary conditions.

Further, interactions within and between systems (including foreign autonomous systems) can induce unintended consequences and are even more complex to predict or understand. The potential for unintended engagement or escalation is even greater when U.S. and/or adversary systems have the sorts of advanced autonomy features that deep learning can enable, and their interaction cannot be studied or fully tested in advance of deployment.

All of these challenges undermine the establishment of trust between operator and system, which is essential given the U.S. military is likely to deploy DL as part of human-machine teams. Critical to building this trust will be the ability to accurately characterize the bounds of a system's behavior—that is, when it will work and when it will not. If DOD has an image classifier that only works in a desert environment and operators know it will only work in a desert environment, then they are more

likely to trust it. Operators don't need to know exactly how a system works, only under what conditions it will and won't work. This will require new methods of testing and assurance to predict system failure and govern system performance.

Testing ML/DL requires large, representative data sets.

While technological advances in "one shot" and reinforcement learning may ultimately enable the Pentagon to test ML/DL without a lot of data or provide alternative approaches to handle out-of-distribution situations, for the next five to 10 years, the Defense Department will likely rely on supervised learning systems, and testing ML/DL systems will likely require large sets of labeled, representative data.

The United States needs a whole-of-government data strategy that allows for data collection, cleaning, curation, and sharing across agencies, especially between DOD and the IC.

Currently, the Defense Department lacks sufficient available data that mimics the conflict condition in which these systems may operate in the future. This will limit its ability to test system performance against realistic conditions. It will also hamstring efforts to identify edge cases and develop fail-safe mechanisms to prevent catastrophic outcomes.

The Pentagon lacks the ability to effectively collect, manage, store, and share testing data across the enterprise, which would enable this approach to scale. Finally, DOD leadership will need approaches to continuously test the quality of the data itself, as testing data could be compromised or revealed unintentionally or intentionally by adversaries.

ML/DL will be integrated into a system of systems.

ML/DL will be integrated into a range of DOD software and hardware systems, so it is imperative that developers, testers, and policymakers take a systems architecture view when building and evaluating these systems.

The Defense Department cannot simply test all components separately and assume that the system as a whole will work as intended. The accuracy and precision of ML/DL systems is typically a composite effect that arises from a combination of the behaviors of different components, such as the training data, the learning program, and even the learning framework. These components are then embedded in larger systems, so interactions with the physical, computational, and human components of the system will ultimately affect system performance.

Often, failures come from unexpected interactions or relationships between systems, rather than the behavior of any individual element. These dynamics make the system increasingly vulnerable to malfunction and cyber-attacks. An adversary could attack any number of vulnerable entry points within the hardware or software that could, in turn, compromise the entire system.[11]

The Defense Department needs to greatly advance its ability to conduct integrated systems testing that takes into account the interactions with and between systems, testing both machine-machine and human-machine interactions. It should also prioritize testing for how failure in a given subsystem could impact the performance of the system as a whole.

The black box challenge: unique features of DL traceability and interpretability.

It is critical that all ML/DL systems are trustworthy, traceable, and transparent to the greatest extent possible. Deep learning presents challenges for each of these features. Unlike most previous types of computer systems, it may not be possible to trace why a deep learning system made the decision it did in a particular scenario. Not being able to determine what led to an error can obviously create significant challenges for TEVV. It can also undermine user confidence in any solution devised to address the problem identified. Challenges with interpretability in real-time will also hamper human-machine teaming—operators are more likely to trust a system and interact with it effectively if they understand roughly why it is taking certain actions or decisions.

Further, the opacity of deep learning systems makes it difficult to identify or trace back certain kinds of adversary attacks, such as some forms of data poisoning. Some forms of attacks are not obvious to human intuition and, therefore, difficult to imagine and test against (i.e., a 3-D printed turtle that fooled Google's image classifier into classifying it as a rifle.)[12]

The lack of interpretability, traceability, and explainability of DL systems has the potential to undermine trust and exacerbate challenges associated with developing, deploying, and governing ML/DL at scale.

BUREAUCRATIC BARRIERS

In addition to these technological features, there are a number of bureaucratic barriers—ranging from leadership and process to human capital and infrastructure—preventing DOD from accelerating the development of new approaches to TEVV for ML/DL.

Responsibility for ML/DL TEVV is shared and not well coordinated.

While responsibility for TEVV is shared across multiple parts of the Office of the Secretary of Defense (OSD) and the services, greater coordination is needed to streamline investment and R&D on new testing approaches, increase cross-program visibility, and proliferate standards and best practices.

There is a growing community of stakeholders within DOD and the broader U.S. government that will be critical to adapting the ML/DL TEVV enterprise. The Director of Operational Test & Evaluation (DOT&E) oversees policy and procedure for operational testing of major defense acquisition programs (MDAPs). DOT&E can play a key role in promulgating testing standards but tends to be cautious in setting new standards. It is accustomed to a rigid, sequential TEVV process that works well for MDAPs, but not for emerging technologies like ML/DL. The Testing Resource Management Center (TRMC) oversees infrastructure and spending, and develops investment roadmaps for new technology programs. TRMC will also be critical to adapting infrastructure for ML/DL DevSecOps. TRMC has included the Autonomy and Artificial Intelligence Test Technology Area in its T&E/S&T portfolio.[13]

The Joint Artificial Intelligence Center (JAIC), the Under Secretary of Defense (USD) for Research and Engineering (R&E), and the Director of the Defense Advanced Research Projects Agency (DARPA) all have important roles to play in the development of AI TEVV metrics, methods, and standards for DOD systems. The JAIC is actively engaged in setting standards, sharing best practices, and conducting testing. These programs promote, for example, designing DOD ML/DL systems and tagging data in ways that make it possible to understand how any particular decision is made. In April 2020, JAIC issued a request for information for new T&E capabilities for AI technologies.[14] It is also already leading on implementation of the Defense Department's AI ethics principles and integration of TEVV throughout the product development life cycle. The JAIC has established a DOD-wide responsible AI subcommittee, with representation from the services and the Joint Staff, DARPA, R&E, T&E, A&S, Policy, and the Office of the General Counsel to develop detailed policy documents, which will map the AI principles to the AI product life cycle and acquisition process.[15] However, the JAIC is too small to scale these solutions throughout the Department. Meanwhile, the USD R&E is responsible for prototyping systems and developing large system of systems that will increasingly be AI-enabled. Finally, DARPA's Explainable AI program is working to produce more explainable models that facilitate trust and human-machine collaboration.

The armed services each have their own AI programs, which include testing components and research on AI TEVV at the service labs. The services have the operational knowledge and program acquisition offices and have traditionally led on developmental testing for major programs of record. They also know there is power in owning the test data and, understandably, want to evaluate the capabilities they are sending to their servicemembers themselves. However, the services don't tend to have the S&T expertise and personnel to develop new approaches to TEVV for ML/DL.

The Defense Department will need to designate an office or organization with overall responsibility for the TEVV process and establish a coordination mechanism

that leverages the unique value-add of each of these entities, breaks down bureaucratic siloes, and streamlines investment in research and infrastructure to support new TEVV approaches.

DOD policy, standards, and metrics for testing performance and evaluating risk need to evolve.

DOD needs a policy framework for determining safety standards for a range of ML/DL applications based on the use case, mission, and anticipated environment in which the system will operate. These standards then need to be translated into requirements for system design and metrics for measuring system performance that are operationally-relevant; transparent to developers, testers, and users; and reflect DoD's AI ethics and U.S. values.

DOD will first need to establish a testing framework that provides guidance on the degree of acceptable risk and limits for a given ML/DL use case based on a potential range of outcomes and errors. For example, if a potential outcome has lethal consequences, the acceptable risk is likely to be extremely low, whereas if the outcome has no clear negative consequences, the acceptable risk will almost certainly be higher. The risk of fielding these systems will also need to be weighed against the risk associated with not adopting the system. For example, a 5 percent error rate may be palatable if the existing system has a 10 percent error rate. These risk and error rates will also need to incorporate the potential for adversary attacks or interactions with adversary systems. For example, an error that happens .001 percent of the time naturally, but which an adversary is able to consistently exploit, could create significant challenges for the Pentagon.

Further, policymakers must acknowledge that with technology, there might be less margin for error than with humans, and less clarity about who is accountable for such errors. For example, the United States may determine that as a society, we are not willing to accept a scenario in which an algorithmic error in an autonomous vehicle causes a loss of life even if it saves thousands of lives overall. Ultimately, these technologies will never be perfect, and testing to a near-perfect standard will inhibit DOD's ability to field these systems at all. Therefore, it needs a dedicated process to develop policies to determine how much risk it is willing to accept in a given case, weighing operational need and potential consequences against DOD ethics, principles, and policies.

DOD will need to translate this testing and safety framework into functional, specific requirements language. For example, the JAIC could put out a request for proposal saying it needs a DL that can identify a target from X range, in this season, in these weather conditions.

Finally, DOD will then need metrics and methods to evaluate operational performance in easily understandable, operationally relevant terms. For example, if U.S. Special Operations Command uses a deep learning algorithm to translate documents from a raid on a terrorist compound and finds time-sensitive information, how do you measure operational impact? Determining impact isn't just about statistical analysis on the level of precision-recall, but the impact compared to a human being's ability and the efficiency created for the operator.

DOD lacks an iterative, continuous approach to development, testing, and sustainment that bridges the gap between acquisition and T&E.

For ML/DL, the Defense Department will need to replace its classic approach to TEVV of formulating a T&E Master Plan for a given capability up front with a more automated, iterative, and continuous approach to testing in line with DevSecOps. Assuring that ML/DL systems function as expected and do not engage in behaviors outside their intended use and operational parameters will require testing across the system's entire life cycle—from development to operational deployment to sustainment.
It will also require new methods for capturing lessons learned and integrating these into iterative development and testing. Because of the difficulty predicting and binding system performance, one should consider every deployment of an ML/DL system as an experiment and opportunity to collect data and insight on performance.[16]

To support this approach, DOD will need to expand coordination between program managers and testers to ensure testing milestones are built in throughout the acquisition program. Program managers often see TEVV as an obstacle to be surmounted at the end of the development process, rather than a necessary process to be integrated throughout the development life cycle. Of course, this problem is not unique to ML/DL programs, but it is exacerbated when it comes to emerging technologies that do not yet have established testing methodologies. Further, the Pentagon needs to invest in and scale automated TEVV capabilities for operational platforms, such as the Navy's automatic test and retest program,[17] which will significantly speed up the testing process.

An agile approach of iterative testing, updates, and releases will place significant burdens on TEVV and require infrastructure and research investments, as well as incentivizing program managers to see testing as an integral part of the development process rather than a barrier. Program managers should be responsible and rewarded for delivering a well-functioning product, not just staying on budget and schedule.

Current TEVV methods and infrastructure aren't well suited for ML/DL and may require new funding approaches.

Adapting the TEVV enterprise for ML/DL will require targeted investment in developing new testing methods and adapting current testing infrastructure to support DevSecOps and iterative testing. The Defense Department needs new approaches, such as automated testing and digital twinning,[18] as well as new testing infrastructure, including test beds, test ranges, and advanced modeling and simulation (M&S). DOD also needs computing support, cloud-based resources, data capture for continuous development, and generation and use of synthetic data, particularly for DL applications.[19] Finally, it needs tools for traceability that capture key information about the systems development and deployment to inform follow-on development, testing, and use.

The JAIC has adopted commercial best practices for AI DevSecOps. Its Joint Common Foundation (JCF)—an infrastructure environment designed specifically for training, testing, and transitioning AI technologies, which is intended for use by all the services—is an important down payment on these efforts that will make it easier to secure and rapidly test and authorize AI capabilities.[20] The JAIC should be given the resources and top-cover it needs to scale this effort. The Pentagon should build on the JCF and other efforts to promote a secure, cloud-based DevSecOps ecosystem that facilitates the rapid commercial development and iterative testing of ML/DL and the proliferation of testing tools, data, and standards across OSD and the services.

The Defense Department also needs to increase resources, bandwidth, and personnel dedicated to adversarial testing. It can and does use Federally Funded Research and Development Centers (FFRDCs), but there is concern among some experts that it is too heavily reliant on just one—MITRE—for adversarial testing. DOD needs to invest in creating a catalogue of adversarial testing tools and proliferate these capabilities across the service labs and FFRDCs that support testing. Finally, DoD needs to work more closely with the intelligence community to simulate realistic threats.

Finally, new approaches to TEVV for ML/DL will require new funding approaches. DoD, in coordination with Congress, should consider new approaches that incorporate T&E funding into the cost of development, given that TEVV must be integrated into an iterative development process. DoD and Congress should also consider establishing a new appropriations category that allows AI/ML to be funded as a single budget item, with no separation between RDT&E, production, and sustainment, as recommended by the Defense Innovation Board Software Acquisition and Practices Study.[21]

DoD lacks the ability to recruit, train, and retain the right talent.

For many organizations within the DoD TEVV ecosystem, recruiting and retaining talent is often a bigger challenge than securing funding. These organizations need diverse, interdisciplinary teams that understand both testing and the technology itself. DoD needs data scientists, statisticians, and computer scientists that can develop new testing and verification mechanisms; computer science and ML/DL experts to develop the technology; and operators that understand the technology enough to trust, deploy, and integrate it operationally. Finally, it needs experts in human cognition and psychology that understand human-machine interaction and can build interfaces that enable greater trust.

Many of the challenges of recruiting and retaining such technical talent are not unique. Existing DoD programs to recruit recent science, technology, engineering, and math graduates are too small, non-traditional hiring authorities for STEM talent are underutilized, and the service academies do not feed enough STEM talent directly into technical roles. DoD lacks dedicated career paths for technologists and testers, which further constrains the Department's ability to retain what talent it does manage to recruit or grow in-house.

Not all of this talent needs to be cutting-edge researchers; the Department will need a cadre of professionals—program managers, requirements writers, lawyers, operators, policy officials, and others—who have a baseline understanding of the technology and testing procedures, and can bridge the gap between DoD leadership and policy teams on one hand and the technical developers and testers on the other. Further, DoD should leverage its expansive network of FFRDCs and academic partnerships to expand its access to technical personnel. Many of the FFRDCs, such as the Lawrence Berkeley and Lawrence Livermore National Laboratories in California and the MIT Lincoln Laboratory in Cambridge, are located near hotspots for AI talent and have fewer challenges with hiring.

The Department has developed policy and ethical guidance on autonomous systems and AI, but these guidelines have yet to be translated into TEVV implementation guidance.

DoD has established important foundational policy guidance for the use of autonomous systems and artificial intelligence with DoD Directive 3000.09 on Autonomy in Weapon Systems and the Defense Innovation Board's AI Ethics principles, adopted by DoD in February 2020. These policy documents have important implications for testing and evaluating of ML/DL. For example, 3000.09 calls for systems to go through "rigorous hardware and software [verification and validation] and realistic system developmental and operational T&E, including analysis of unanticipated emergent behavior resulting from the effects of complex oper-

ational environments on autonomous or semiautonomous systems." It also states that interfaces should be "readily understandable to trained operators,"[22] making explainability an important component of implementing this policy. Meanwhile, the AI Ethics Principles commit DoD to develop and deploy AI that is traceable (including with transparent and auditable methodologies, data sources, and design procedure and documentation), reliable (explicit, well-defined uses, with the safety, security, and effectiveness of such capabilities subject to testing and assurance across their entire life cycle), and governable (with the ability to detect and avoid unintended consequences, as well as disengage or deactivate deployed systems that demonstrate unintended behavior).

These policies are an important start and provide a useful framework for driving TEVV for AI and autonomous systems. However, these goals are incredibly broad, and many are currently technologically infeasible, given existing testing methodologies. DoD needs to develop TEVV implementation guidance for both 3000.09 and AI ethics principles. In particular, these principles must inform ML/DL design and be incorporated into the standards, specifications, and requirements against which systems will be tested.

Finally, as ML/DL development and testing capabilities are still evolving, policy and implementation guidance should not be overly prescriptive or rigid before DoD knows how these systems will function and in what contexts they will be deployed. Nonetheless, the development of implementation guidance and processes that integrate ethics into the design and testing process will help accelerate the deployment of reliable, safe, and transparent ML/DL.

There is insufficient coordination between DoD, the private sector, and academia.

DoD needs a hybrid approach to TEVV that leverages DoD, academic, and industry research, infrastructure, and talent. The majority of ML/DL innovation will come from the private sector and academia, as will most of the insight into how to test, benchmark, and assure these systems. However, DoD has an important role to play in integrating, scaling, and deploying these solutions. Further, it can dedicate significant resources to basic and applied research and use its market power to influence the development and promotion of national standards for at least certain industries. DoD also has the unique capability to do adversarial testing, with access to threat intelligence and operational knowledge that can inform realistic modeling and simulation. DoD should, therefore, focus on unique use cases where there is no commercial relevance or where sharing the data or algorithm would reveal sensitive or classified information.

DoD should, when possible, leverage commercial TEVV methods and tools, such as Microsoft Azure and Amazon Web Services secure environments and tooling.[23] In many cases, however, industry methods will not be applicable, given the safety-critical application and unique classification of DoD data, requiring a hybrid model of development and testing informed by academic research.

Further, DoD needs to engage the private sector to develop an intellectual property strategy both parties can live with that includes access to sufficient data for continuous testing.

DoD should also engage in a sustained dialogue with commercial developers to inform how DoD defines the requirements for ML/DL testing and performance based on what is technologically feasible, now and in the future. There are some successful models for this cooperation, such as the Army's AI Hub—a consortium of industry, government, and academic partners based at Carnegie Mellon University—which works with the JAIC and other DoD AI entities to provide independent assessments of key research questions.[24] Scaling this effort will require a senior DoD champion, such as the Under Secretary for R&E, who values this work and can promote it across the Department.

Recommendations for Adapting DoD's TEVV Enterprise for AI/ML

1. Create an OSD coordinating body to lead on AI/ML TEVV and incentivize strong cooperation with the services.

Accelerating and streamlining TEVV methods and processes for AI/ML will require greater coordination across the TEVV ecosystem, including the JAIC, USDE (R&E), USD (A&S), TRMC, DOT&E, and the service program offices, test commands, and T&E organizations.

The Director of the JAIC and the Director of Operational Test and Evaluation should co-chair a new AI/ML TEVV Cross-Functional Team (CFT) that reports biannually to the Deputy Secretary's Management Action Group (DMAG) and coordinates AI/ML TEVV research and investment across the Department.[25] This forum would include representation from R&E, DARPA, TRMC, and the service labs, test commands, and T&E organizations, building on the work of the OSD-led Autonomy Community of Interest TEVV group.[26] The CFT would also work with the Defense Science Board and Defense Innovation Board, which would provide expert support and connect DoD TEVV efforts with those in the private sector.[27]

This body would spearhead the development of policy, standards, requirements, and best practices for AI/ML TEVV, which would incorporate the AI ethics principles and 3000.09 and serve as testing implementation guidance. The CFT would be responsible for assessing and certifying the service AI/ML TEVV budgets, just as the Cost Assessment and Program Evaluation (CAPE) office advises the Secretary and Deputy Secretary on the budget.

The CFT would also create an AI/ML T&E action plan to delegate and coordinate responsibilities across the Department. The JAIC should serve as a center of excellence for AI/ML TEVV and lead on the development

of testing tools and a testing framework. Further, R&E should lead an integrated research plan for new TEVV methods, the service labs on modeling and simulation and operational testing, TRMC on infrastructure investment, and DOT&E on policy and standards proliferation and coordination.

DoD would benefit from an expansive testing ecosystem that pulls together OSD standards, policy setting, and best practices with the services' operational knowledge and acquisition infrastructure.

2. Invest in priority areas of research in partnership with industry and academia.

Research on new tools, methodologies, and metrics is key to implementing new ML/DL testing framework and standards. Many of the following recommendations will fail if this one is not successful.

Removing this critical barrier will require coordinating and prioritizing research on the science of ML/DL TEVV, backed by sustained, focused DoD funding. While some TEVV challenges for ML/DL are well understood by industry and DoD simply needs to adopt commercial best practices, there are many problems without existing solutions that DoD has a unique interest in solving given its operational requirements. This work must be ongoing, as the technology evolves and new challenges with ML/DL are identified.

The ML/DL TEVV CFT should task the Defense Science Board to conduct a thorough review of all current research programs for ML/DL TEVV.

Based on this review, the Committee, with strong input from R&E, should develop a coordinated research plan and seek funding for DARPA, TRMC, and the service labs.

1. DoD should prioritize research on automated and repeatable testing. TEVV is currently slowing down development and deployment, as testing processes move much slower than development. Advancing and scaling automated testing could help standardize the testing process, help DoD keep pace with industry, and accelerate fielding and scaling of these systems.
2. DoD should research methods for bounding, governing, and interrupting system performance, including monitoring systems that can detect performance issues and edge testing to prevent unacceptable errors.
3. DoD should research and develop performance metrics in operationally relevant terms that are easily traceable and understandable to the user and support risk assessments. These metrics are key to translating safety and assurance guidance into requirements for various design features, such as explainability and traceability. Such metrics could include whether the

system satisfies specific mission requirements, how it utilizes resources over time during the mission, how a system's output impacts human decision-making, and whether safe actions are selected in the presence of unexpected or hostile inputs. For example, if a certain class of system has an explainability requirement given its context of use, how does one actually characterize levels of explainability?

4. DoD should work closely with industry (particularly the commercial autonomous vehicle community) to continue research on new techniques for synthetic data creation, modeling, and simulation.
5. DoD needs to increase research on a range of issues related to human-machine teaming and interaction. DoD needs a more human-centric approach to considering ML/DL development and testing. Humans are central throughout a system's life cycle, from development to deployment, and DoD will need to account for human psychology and bias at each stage.
6. DoD should expand research on increasing system robustness to overcome adverse conditions or enable systems to withstand or respond when targeted by an adversary attack.[28]
7. DARPA should continue its Explainable AI program, which is important for increasing the transparency and accuracy of ML/DL while strengthening trust with the end user.

As a general matter, investments in the science of ML/DL TEVV are a critical prerequisite to developing the most effective and efficient standards, tools, and methodologies needed to assure system performance. As these technologies and their applications evolve, new areas of research will undoubtedly arise that warrant investment.

3. Develop a tailored, risk-based framework for ML/DL testing and safety.

The AI/ML TEVV CFT should lead on the development of a framework that establishes architecture and testing standards for TEVV. DoD cannot have a one-size-fits-all approach; it needs a flexible testing framework that is mission and use-case dependent.

A DoD-wide testing framework for AI/ML will help shorten the testing cycle and make test results interpretable and comparable across the Department. This framework should also incorporate DoD's legal and ethical requirements, serving as implementation guidance for the AI ethics principles and 3000.09.

The DIB AI ethics principles call on the JAIC to create a taxonomy of DoD use cases of AI, based on their ethical, safety, and legal risk considerations. The CFT

should leverage this taxonomy and develop corresponding testing criteria and safety standards. Testing standards, for levels of interpretability or assurance, for example, should be determined based on several dimensions of risk and performance, including: the likelihood of error detection, the consequence of the error given the complexity of the operating context, the potential for unintended escalation, the size of the attack surface, system performance relative to that of a human operator, impact on human decision-making, and the risk associated with not adopting a system (e.g., the risk may be a 10 percent error for DL, but a 30 percent error without the system). For example, AI for business process automation would likely score low on all of these risk criteria, while AI for critical network cybersecurity would likely score high on all and, therefore, necessitate stricter and more expansive TEVV requirements.

For higher-risk applications, DoD may need to require systems to be designed with fail-safe systems or operated only as part of a human-machine team to help mitigate risks and govern system performance. Researchers are currently developing methods for monitoring system performance and constraining a system to a set of allowable, predictable behaviors and mitigate the risk of failure and unintended escalation.

A DoD-wide AI TEVV framework will help decision-makers manage the tradeoff between the risks of failure and the value of deployment, while advancing the development of clear and consistent requirements for system design and metrics for performance evaluation.

4. Translate the testing framework into testable, verifiable requirements to be used by the private sector and build an integrated team to leverage this approach.

DoD should establish a process for translating the testing framework into testable and verifiable requirements.

DoD requirements would help standardize processes for industry contractors who develop AI for DoD and promote a faster and cheaper TEVV process by enabling the private sector to do some testing throughout development. The development of such requirements and standard processes would allow DoD to leverage the talent and expertise in the private sector, while maintaining DoD's safety and risk standards and employing DoD's operational knowledge and adversarial testing capabilities. Such requirements could additionally serve as the basis for standards that are promoted by the U.S. Government across the private sector and internationally, as discussed in a later recommendation.

To realize this approach, DoD should build integrated, multi-disciplinary teams that reflect the entire development, testing, and sustainment life cycle. One model is

the JAIC's project manager model, which incorporates experts on product, policy, legal, test and evaluation, and requirements into one team. A similar approach is the Navy's AI "DevRon" concept, a single entity accountable start to finish for the life cycle of capability development.[29] Scaling this approach would help ensure requirements take into account the unique challenges of ML/DL TEVV.

Requirements should also be written to advance the integration of ethics into the design and testing process. One case study for how this is already being done is DARPA's Urban Reconnaissance through Supervised Autonomy (URSA) program, which is developing AI-infused drones that can help prevent friendly fire and civilian casualties in urban battles.[30] This program brought in ethicists to anticipate challenges before development even began, with the aim of integrating ethics into the development loop. This URSA program offers useful lessons on how to conduct ethically aligned design of military systems that enables testing for robustness, safety, and security and ensures these systems are reliable and governable.

Developing requirements and standards for ML/DL will help strengthen the link between DoD and industry, ensure a hybrid approach that leverages both private and public sector resources, and inject DoD's safety and ethical requirements into industry development practices.

5. Bridge the gap between development and testing.

ML/DL systems will require testing and verification across their entire life cycle, which will require stronger links between program managers and testers as well as methods to capture lessons learned throughout deployment.

There are already good models for how this could be done. For example, developers for Project Maven need to submit to T&E in every sprint cycle, or they cannot move forward to the next stage of development. We recommend replicating and scaling this approach in other programs.

DoD could also look to the Joint Improvised-Threat Defeat Organization, established in 2006, to rapidly field new technologies to counter IEDs. The organization's approach to fielding prototypes and testing them in the field provides a useful model for the rapid deployment of ML/DL.

Another way to bridge this gap is to leverage the testing framework and requirements language to inform the acquisition process. The JAIC is already using its acquisition process to impact development of AI technologies by using the AI ethical principles as "applicable standards." One could envision a similar process with AI testing and governance standards. To do so, DoD will need to maintain a robust dialogue with industry on what is feasible to help inform the process.

In addition, program managers should be incentivized to build a test program that verifies performance throughout the development and fielding life cycle and

holds developers accountable for meeting the requirements. Program managers should be recognized and rewarded for testing and for overall performance of the system, not just on whether they can meet budget and schedule targets. Delivering a system that is too risky to field but on time and budget does not make sense. Program managers should also be rewarded for canceling programs that do not work and for delivering programs that do.

6. Increase and integrate spending for T&E research and infrastructure.

Advancing TEVV for ML/DL will require a substantial investment in both research and infrastructure. TRMC should lead on assessing current gaps in infrastructure and be given increased funds to invest in service and DoD T&E live, virtual, and constructive (LVC) test ranges, test beds, and modeling and simulation for testing adaptive systems.

DoD should significantly increase investment in modeling, digital twins, and simulation, working with the private sector—particularly commercial autonomous vehicle companies—to implement industry best practices. These technologies can be used to develop representative testing environments and conduct edge testing to determine a system's operational envelope. This investment is also key to reaching the goal of automatic, repeatable testing, which is critical to DevSecOps, and creating synthetic data that can help offset a lack of usable, operational data.

DoD could invest in test beds to be hosted at FFRDCs and university-affiliated research centers, which attract top talent and work with DoD regularly. TRMC should also help scale the Navy's automatic test and retest program, which uses cloud-based digital twins to provide near real-time feedback and automatic testing of thousands of simulated environments.[31] To do so, we support the National Security Commission on AI's recommendation that Congress should raise the authorized cap for laboratory infrastructure investments, currently set at $6 million, in order to provide laboratories with the ability to invest in equipment and testbed infrastructure necessary for robust AI research, prototyping, and testing.[32]

Finally, the Department should consider new approaches to fund AI/ML TEVV. For example, DoD could require that TEVV cost is factored into development, rather than having as a separate T&E item.

Congress could also consider a new type of funding authority that bridges the gap between AI S&T and T&E, allowing for both development and testing of new technology. DoD does not yet have well-established methods of testing for ML/DL, and will therefore be developing the capability and the ability to test it in parallel. This will require S&T dollars for research on new T&E approaches. Congress has already authorized a similar model for cyber, in which funds are authorized for creating, testing, fielding, and operations.[33] The Department, working with Con-

gress, should explore the potential of replicating this model for AI development and testing, consistent with the Defense Innovation Board Software Acquisition and Practices study recommendation for a single budget item for AI/ML.[34]

7. Develop industry/U.S. government TEVV standards and promote them internationally.

DoD, working with the National Institute of Standards and Technology (NIST) and industry, and building on DoD requirements and processes developed for industry, should develop standards for ML/DL testing for the private sector that can be publicly promoted and help inform private sector development of ML/DL systems.

Such standards would focus on a range of issues, including robustness, interpretability, performance metrics, fail-safe design, traceability for data collection and management, privacy, and testability. These should be broad standards that serve as guidance for both government and commercial developers to develop operationally specific design requirements and testing metrics.

DoD entities—including the JAIC and R&E—and the IC are already playing a role in the development of U.S. government AI standards, led by NIST and the Office of Science and Technology Policy. NIST's 2019 plan on U.S. leadership on AI provides an important foundation and calls for the development of standards and metrics for trustworthiness (e.g., accuracy, explainability, resiliency, safety, reliability, objectivity, and security), complexity, domain-specific and context-dependent risk, and uncertainty.[35]

NIST and the State Department should play a leading role in promoting U.S. government standards for AI and ML development and testing domestically and throughout international standards-setting organizations, such as the International Standards Organization and Institute for Electrical and Electronics Engineering, and multilateral institutions, such as the OECD AI Policy Observatory. The promotion of U.S. standards globally will help bolster U.S. economic competitiveness, create a level playing field for U.S. industries who are collaborating with DoD and consequently subject to these standards, incorporate U.S. values and ethical principles into AI and ML development, and ensure that the United States and its allies are interoperable. As we've seen in other critical technology areas, the United States must ensure that competitors do not set standards, which make it harder to manage vulnerabilities and hinder U.S. efforts to establish the highest degree of ethics, safety, and risk management.

8. Test, train, and certify human-machine teams through wargaming, simulation, and experimentation.

ML/DL systems are likely to be deployed in human-machine teams and DoD will need new approaches to test, train, and certify these teams as a whole. DoD will need to understand and address the issues that could arise both when humans are the operators of a system, including issues of handoff, and when humans are part of the operating environment.

Operators, analysts, and commanders will need to understand and trust ML/DL-enabled systems; be trained on how they will impact, enable, or detract from operator capabilities; and how they will contribute to the overall mission.

DoD should develop a plan to better incorporate ML/DL-enabled applications into exercises, wargames, and tabletop exercises. Increased investment in live, virtual, and constructive environments that allow for user interaction and experimentation in realistic, simulated environments will further help with training and validating human-machine teams.

DoD will also need to develop operator training and certification programs for specific systems and use cases, as well as refresher trainings and re-certification for when systems are deployed in new environments or when the system is updated.

DoD will need to consider each real-world deployment as an experiment. Through wargaming, experimentation, and simulation, DoD can create the processes and methodologies to capture lessons learned and data, providing this feedback to developers, testers, and operators to improve systems and build trust.

9. Accelerate recruitment and training of ML/DL and TEVV talent.

Recruiting and retaining diverse, interdisciplinary teams is an essential prerequisite to advancing TEVV for ML/DL systems. DoD needs those with a fundamental academic grounding in test and evaluation, as well as the systems engineers, computer scientists, and ML/DL experts that understand the technology itself. It also needs statisticians, data scientists, and applied mathematicians who can perform mathematical testing. Finally, it needs experts who understand human-machine interaction, such as psychologists and ethicists.

In addition to these subject matter experts, it also needs operators, requirements writers, acquisition professionals, and lawyers who have a basic degree of technological literacy and understand why this technology matters.

To build up this talent, DoD should rely on its vast network of national and service labs, FFRDCs, and university-affiliated research centers for technical talent. DoD could consider designating a current FFRDC for ML/DL or creating a new FFRDC to help focus resources and talent. In addition, DoD should establish a dedicated T&E career path, including education, training, and rotational assignments.[36] The services should develop corresponding programs for technologists, allowing STEM graduates from the service academies and ROTC programs to serve

in assignments where they can leverage AI/ML or other technical expertise. Further, DoD should better leverage existing authorities Congress has provided to attract entry-level tech talent, using incentives such as scholarships and debt forgiveness, and more experienced talent using vehicles like the Highly Qualified Expert and Intergovernmental Personnel Act programs. The JAIC could also develop a best practice guide for recruiting AI talent, from developers to testers.

Further, DoD should expand in-person and virtual technical training across OSD, the services, and other components to bolster technological literacy in the workforce. The Department could tap into an extensive repertoire of technical education now available online, including on machine learning, advancing testing techniques, DevSecOps, and human-machine interaction. DoD could also create a training module in-house, focused specifically on adversarial testing.

10. Increase resources for and attention on adversarial testing and red-teaming.

DoD's unique operational context will require particular emphasis on adversarial testing. Not only will DoD need to worry about adversary aggression, but it must also consider the unintended consequences that may arise when U.S. systems with ML/DL interact with adversary systems that have intelligent and/or autonomous features.

DoD should significantly increase wargaming and red-teaming focused on spoofing ML/DL, drawing on offensive cyber experiences like "Hack the Pentagon" exercises, bug bounties, and NSA R6 (a dedicated red team within NSA's research directorate). DoD will also need to increase funding to fully develop threat models for future battlefield environments involving near-peer adversaries.

DoD and the Office of the Director of National Intelligence (ODNI) could consider standing up a national AI and ML red team as a central hub to test against adversarial attacks, pulling together DoD operators and analysts, AI researchers, T&E, CIA, DIA, NSA, and other IC components, as appropriate. This would be an independent red-teaming organization that would have both the technical and intelligence expertise to mimic realistic adversary attacks in a simulated operational environment.

11. Promote greater cooperation on ML/DL between DoD and the IC.

Development and testing of ML/DL systems would greatly benefit from stronger cooperation between DoD and the IC. The IC is more advanced in terms of development and fielding of ML/DL. It also has more flexible authorities to develop novel applications, conduct testing, and acquire commercial technology. However, IC and DoD cooperation in this arena is rare—the IC does not want to share its

most exquisite capabilities, and the services do not want to deploy systems they have not built themselves.

Both sides stand to benefit from greater cooperation on testing, particularly adversarial testing. If the IC shares some of its capabilities, DoD can add resources to bring them to scale and then collect data it can then share with the IC. If the IC wants to influence industry standards, it can do so far more effectively through DoD.

While DoD and the IC may not be able to share or transfer every system developed, they can share the fact that a system exists and its basic characteristics and capabilities. DoD can then acquire its own version or the IC can use its own capability in support of DoD operations.

We endorse the National Security Commission on AI recommendation that DoD and ODNI stand up a steering committee on emerging technology, tri-chaired by the Deputy Secretary of Defense, the Vice Chairman of the Joint Chiefs of Staff, and the Principal Deputy Director of ODNI.[37] We recommend this committee have a special line of effort dedicated to adversarial testing for emerging technologies, focused on ML/DL.

Greater collaboration will help bolster adversarial testing capabilities with realistic threat modeling; create a division of labor for developing testing infrastructure and methods, reducing costs to both; and strengthen U.S. government coordination to scale development and testing standards.

CONCLUSION

The future of U.S. leadership on ML/DL and DoD's ability to harness these critical technologies depends on DoD investing in the science of ML/DL TEVV to develop new approaches and metrics, as well as standing up the coordination and governance mechanisms to accelerate progress and scale solutions. It will require developing the testing frameworks, requirements, and standards to bridge the gap between industry and government and shape a more iterative development and testing approach; shifting culture and practice toward the testing and certification of human-machine teams; and securing the talent, infrastructure, and resources to implement this new approach. Finally, DoD will need to deepen partnerships with the private sector, academia, non-governmental organizations, international organizations, and international partners to realize a multi-stakeholder approach to ML/DL development, testing, and deployment.

Adapting the TEVV enterprise for ML/DL is critical to increasing trust in and, consequently, accelerating the deployment of these systems on a timeline consistent with the rate of innovation, operational need, and U.S. ethics and principles. The steps DoD and the broader U.S. government take now to adapt the ML/DL testing ecosystem will determine the long-term safety, reliability, and relevance of these systems in the coming decades.

Endnotes

1. U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.
2. U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy* (Washington, DC: Department of Defense, 2018), https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
3. Patrick Tucker, "SecDef: China is Exporting Killer Robots to the Mideast," *Defense One*, November 5, 2019, https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/.
4. Roger McDermott, "Moscow Pursues Artificial Intelligence for Military Application," *Eurasia Daily Monitor* 16, no. 89 (June 2019), http://jamestown.org/program/moscow-pursues-artificial-intelligence-for-military-application/.
5. Margarita Konaev and Samuel Bendett, "Russian AI-Enabled Combat: Coming to a City Near You?" *War on the Rocks*, July 31, 2019, https://warontherocks.com/2019/07/russian-ai-enabled-combat-coming-to-a-city-near-you/.
6. Department of Defense, *Summary of the 2018 Artificial Intelligence Strategy*.
7. Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, DC: Department of Defense, 2019), https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.
8. Greg Allen, "Understanding AI Technology" (Joint Artificial Intelligence Center, April 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf.
9. U.S. Department of Defense, *OSD DevSecOps Best Practice Guide* (Washington, DC: Department of Defense, 2020), https://www.dau.edu/cop/it/DAU%20Sponsored%20Documents/DevSecOps_Whitepaper_v1.0.pdf.
10. Andrew J. Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance," *arXiv*, September 2, 2020, https://arxiv.org/abs/2009.00802.
11. Sven Herpig, "Securing Artificial Intelligence" (Stiftung Neue Verantwortung, October 17, 2019), https://www.stiftung-nv.de/en/node/2650#collapse-newsletter_banner_bottom.
12. James Vincent, "Google's AI thinks this turtle looks like a gun, which is a problem," *The Verge*, November 2, 2017, https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed.
13. Defense Innovation Board, *AI Principles*.
14. Frank Konkel, "Pentagon Needs Tools to Test the Limits of its Artificial Intelligence Projects," *Nextgov*, April 16, 2020, https://www.nextgov.com/emerging-tech/2020/04/pentagon-needs-tools-test-limits-its-artificial-intelligence-projects/164687/.
15. JAIC, "The DoD AI Ethical Principles – Shifting From Principles to Practice," *AI In Defense* (blog) on JAIC, April 1, 2019, https://www.ai.mil/blog_04_01_20-shifting_from_principles_to_practice.html.
16. Interview with Ashley Llorens, Chief of the Intelligent Systems Center at the Johns Hopkins University Applied Physics Laboratory (JHU/APL), July 16, 2020.
17. "Navy SBIR/STTR Success Story," *Innovative Defense Technologies*, https://idtus.com/wp-content/uploads/2020/02/SBIR-Success-Story-NAVSEA-N05-163-IDT-2020-02-04.pdf.

18. Phil Goldstein, "Digital Twin Technology: What Is a Digital Twin, and How Can Agencies Use It?" *FedTech*, January 31, 2019, https://fedtechmagazine.com/article/2019/01/digital-twin-technology-what-digital-twin-and-how-can-agencies-use-it-perfcon.
19. National Security Commission on Artificial Intelligence (NSCAI), *Second Quarter Recommendations* (Washington, DC: National Security Commission on Artificial Intelligence, July 2020), https://drive.google.com/file/d/1LDrd6T7H50ry9uXNA6cwhsrtnpQ63EWH/view.
20. JAIC, "The JAIC Pushes the Envelope with DevSecOps through the Joint Common Foundation," *AI In Defense* (blog) on JAIC, July 16, 2020, https://www.ai.mil/blog_07_16_20-jaic_pushes_the_envelope_with_devsecops_jcf.html.
21. Defense Innovation Board, *Software Acquisition and Practices (SWAP) Study* (Washington, DC: Department of Defense, May 2019), https://media.defense.gov/2019/May/01/2002126693/-1/-1/0/SWAP%20MAIN%20REPORT.PDF.
22. U.S. Department of Defense, *Autonomy in Weapon Systems*, DODI 3000.09, May 8, 2017, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.
23. Defense Innovation Board, *AI Principles*.
24. Gary Sheftick, "AI Task Force taking giant leaps forward," U.S. Army, October 7, 2019, https://www.army.mil/article/225642/ai_task_force_taking_giant_leaps_forward.
25. A HASC subcommittee NDAA draft would elevate the JAIC to report directly to the Deputy Secretary of Defense.
26. Defense Innovation Board, AI Principles.
27. Interview with Bob Work, former Deputy Secretary of Defense. August 17, 2020.
28. NSCAI, *Second Quarter Recommendations*.
29. NSCAI, *Second Quarter Recommendations*.
30. Bartlett Russell, "Urban Reconnaissance through Supervised Autonomy (USRA)," *DARPA*, https://www.darpa.mil/program/urban-reconnaissance-through-supervised-autonomy.
31. "What is ATRT: Automated Test and ReTest?" *Innovative Defense Technologies*, https://idtus.com/atrt-automated-test-and-retest/.
32. NSCAI, *Second Quarter Recommendations*.
33. Mark Pomerleau, "Senate committee wants more cyber pilot programs," *Fifth Domain*, June 11, 2020, https://www.fifthdomain.com/congress/2020/06/11/senate-committee-wants-more-cyber-pilot-programs/.
34. Defense Innovation Board, *Software Is Never Done: Refactoring the Acquisition Code for Competitive Advantage* (Washington, DC: Department of Defense, May 2019), https://media.defense.gov/2019/May/01/2002126690/-1/-1/0/SWAP%20EXECUTIVE%20SUMMARY.PDF.
35. National Institute of Standards and Technology, *U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools* (Washington, DC: Department of Commerce, August 2019), https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
36. Office of the Director, Operational Test and Evaluation, *FY19 Test and Evaluation Resources* (Washington, DC: Office of the Secretary of Defense, January 2020), https://www.dote.osd.mil/Portals/97/pub/reports/FY2019/other/2019teresources.pdf?ver=2020-01-30-115558-813.
37. National Security Commission on Artificial Intelligence (NSCAI), *First Quarter Recommendations* (Washington, DC: National Security Commission on Artificial Intelligence, March 2020), https://drive.google.com/file/d/1wkPh8Gb5drBrKBg6OhGu5oNaTEERbKss/view.

WESTEXEC ADVISORS

westexec.com