AI Engineering for Defense and National Security
A Report from the October 2019 Community of Interest Workshop
ARTIFICIAL INTELLIGENCE HOLDS GREAT PROMISE FOR DEFENSE AND NATIONAL SECURITY, and in many cases AI systems are already being developed and deployed. However, these systems are difficult to specify, build, replicate, verify, validate, and monitor. Whereas fields such as civil, mechanical, electrical, and traditional software engineering have established disciplines for creating safe and reliable structures, machines, programs, and systems, there is not yet an engineering discipline for AI. Without an AI Engineering discipline, we face a chaotic landscape far from the DoD’s vision of safe, ethical, and secure uses and applications of AI.

Carnegie Mellon University’s Software Engineering Institute (SEI), a federally funded research and development center, is leading a movement to create an AI Engineering discipline to enable the United States Department of Defense (DoD) to realize the full benefit of AI for defense and national security and to provide a foundation for creating viable, trusted, and extensible AI systems.

Why AI Engineering?

In 1956, John McCarthy defined AI as “the science and engineering of making intelligent machines.” Today, AI includes modern machine learning as well as knowledge representation, reasoning, heuristic search, planning, and other traditional or future AI techniques (e.g., neuro-symbolic reasoning, reinforcement learning, and meta-learning). To date, most research and development has been devoted to creating AI capabilities, and little progress has been made on the engineering of those capabilities to ensure that they are safe and reliable.

There is much to be reused and adopted from traditional software and systems engineering for the development and deployment of AI systems—systems that include AI components and traditional software components. However, AI extends and challenges traditional software and systems engineering in several important ways.

Currently, data is the lifeblood of modern machine learning techniques and the foundation for many traditional AI techniques. When using modern machine learning techniques, data, data engineering, and data management play an essential role in system development, deployment, management, and evolution. Furthermore, it is the data, rather than a formal or functional specification, that shapes the behavior of systems that use machine learning.

While extremely powerful, modern machine learning and the models they produce are opaque, difficult to interpret, probabilistic, and typically non-deterministic—and they introduce new security and robustness concerns (e.g., adversarial attacks on ML systems). Typical software engineering tools and techniques such as static analysis, dynamic analysis, and reverse engineering do not apply to models created by machine learning algorithms. Similarly, failure modes of all kinds are difficult to isolate or debug in models produced from modern machine learning techniques like deep learning.

Complementary to techniques like deep learning in AI is knowledge representation and reasoning. Knowledge in an AI system is an organized collection of rules, facts, and relationships that can be used to drive reasoning or augment machine learning. Knowledge can be explicit, sometimes transparent, and separate from the inference mechanism in an AI system. Furthermore, knowledge can be hand-coded, learned from data, or based on experiences and observations. Knowledge in an AI system takes on some characteristics like software and some characteristics like data; both types of characteristics require new and different types of management and maintenance practices, especially as both the system and the knowledge the system uses evolve over time. This complexity poses a unique challenge.

Last, and more generally, AI techniques are augmenting humans by addressing tasks that require human attention. These tasks range from mundane or routine tasks to complex decision making under stressful or urgent conditions. Additionally, applications of AI can be significant in certain system deployment scenarios. AI systems require thorough engineering to be safe, ethical, secure, and reliable—especially for defense and national security applications.

Workshop Purpose

In late October of 2019, the SEI convened the first-ever workshop on AI Engineering for Defense and National Security. Bringing together thought leaders in defense and national security, industry, and academia, the workshop laid a foundation for identifying challenges and opportunities for future initiatives. This report identifies recommended areas of focus and describes general discussion topics.

Summary Recommendation

The workshop recognized that there is both a major need and a current gap in the processes, practices, and tools necessary for engineering AI technologies and systems. Further, the workshop confirmed and identified that defense and national security applications of AI create particular challenges and needs beyond what typical commercial applications of AI demand. That finding does not dismiss processes, practices, and tools from commercial capabilities—in fact, there are many lessons and best practices to adopt from what has been learned in commercial applications of
AI—but rather indicates several areas of special focus for the defense and national security community. This leads to the overall recognition and recommendation from the workshop:

Where possible, the DoD and related organizations should identify opportunities to build, share, evolve, and mature processes, practices, tools, and technologies for reliably engineering AI systems.

This overall recommendation is in line with recommendations released by the Defense Innovation Board (DIB) in November 2019: “Cultivate and grow the field of AI engineering. The Office of the Under Secretary for Research and Engineering (OUSD(R&E)) and the Service Labs should support the growth and maturation of the discipline of AI engineering by building on sound engineering practices that DoD has long fostered, engaging the broader AI research community more extensively, providing specific opportunities for early-career researchers, and adapting the Department’s legacy of safety and responsibility to the field of AI to integrate AI technology into larger complex engineered systems.”

The DoD adopted the DIB’s recommendations in February 2020, with the Joint Artificial Intelligence Center (JAIC) leading the implementation of the principles.

**Background and Context**

With the workshop’s focus on understanding and shaping a new discipline of AI Engineering, specifically for defense and national security applications, it was important to have a common understanding of what is meant by the terms AI and engineering. Many definitions exist for each of these terms, and they all convey important meaning.

For the purposes of the workshop and this report, we use the definition of AI given in the Summary of the 2018 Department of Defense Artificial Intelligence Strategy: “AI refers to the ability of machines to perform tasks that normally require human intelligence—for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action—whether digitally or as the smart software behind autonomous physical systems.”

The workshop took a broad understanding of artificial intelligence that spans from the device and systems level, up through data management and machine learning, and on to knowledge representations, reasoning, planning, and autonomy.

We use Mary Shaw’s analysis of historical definitions of engineering and synthesis of those ideas into this definition of engineering: “Creating cost-effective solutions to practical problems by applying scientific knowledge to building things in the service of mankind.”

**Participants**

Workshop participants indicated affiliations that included the following organizations:

- Dr. Peter Santhanam, IBM Research
- Google
- Drew Conway
- Brett Vaughan, Navy Chief AI Officer, SECNAV / OPNAV / ONR
- MIT Lincoln Laboratory, www.ll.mit.edu
- Defense Advanced Research Projects Agency (DARPA)

**Initial Needs and Recommendations**

The topics raised and discussed at the workshop organize around three central themes: Robust and Secure AI, Scalable AI, and Human-Centered AI. While the themes may expand or evolve, they provide an organizational framework for a discipline of AI Engineering. Below we provide a summary of the main ideas behind each theme and highlight several needs specific to defense and national security capabilities that were surfaced at the workshop.

**Robust and Secure AI**

The primary quality used for evaluating AI systems today (most machine learning or deep learning systems) is accuracy. While accuracy is obviously important, the community wanting to develop, adopt, and deploy AI technologies must move beyond just accuracy; and this is especially true for most defense and national security applications of AI. This idea led to many conversations at the workshop about Robust and Secure AI. Here, the quality of robust describes AI systems that continue to operate at expected levels of performance when faced with uncertainty, novelty, or other changes to the operating environment. The quality of security complements the quality of robust and implies free from danger or threat—that is, a secure AI system has mechanisms and mitigations to prevent, avoid, or provide resilience in the face of dangers from a particular threat model. Obviously, security and robustness overlap, but the workshop participants indicated the importance of considering both of these qualities explicitly. And finally, robust and secure implies many other related qualities such as safety, reliability, dependability, and stability.

In discussing the robustness and security of AI—and specifically machine learning—systems, it is important to mention adversarial machine learning. Adversarial machine learning is a field of study where researchers seek to understand both how machine learning models can be attacked and how to defend against those attacks (see, for example, the DARPA GARD program). Here, attack means manipulated in some way to be deceived or to extract information about the model, compromising privacy.
Manipulations of machine learning systems can happen at training time through data poisoning or at inference time through the introduction of deliberate patterns of erroneous data to force misclassification or deception. Similarly, deployed models can be probed to extract both general and specific information about the data they were trained on.

The discussion at the workshop around Robust and Secure AI certainly included adversarial machine learning, but treated the topics of robustness and security in a much broader context.

**Need 1: Tools to Build Robustness and Security into AI Systems**

There are many ways that machine learning and AI systems can fail, and there are many ways that they can be attacked, deceived, or defeated. To use AI capabilities in many defense and national security applications, concerns about robustness and security must be addressed very early in the system lifecycle. To avoid the downsides of system failures or vulnerabilities, it is imperative to consider robustness and security at design and development time—long before the deployment or operational phases of the lifecycle. These lessons are well known from decades of traditional software engineering, and we must not forget them as we develop and deploy new AI systems.

Beyond considering robustness and security early in the lifecycle of AI systems, we need tools, techniques, and methods for actually building robustness and security into AI systems. The field of robust machine learning is nascent and is further challenged by the rapid advancement of the overall field of machine learning. In addition to general approaches to building robustness and security into AI systems, we need specific approaches for making certain learning, reasoning, planning, and other AI algorithms and techniques robust and secure in the face of noise, uncertainty, novelty, and active adversaries.

One possible approach to building more robust AI systems is to appropriately extend, adapt, and enhance Agile and DevOps methodologies to apply to the development of AI systems. Agile and DevOps provide an existing foundation to build upon and should be extended to apply to AI systems by incorporating ideas around data and training of machine learning capabilities, verification and validation of AI systems, and continuous monitoring of AI system behavior. MLOps is a new and expanding approach to developing, deploying, and evolving systems built with machine learning. Fundamentally, the concept of incremental, iterative development with a focus on a functional system that is continuously monitored can provide a process approach to improving robustness and security of an AI system over time. This is also linked to the following need on testing, monitoring, and mitigating AI system robustness.

**Need 2: Tools for Testing, Monitoring, and Assuring AI System Robustness**

The aspiration of building robustness and security into systems is important and worthy, but it is foolish to believe that in the very dynamic field of applied artificial intelligence and machine learning that systems developers will get perfect robustness and security in real-world settings. Using best practices to build robustness and security into AI systems will still leave failure modes and attack vectors that could lead to unintended and undesirable circumstances. As identified above, methods to “build in” robustness and security should be developed and utilized to improve AI systems, but verification and continuous monitoring for system robustness and security are still required.

There is a need for tools to understand system behavior and functionality later in the lifecycle. Testing mechanisms for robustness and security—in addition to normal performance and accuracy testing—are essential during the testing and acceptance phase of an AI system’s lifecycle. Even more complicated for systems that have AI and machine learning components is that these systems must be continuously monitored to see how they are behaving in the “real world.” These systems’ behaviors are often probabilistic/non-deterministic, and they could start to fail in unknown ways based on changes in the environment, the type of data that is being observed, or even manipulation (of data, the environment, or the system) by an adversary.

Finally, beyond testing and continuous monitoring of AI system performance, robustness, and security, there is a need to develop techniques, patterns, and tools for mitigating failures in AI systems. It may be the case that mitigation strategies must be custom designed for specific systems and the context in which they operate, but there are likely general patterns and best practices that systems developers and operators could employ. One such example has been described as algorithmic agility, where a system is deployed with three algorithms running side by side: the current operational version of an algorithm that is being used, the previous stable version of an algorithm used as a fallback if something goes wrong, and the development version of a future algorithm for live-world testing and monitoring. In this setting, the system would be configured to switch between the algorithms based on monitoring and other environment sensing capabilities. The development of other strategies, patterns, techniques, and tools will be essential to the reliable deployment and operations of AI systems and the rapid, iterative, incremental development and deployment processes necessary to mature AI systems over time.
Need 3: Sharing of AI “Incidents”
Incident response, security update processes, and responsible vulnerability disclosure and coordination are mainstay functions of the global software ecosystems. Rich networks of responders, analysts, and coordinators help identify potential problems (e.g., vulnerabilities) and coordinate the response to those problems and the fixing of systems that are susceptible.

As AI technologies and systems are more broadly adopted, the need for coordinating incidents, vulnerabilities, and mitigations related to AI systems needs a similar rich and vibrant infrastructure and ecosystem. However, some systemic challenges must be addressed before this vision can be realized. One major challenge is that AI technology development and system integration remains largely a craft where developers and modelers pick and choose from a large array of available tools, frameworks, and models and then tailor those capabilities for their specific need. It is hard-to-impossible to assign version numbers and track these capabilities and identify where particular problems or vulnerabilities might exist. Another obvious challenge is that the concept of a “patch” is unclear for many types of AI systems. For example, it is unclear what it might mean to patch a particular kind of deep neural network (Is it an incremental update to weights? Is it a full retrain of the entire model? Is it use of a different training algorithm?). Therefore, the mechanism for managing the ecosystem and the coordination processes must be developed and co-evolved.

Scalable AI
For successful adoption of AI technologies across the many and varied needs of the defense and national security community, AI technologies must be scalable. But when it comes to AI technologies, scalable is a multifaceted concept. AI technologies must be able to scale to the size of mission needs and the data that can support those missions. AI technologies must be able to perform at the speed required by mission and operational constraints. AI technologies must be able cope with and leverage the complexity of real mission scenarios and the unique modalities and phenomenologies that defense and national security applications afford. And, of course, while dealing with different dimensions of scalability—size, speed, complexity—AI technologies and systems must remain buildable, deployable, usable, reliable, and trustworthy.

Beyond system scalability concerns, the broad adoption of AI technologies across the defense and national security community also raises issues of realizing AI at enterprise scales, affordable development and acquisition of capabilities, workforce readiness and capacity-building challenges, and ways to democratize the effective development, adoption, and use of AI technologies. Efforts must be taken to address each of these areas as well as solutions that enable mission-motivated system scalability of AI components.

Need 4: Scalable Oversight for Defense Applications
A major difference between most defense and national security applications of AI and most commercial applications of AI is the ability to collect, curate, or generate enough data to support the development of a reliable AI solution. In most commercial applications of machine learning, Internet companies can use the clicks and other interaction patterns of their millions (or billions) of users to create datasets to develop future versions of prediction, detection, or recommendation systems. In most defense and national security applications, having millions or billions of users to “label” datasets through their regular interactions with the system is not possible or feasible. Even giant Internet companies realize that so-called scalable oversight of machine learning systems is a major challenge.

This problem is further compounded in defense and national security application by both the complexity of the operational environment and the exquisite and complex sensing capabilities that are available. There are several approaches to address the challenge of scalable oversight. One possibility is creative use of the defense and national security workforce to encourage labeling of operational data in daily workflows. Over time, this approach could create very useful datasets for training machine learning models or even a knowledge graph to inform reasoning and other AI capabilities. A second is to develop methods and algorithms that require less data to develop system capabilities (e.g., low-shot learning). The latter is an active area of research in the machine learning community. Another approach that can be useful for a variety of problem domains is the development and use of game engines or other simulators to create data, to provide experience and observations, or to allow for exploration of a proxy environment for a real-world problem. Of course there may be other approaches as well. To date, the most common approach is to leverage brute force methods for creating datasets for specific missions or applications. While this solution works—at a significant cost—at the individual application level, it is not a scalable approach that will allow for the broad adoption of AI technologies for both business and mission applications.

Need 5: Data and Model Management and Sharing
Where scalable oversight is about creating the data that feeds and drives AI and machine learning capabilities, there is also the concern over data and model management, reuse, and sharing. One way to scale the adoption of machine learning models is through a process called transfer learning, where a model is trained for a particular problem and then reused—perhaps with some minor retraining—for a different problem. This is a powerful technique, but scaling it across the defense and national security enterprise will require developing and institutionalizing mechanisms for managing, tracking, versioning,
and even analyzing these reused and derivative capabilities. This is also true of data and datasets that can be repurposed across applications. Finally, in many defense applications of AI, data and model management will need to become a warfighter activity as new capabilities become integrated into operational capabilities from the enterprise to the edge.

Also of concern for scalable reuse of data and AI components are patterns for system integration (e.g., application programmer interfaces [APIs], information requirements), composition of AI components, mechanisms for updating shared or reused components, and metadata management for appropriate reuse. Addressing these enterprise-wide concerns early on will significantly aid the successful and scalable adoption of AI technologies. These sorts of tools and administrative mechanisms should be established early in the adoption of AI technologies and iteratively evolved over time to support broad and diverse mission applications.

Defense applications will also require that AI capabilities be deployed in delayed/disconnected, intermittently-connected, low-bandwidth (DIL) environments.

**Need 6: Available, Scalable, and Adaptive AI Computing Infrastructure**

A major enabler of commercial applications of AI and machine learning technologies is the ready availability of massive-scale computing capabilities. To realize broad development and adoption of AI technologies across the defense and national security communities, similar access and use of computational resources will be essential. Wherever possible, shared and reusable computing infrastructure will enable multiple organizations, agencies, and teams to realize AI-enabled capabilities more readily and will most likely limit the cost to realize AI capabilities. It is essential that teams working to develop and deploy AI technologies are not hamstrung by the inability to access and use appropriate computing capabilities and that computing capabilities are managed and shared appropriately and responsibly.

Additionally, some defense and national security applications will have special computing requirements (e.g., edge-enabled capabilities). It is most likely necessary that the requirements of special computing constraints will require the custom advances in resource-constrained machine learning, federated machine learning, edge computing, knowledge representation and reasoning, and other areas of AI that may be currently underserved in both the research and engineering communities. Specialized computing architectures and hardware—if developed and appropriately provisioned—are likely to provide solutions and computational advantages to mitigate many of these challenges.  

**Human-Centered AI**

A central topic of the workshop was the need for an AI-ready culture. This focus on culture includes a variety of aspects such as reducing perceived fear of AI technologies, understanding the behavior and functionality of AI technologies, education and training, ethics and privacy, and focusing on augmenting human capabilities (rather than replacing them). It also focused on more practical considerations to create an AI-ready culture and workforce such as access to computing infrastructure and tools, data preparation and availability, possible changes to acquisition policies and practices, and increasing comfort with uncertainty and risk tolerance.

For defense and national security organizations to successfully adopt AI technologies, these cultural aspects regarding the understanding, comfort with, and adoption of AI technologies must be considered. In some cases, technology concepts and solutions can and will aid in creating an AI-ready culture and the successful adoption and integration of AI technologies into business and mission workflows and capabilities. However, these culture considerations cannot be addressed with technical solutions alone; they require organizational and socio-technical solutions.

**Need 7: An AI-ready Workforce**

There is an obvious need for training and education across the broad and diverse defense and national security workforce. This is a far-reaching problem. Obviously, the acquisition and engineering workforce needs to better understand how to ask for (requirements) and buy (acquisition) AI capabilities. This includes smart and perhaps novel approaches to testing and evaluation (T&E) at both development and operational stages. But the need for education goes well beyond just the engineering and acquisition workforce.

At the leadership and management levels, AI education is necessary to ensure that expectations and future promises of AI capabilities are appropriately managed and messaged. Of course, this does not require a deep technical or algorithmic understanding of AI technologies, but it does require a continuously updated awareness of what is practically possible with current AI technologies and the requisite limitations of those technologies.

The operational community also has to have an understanding of what is possible with AI technologies and what they can and can’t do. A basic understanding of different types of machine learning and the types of capabilities they can support (e.g., decisions, classification, predictions) will greatly improve the adoption and comfortable use of these technologies in operational settings.

Beyond education and creating an AI-ready workforce, there are also organizing principles that can improve the
development, adoption, and deployment of AI capabilities. The Navy has established the “AI DevRon” concept where acquisition experts and transition partners are involved in the full lifecycle of capabilities. Similarly, the Army is using the concept of Tactical Data Teams to capture and pursue near-term operational needs. These models should continue to be explored and evolved over time and used to inform other activities across the community.10

**Need 8: Mechanisms and Frameworks to Enforce Ethical Principles**

Shortly after the workshop, in November 2019, the Defense Innovation Board (DIB) published a recommended set of Principles for the Ethical Use of AI. In February of 2020, the JAIC and DoD officially adopted the Ethical AI Principles recommended by the DIB. These principles—responsible, equitable, traceable, reliable, and governable—are properties and requirements for the ethical development, use, and operation of AI solutions. To take these principles forward, we need tools, mechanisms, and frameworks for the operationalization and enforcement of these principles across the lifecycle of AI capability development and deployment.

Notably, when the DIB recommended the five principles listed above, they also recommended the development and maturation of a discipline of AI Engineering. A sound discipline of AI Engineering will incorporate the necessary tools and mechanisms to ensure the safe and ethical development and use of AI technologies for defense and national security.

**Need 9: Instrumentation, Monitoring, Evidence-Production, and Interpretability**

To provide the ability to enforce ethical principles, AI systems must be instrumented in order to produce telemetry data about their behavior and functions; monitoring systems must be put in place to capture and analyze information from instrumentation; analysis and synthesis techniques must be developed to support evidence production to assure proper function and aid in understanding system output (e.g., decision making and justification for those decisions); and tools and techniques must be developed to provide various levels of interpretability on the behaviors and outputs of AI systems. These tools surrounding the development and operational use of AI technologies are currently overlooked, as the current focus of AI activities is about demonstrating some specific application of AI.

Of course, these tools will support much more than only enforcement of ethical principles, although that is a very important use of these types of tools. In addition, monitoring and interpretability tools will support testing and evaluation, continual verification practices (e.g., to detect when a systems behavior has degraded due to environmental or other circumstances), promoting trust in human-machine teaming.
and overall adoption of AI technologies in both business and mission workflows. At minimum, these types of tools must be considered when adopting and deploying AI technologies. Even better would be the adoption of a small number of general frameworks that provide these types of tools and infrastructure at the enterprise level for broad development and deployment of AI capabilities.

Conclusion
As defense and national security organizations increasingly invest in AI solutions, AI Engineering will enable the DoD to achieve its vision of creating viable, trusted, and extensible AI systems. The path forward for establishing a discipline builds on the foundation in this report and requires the continued collaboration of a varied field of experts. As we find lessons in the experiences of industry, academic, and defense researchers, developers, and implementers, we will capture these lessons and integrate them to create robust and secure, scalable, and human-centered AI systems.
References


About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of art in software engineering and cybersecurity to benefit public interest.

Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612
sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu