National
AI Engineering Initiative

Scalable AI

Carnegie Mellon University
Software Engineering Institute

Scalable AI

Contributors:

Hollen Barmer, Rachel Dzombak, Matt Gaston, Jay Palat, Frank Redner, Tanisha Smith, John Wohlbier

Acknowledgements:

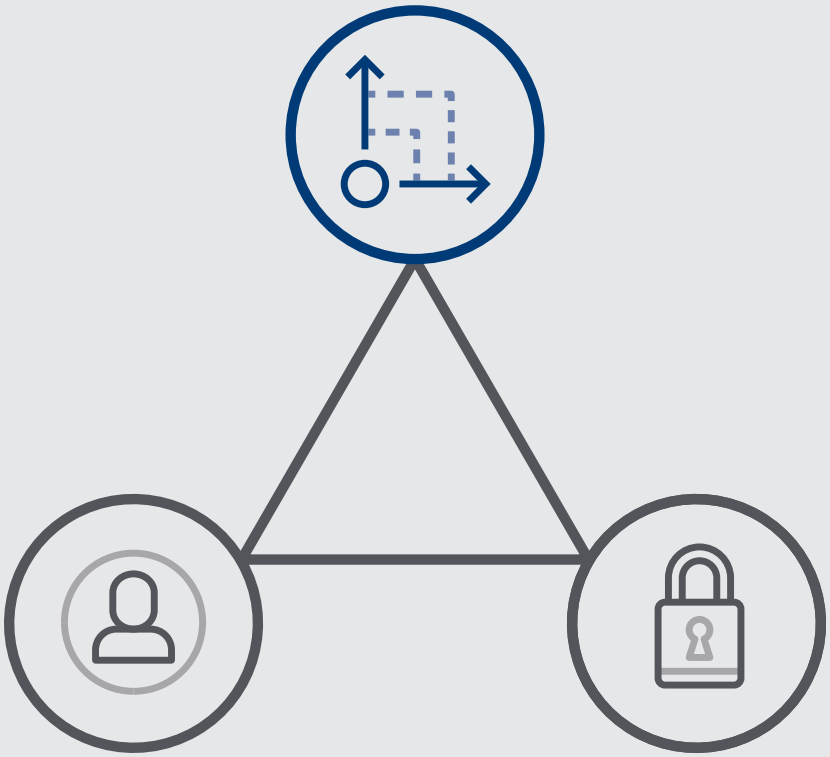
The team thanks Charles Holland, Ipek Ozkaya, and Joshua Poore for their review; and Nancy Ott for editing the paper.

The Pillars of AI Engineering

1. Human-Centered
2. Robust and Secure
3. **Scalable**

The emergent discipline of AI Engineering is focused on three pillars: human-centered AI, robust and secure AI, and scalable AI.

To learn more about AI Engineering, visit **sei.cmu.edu/our-work/artificial-intelligence-engineering**.



Scalable AI

"In many instances, the whole seems to take on a life of its own, almost dissociated from the specific characteristics of its individual building blocks."

—Geoffrey West in *Scale*

Massive improvements in computing resources and data storage capacity enable us to create artificial intelligence (AI) models that encompass billions of parameters. Individuals, organizations, and governments are increasingly applying AI models to address complicated, wide-ranging problems—for example, managing the national electricity grid, tracking outbreaks of disease during pandemics, and combatting online bullying. These multifaceted problems involve large numbers of people, multiple data sources, many geographic locations, varying time scales, and other complex inputs. Solutions to such problems must consider the embedded interconnections and scale to meet the scope of the challenges. It's not as simple as finding what works for small instances of a problem and applying the same methods for larger instances. (If that were the case, there would be no difference between managing a team of 10 employees and managing a team of 100.) Instead, when creating scalable systems of any kind, we must recognize that system properties change as size changes. As a result, we need to carefully consider how to create and guide system development at a scale that is proportional to the scale of the problems we face [1].

***Scalable AI* is defined as the ability of algorithms, data, models, and infrastructure to operate at the size, speed, and complexity required for the mission.** This is not a trivial problem. For example, many academic research teams have developed AI applications with advanced capabilities, such as computer vision systems that detect vehicles in satellite imagery and machine learning systems that predict leakage from gas wells. However, the rocky process of transitioning these systems from research prototypes to enterprise-wide implementations shows how difficult it is to design AI systems for scale. Gartner estimates that as many as 85% of AI projects will fail to deliver their intended value [2]. This failure is often caused by a lack of emphasis on the technological, financial, and organizational factors that change when an AI system is implemented at a large scale. For example, training GPT-3, the world's largest language model, is estimated to have cost OpenAI $12 million dollars due to the expense of its computational requirements [3]. Although this is not a direct failure, the unexpectedly high cost of training this model shows the lack of repeatability over time.

Creating scalable AI systems requires disciplined engineering approaches to guide their development, deployment, and maintenance. From an AI Engineering perspective, the field must work to overcome challenges pertaining to scarcity of data sets and data labels for training and using AI models, obsolescence and reusability of AI tools, and deficiencies in the infrastructure needed to develop and deploy AI capabilities. These challenges are often addressed in an ad-hoc manner. For example, many organizations employ individuals to manually label training data. But what will happen when the data required for training grows by an order of magnitude? What if the time scheduled to complete this task is cut in half? Scalable AI's focus must also include the processes, policies, practices, and tools to support the enterprise scalability of AI capabilities.

Scalability is a critical concept in many engineering disciplines and is crucial to realizing operational AI capabilities. We identify three areas of focus to advance scalable AI:

- **Scalable management of data and models** to overcome data scarcity and collection challenges, and promote reusing and recombining capabilities to scale across missions
- **Enterprise scalability of AI development and deployment** including establishing production pipelines, extensible system architectures, and modern policies and acquisition practices to maintain advanced capabilities and take advantage of rapid innovation in AI technologies
- **Scalable algorithms and infrastructure** to fully apply the power of AI to critical missions, including centralized data center capabilities and distributed cloud-enabled and network-enabled applications for edge devices

For each area, we identify ongoing work as well as challenges and opportunities in developing and deploying AI systems at scale.

Scalable Management of Data and Models

Effective approaches to managing data and models are critical to the scalable and responsible adoption and use of AI. Data is the driving force of modern machine learning and is equally as critical for other AI approaches (e.g., knowledge graphs to support reasoning). Models encode decision or inference processes of various AI techniques. In modern machine learning, a model is the output of a learning algorithm (e.g., a deep neural network) that is trained on a set of data. It can then make predictions about data that shares characteristics with the model's training data. The disciplined management of data and models includes carefully collecting and curating data sets, versioning both data and models, reusing and recombining capabilities, and creating policies and procedures that support discovering and sharing data and models.

Scalable oversight is a major challenge for the careful collection and curation of data sets for use in AI systems [4]. Creating useful datasets for AI can be time consuming, expensive, error-prone, and labor-intensive. Scalable oversight encompasses methods that reduce the time, cost, errors, and labor required to collect and curate data sets that can be used to train or otherwise inform AI techniques. Realizing scalable oversight is particularly challenging in applications where data is scarce and difficult or expensive to collect, or requires specialized expertise to label or synthesize. One example is prediction and pattern-of-life models that require distributions of events that account for both normalcy and anomaly [5]; the latter is a rare event that is by definition hard to capture and accurately label during collection.

In many industry applications of AI, two common approaches to scalable oversight are crowdsourcing and gathering the regular interactions of large numbers of users with Internet applications. At first glance, these approaches only apply to the collection, creation, and curation of useful AI data sets in private technology companies. However, public and government organizations have significant opportunities to use their current and future systems to gather information about the everyday interactions of their workforce (e.g., analysts working on intelligence missions). Proper instrumentation and data collection strategies to capture the interactions of users, analysts, and operators could significantly reduce the effort required to create and label datasets for developing new AI-enabled mission capabilities. Even rudimentary annotations would make data labeling much easier, especially if these annotations can be programmatically repurposed for a variety of use cases. Historically, security concerns prevented the adoption of scalable oversight process in operations. But in recent years, proactive policies enabled these approaches to be deployed.

Successfully scaling AI requires both datasets and models to be discovered, shared, reused, and recombined across a variety of mission capabilities. This involves developing and institutionalizing policies and mechanisms for managing, tracking, versioning, and analyzing reused and derivative capabilities. The implementation and adoption of scalable management mechanisms allows for the use of powerful techniques for using machine learning in applications where there is limited—or in some cases no—labeled data (for examples, see: [6]–[8]).

Transfer learning is commonly used in industry and academia to scale models across domains. A model is trained for a particular problem on a source domain and then reused—perhaps with some modest retraining—for a different problem on a target domain [9]. A typical scenario for transfer learning is when the source domain has an abundance of labeled data and a well-trained model, but the target domain has a limited amount of labeled data. For example, transfer learning could be used to classify vehicles in radar data by applying what is known about classifying vehicles in electro-optical data. *Few-shot learning* is another scaling technique that can be used when labeled data is scarce. It applies machine learning to datasets with a small number of labeled instances [10].

Policies and mechanisms that promote sharing, reusing, and recombining AI capabilities across a variety of missions facilitate the use of powerful techniques like transfer learning and few-shot learning. Furthermore, organizational support for scalable management of data and models democratizes the adoption and use of AI capabilities. This leads to more generalizable and robust applications of AI across a broader variety of missions.

Enterprise Scalability of AI Development and Deployment

As the discipline of AI Engineering grows and matures, organizations that follow its practices will be able to further democratize the responsible development, deployment, adoption, and use of AI capabilities across their enterprise. In addition to scalable management of data and models, organizations can adopt processes, practices, tools, and frameworks to support enterprise scalability of AI. They include iterative development practices, reusable development pipelines, extensible AI-aware system

architectures, common frameworks and interoperability standards, and modernized acquisition policies.

Adapting DevOps practices for AI and machine learning supports iterative development practices and reusable development pipelines. So does the careful overall design and management of production pipelines and the development lifecycle.

The recent proliferation of machine learning has led to promising initial work in adapting DevOps practices specifically for applications that use machine learning. This set of practices—now commonly referred to as MLOps [11]—extends DevOps to include data processing and preparation, model training, model evaluation, model deployment, and continuous monitoring capabilities. It ensures that machine learning components in the system continue to operate as expected when deployed in the real world. MLOps supports the continuous delivery of machine learning [12] capabilities in AI systems and provides opportunities to automate capabilities across the AI system development lifecycle.

MLOps is a relatively new set of practices and tools; many challenges and opportunities remain open. It will continue to evolve as more practitioners and organizations focus on the production and operation of AI and machine learning. Adopting and evolving MLOps as a standard practice, including development pipelines that can be replicated and shared across an organization, directly enables the enterprise scalability of AI.

To develop and grow MLOps as a practice, the workforce needs training and education in the design of machine learning systems *for production*. This focus on training is driven by the realization by both the research and practitioner community that there is a significant difference between building a machine learning model and deploying and operating it in production. To meet this need, Stanford and DeepLearning.ai recently offered two of the first courses in machine learning for production [13], [14]. To grow and evolve this body of knowledge, the field of AI Engineering will need to track ongoing education efforts and coalesce curricula.

Scalability considerations also affect the design and development of systems that include AI components and the processes by which organizations acquire AI systems. At the systems level, system and software architectures [15] must be designed to allow AI capabilities to evolve over time. For example, most machine learning models must periodically be retrained to account for new data or shifts in the production or operational context. This retraining process can be facilitated by a system architecture that makes it easy to swap models in and out or operate multiple models for redundancy or roll-back. Recent work on capturing and understanding machine learning design patterns [16] is a helpful start to building a set of best practices for AI systems engineering.

Scalable Algorithms and Infrastructure

Two factors underlie the renewed focus and promise of AI over the past decade: 1) the availability of large amounts of data and 2) computing resources that can support the data processing and computational demands of modern AI techniques (e.g., training very large-scale deep neural networks). In turn, this promise drives the demand for computing resources higher and higher. In 2018, OpenAI identified a trend of exponentially increasing computing resources required for training the largest AI models, with a doubling every 3.4 months and an annual, year-over-year increase of a factor of ten [17].

As the computational demands of training grow, so do the size and power of the models and the cost to train them. As we mentioned earlier, GPT-3, a state-of-the-art natural language generation model, has 175 billion parameters and is estimated to cost $12M to train [3]. The ability to scale computing infrastructure underlies all aspects of AI model development and deployment and drives continuous improvements and innovations. The computing demands of modern AI are particularly challenging for the defense and national security community. These demands include the need to use and maintain the necessary computing infrastructure to realize AI capabilities at the size, speed, and complexity of missions, plus the requirement to support a very wide array of different mission applications. Continuing to scale up this infrastructure may not be sustainable; different computing paradigms and alternative or improved algorithms and methods must be developed to support improvements in the capabilities of AI systems.

"Extrapolating forward this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Thus, continued progress in these applications will require dramatically more computationally-efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods." [18]

AI-specific computing is a growing market with a flurry of innovation. Over the last decade, graphics processing units (GPUs) have been the dominant resource for training of models, while CPUs still handle the majority of inference cycles. Early GPUs could have been considered an Application Specific Integrated Circuit (ASIC), where the application was graphics processing. While GPUs have been immensely successful, ASIC manufacturers are seeing opportunities to improve their performance over GPUs for training modern AI systems and CPUs for generating inferences. In particular, the use of low precision linear algebra in AI models has revived interest in older architectures such as Coarse-Grained Reconfigurable Architectures and Systolic Arrays, and has led to newly named architectures such as Data Processing Unit and Tensor Processing Unit [19].

While scaling up AI is challenging, the defense and national security community also wants AI to be scaled down to be run on edge devices by individual warfighters and scaled out to be run across regionally and globally distributed operations.

Some defense applications require AI capabilities to be deployed in delayed/disconnected, intermittently-connected, low-bandwidth (DIL) environments. Developing and deploying AI capabilities for these resource-constrained settings builds upon ongoing research in communication protocols, AI-specific computing architectures, federated learning, TinyML, and related areas.

The competition in ASICs for edge computing is also rapidly growing, driven by attempts to realize AI for everyone, everywhere. Edge computing falls into three useful categories: push from the cloud, pull from the Internet of things, and hybrid cloud-edge analytics. One trend is to push deep learning networks to the edge. However, deep learning networks are notorious for being large and computationally expensive [20]. A significant amount of work has therefore been devoted to reducing model size. For example, the tinyML Foundation focuses on "Enabling Ultra-Low Power Machine Learning at the Edge."

Federated learning is an emerging approach to AI that engages edge devices in the training process, and aims to keep the training data localized and private [21]. It is a significant departure from large-scale machine learning (which is undertaken in data centers) and attempts to train a global statistical model using remote devices that number from the tens to the millions. The challenges associated with federated learning at scale include communications expense, system heterogeneity, statistical heterogeneity, and privacy concerns. Future directions in federated learning could include extreme communication schemes, communication reduction and the Pareto frontier, novel models of asynchrony, heterogeneity diagnostics, granular privacy constraints, beyond supervised learning, productionizing federated learning, and establishing federated learning benchmarks.

The Future of Scalable AI

The field of AI Engineering is only beginning to understand what scale in AI systems actually means and how to achieve it. Guidance on the implementation, scale-up, and measurement of system impacts is sparse. How can an AI prototype or pilot project be transformed into a production scale system? Current implementation efforts are plagued with obstacles that include everything from data collection, ethics, and bias to the cost of maintenance and customer interaction. To enable AI systems to reach their full potential, we need research, processes, policy measures, and tools to comprehensively address these barriers.

For the Department of Defense (DoD) and other national security organizations, scalability is critical to match the global size and scope of the problems they face. Data collection and data sharing looks far different in a DoD setting than in large technology companies, which can lead to situations where teams do not readily have the data they need and aren't able to get their ideal dataset size. We need strategies and techniques to help teams navigate data-scarce situations and still ensure system quality. Further issues include the cost of computing resources and how to financially sustain model training over time, as well as how to push technologies to edge computing.

With a large, distributed workforce, the DoD is also involved in guiding the responsible adoption of AI systems for a broad array of stakeholders. AI systems hold significant promise to support the work of intelligence analysts, warfighters, and support professionals, but *what gets adopted and when* remain open questions. Achieving enterprise scalability will require changes to the existing acquisition process to ensure that systems improvements are possible over time. While the government context has its own unique attributes, organizational uniqueness bias can often prevent learning from industry peers. "In plenty of workplaces, leaders are so focused on what makes their industry or culture different from others that they overlook all the ways it's similar to others " [22]. Adoption of AI systems requires leaders to rethink organizational structures,

management strategies, and a host of other elements. Over the past two decades, many companies have executed new organizational models to accommodate AI systems. The DoD could benefit from their knowledge as it seeks to undergo further digital transformation.

At the acquisition level, AI demands more flexible and nimble processes for identifying, vetting, testing, integrating, and updating systems that include AI components. This includes how requirements are captured and shared, migrating to more continuous testing and monitoring of systems, use of modern software practices [23], and improving the skills, of the acquisition workforce. A more modern and flexible approach to acquisition is especially important for the defense and national security communities, as pointed out by the National Security Commission on AI [24] and current work by the DoD's Joint Artificial Intelligence Center [25].

As we defined earlier, scalable AI is the ability of algorithms, data, models, and infrastructure to operate at the size, speed, and complexity of mission needs. At the same time, AI systems require a holistic approach. Scale is just one critical dimension to track. Many companies failed to successfully scale up their systems because they tried to do too much too fast, or lost sight of other outcomes.

> *"The moral is to watch out for the rules you set up, because you are likely to get what you specify and only that."* **[26]**

While an AI model itself may optimize for a single outcome, we need to remain cognizant of mission needs when designing broader AI systems and recognize that scale is one of the strategies that can help us to achieve them.

References

[1] C. B. Weinstock and J. B. Goodenough, "On System Scalability," CMU Software Engineering Institute, CMU/SEI-2006-TN-012, Mar. 2006.

[2] Gartner, "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence," Feb. 13, 2018. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

[3] K. Wiggers, "OpenAI's massive GPT-3 model is impressive, but size isn't everything," *VentureBeat*, Jun. 01, 2020. [Online]. Available: https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything/

[4] D. Sculley et al., "Machine Learning: The High-Interest Credit Card of Technical Debt," in SE4ML: *Software Engineering for Machine Learning*, 2014, pp. 1–9.

[5] A. Ng and S. Russell, "Algorithms for inverse reinforcement learning," *Icml*, vol. 1, pp. 1–8, 2000.

[6] J. Y. Koh, *Model Zoo*. 2021. Accessed: May 17, 2021. [Online]. Available: https://modelzoo.co

[7] The Linux Foundation, *ONNX: Open Neural Network Exchange*. 2021. [Online]. Available: https://onnx.ai/about.html

[8] The Apache Software Foundation, Apache MXNet. 2021. Accessed: May 17, 2021. [Online]. Available: https://mxnet.apache.org/versions/1.3.1/model_zoo/index.html

[9] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," *Proc. IEEE,* vol. 109, no. 1, pp. 43–76, Jun. 2020.

[10] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *ACM Comput. Surv.* CSUR, vol. 53, no. 3, pp. 1–34, Mar. 2020.

[11] M. Treveil et al., *Introducing MLOps*. O'Reilly Media, Inc., 2020.

[12] D. Sato, A. Wider, and C. Windheuser, "Continuous Delivery for Machine Learning," *martinFowler.com.* https://martinfowler.com/articles/cd4ml.html (accessed May 17, 2021).

[13] A. Ng, R. Crowe, L. Moroney, and C. B. Arámburu, "Machine Learning Engineering for Production (MLOps) Specialization," *DeepLearning.AI*, 2021. https://www.deeplearning.ai/program/machine-learning-engineering-for-production-mlops/

[14] C. Huyen, "CS 329S: Machine Learning Systems Design," *CS329S*, 2021. https://stanford-cs329s.github.io (accessed May 17, 2021).

[15] M. Richards and N. Ford, *Fundamentals of Software Architecture: An Engineering Approach*. O'Reilly, 2020.

[16] V. Lakshmanan, S. Robinson, and M. Munn, *Machine Learning Design Patterns*. O'Reilly Media, Inc, 2020.

[17] D. Amodei and D. Hernandez, "AI and Compute," *OpenAI*, May 16, 2018. System dynamics meets the press (accessed May 17, 2021).

[18] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The Computational Limits of Deep Learning," *ArXiv200705558 Cs Stat*, Jul. 2020, Accessed: May 17, 2021. [Online]. Available: http://arxiv.org/abs/2007.05558

[19] J. Dean, D. Patterson, and C. Young, "A new golden age in computer architecture: Empowering the machine-learning revolution," *IEEE Micro*, vol. 28, no. 2, pp. 21–29, 2018.

[20] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 2, pp. 869–904, 2020, doi: 10.1109/COMST.2020.2970550.

[21] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.

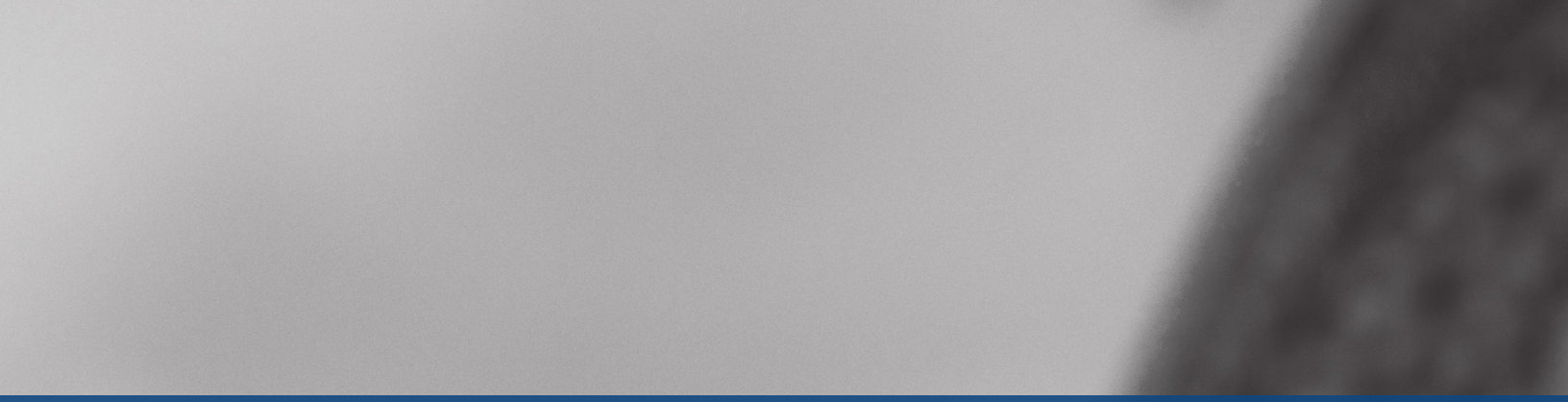
[22] A. Grant, "The Surprising Value of Obvious Insights," *MIT Sloan Manag. Rev.*, vol. 60, no. 3, pp. 8–10, 2019.

[23] J. M. McQuade, R. M. Murray, G. Louie, M. Medin, J. Pahlka, and T. Stephens, "Software Is Never Done: Refactoring the Acquisition Code for Competitive Advantage," Defense Innovation Board, 2019. Accessed: May 17, 2021. [Online]. Available: https://media.defense.gov/2019/May/01/2002126690/-1/-1/0/SWAP%20EXECUTIVE%20SUMMARY.PDF

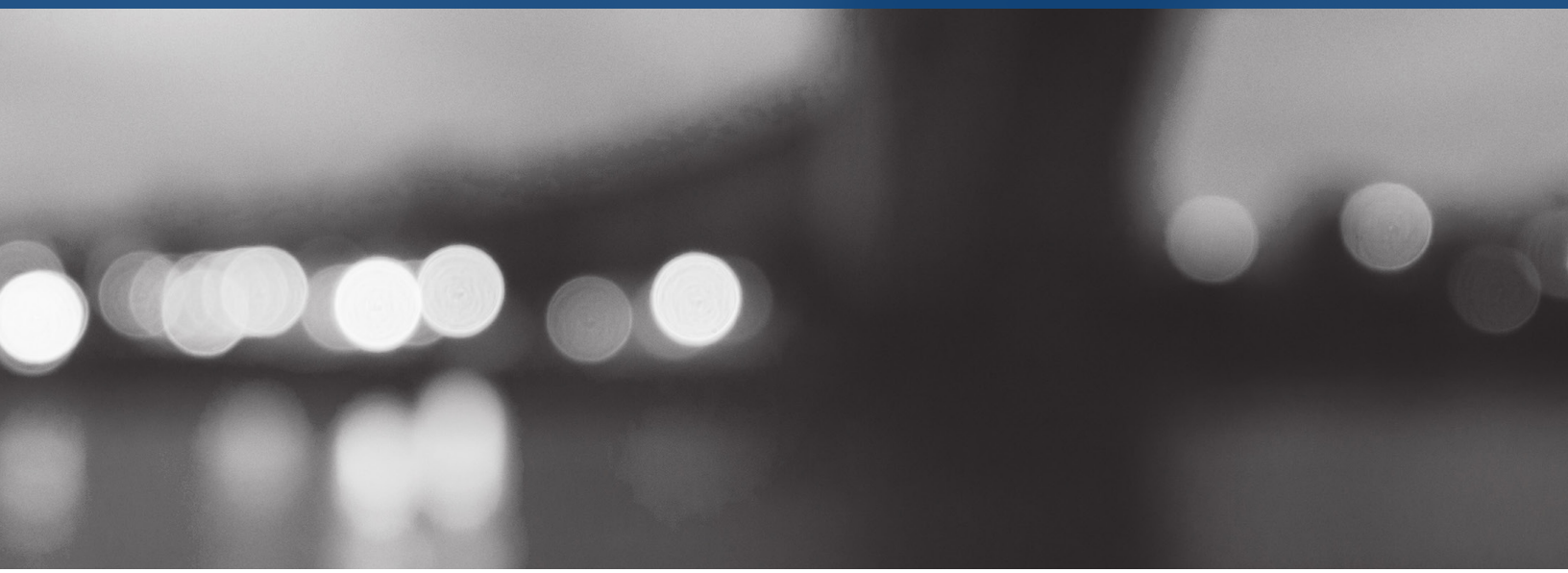
[24] National Security Commission on AI, "Final Report," 2021. [Online]. Available: https://www.nscai.gov/2021-final-report/

[25] J. Serbu, "DoD's JAIC rolling out new contracts to speed up AI acquisition," *Federal News Network*, Feb. 11, 2021. Accessed: May 17, 2021. [Online]. Available: https://federalnewsnetwork.com/artificial-intelligence/2021/02/dods-jaic-rolling-out-new-contracts-to-speed-up-ai-acquisition/

[26] D. Meadows, "System Dynamics Meets the Press," *Syst. Dyn. Rev.*, vol. 5, no. 1, pp. 69–80, 1989.



DM21-0562



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 the art in software engineering and cybersecurity to benefit the 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