National
AI Engineering Initiative

Robust and Secure AI

Carnegie Mellon University
Software Engineering Institute

Robust and Secure AI

Contributors:
Hollen Barmer, Rachel Dzombak,
Matt Gaston, Eric Heim, Jay Palat,
Frank Redner, Tanisha Smith,
Nathan VanHoudnos

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**.

Robust and Secure AI

> *"Failure is central to engineering. Every single calculation that an engineer makes is a failure calculation. Successful engineering is all about understanding how things break or fail."*
>
> —Henry Petroski

All systems fail at some point, no matter how much time and rigor are put into their design and development. AI systems are no different and are susceptible to unexpected and sometimes spectacular failure modes. Some failures show the fragility of system components, such as small stickers that prevent perception systems for self-driving cars from recognizing stop signs [1]. Others show how attackers can use the novel surfaces of AI to drive failure, such as social media "trolls" shifting the personality of an AI chat bot through a barrage of racist language [1], [2]. Still others highlight the lack of versatility in systems, such as when your smart speaker does not respond to the voice of a friend with an accent. Highly competent and well-intentioned developers inadvertently create failure-prone systems across domains and use cases even when operating in closely controlled development, laboratory, and test environments. How can we build robust and secure AI systems for complex and ambiguous contexts such as those in the national security domain, where the potential consequences of failure are catastrophic?

***Robust and secure AI* systems are AI systems that reliably operate at expected levels of performance, even when faced with uncertainty and in the presence of danger or threat.** These systems have built-in structures, mechanisms, or mitigations to prevent, avoid, or provide resilience to dangers from a particular threat model. Our conceptualization builds on existing definitions that focus on correctness of system functions [3], and underscores that system behavior should meet expectations of quality and not generate unexpected emergent behavior as a result of novelty or other changes to the operating environment (e.g., noise, sensor degradation, context shifts). Robustness is not a guarantee against failure, but instead enables users, engineers, and system designers to mitigate common failure modes and know what to do when failure happens.

One of the biggest challenges facing the broad adoption of AI is having confidence that AI systems will work predictably when they are deployed in new and uncontrolled settings.

System developers task AI technologies with complex problems for which there are no guaranteed ways to achieve perfect solutions and no way to construct training data sets that reflect all aspects of the real-world use cases. As a result, the system goal must shift from achieving a perfect outcome to building confidence in AI throughout the entire lifecycle—from design to development to test to operations and around again as the system evolves. Developing new tools, processes, and practices for testing, evaluation, verification, and validation (TEV&V) is critical for building and deploying robust and secure AI systems with confidence.

Current test and evaluation practices predominantly occur during the model evaluation phase of the AI lifecycle and rely on accuracy measurements on test datasets that closely match training data [4]. Accuracy is an important and convenient metric, but focusing on *only* accuracy can impede assessment of whether or not mission outcomes are achieved. Defining relevant mission-specific measurements of performance is critical to guide overall system evaluation, as well as ensure that reliable performance is maintained amidst uncertainty [5]. Expanding test and evaluation capabilities across the AI system lifecycle and across a much richer and mission-relevant set of qualities will enable the responsible adoption and use of AI technologies for defense and national security as well as the effective, iterative, and incremental development of leading-edge mission capabilities [6].

We identify three specific areas of focus to advance Robust and Secure AI for defense and national security:

- **Improving the robustness of AI components and systems** and the need to go beyond accuracy measurements to capture achievement of mission outcomes
- **Designing for security challenges in modern AI systems** including novel attack surfaces and patterns, as well as strategies for risk mitigation
- **Developing processes and tools for testing, evaluating, and analyzing AI systems** and adoption of comprehensive test and evaluation approaches

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

Robustness of AI Components and Systems

There are two general approaches to robust AI: 1) robust against *model errors* and 2) robust against *unmodeled phenomena* [7]. Dietterich characterizes the two approaches as responses to known unknowns, or "uncertain aspects of the world about which the computer can reason explicitly" and unknown unknowns, "those aspects of the world that are not captured by the system's models" [7]. For example, model errors include incorrectly specified hyperparameters, whereas unmodeled phenomena include unpredictable shifts in operating weather due to climate change. For *model errors*, approaches include robust optimization, regularization, risk-sensitive objective functions, and robust inference algorithms. For *unmodeled phenomenon*, approaches include expanding the model, learning a causal model, using a portfolio of models, and monitoring performance. Important prior and ongoing work supports each approach, and significant room for growth and improvement exists. Although these techniques exist in academic papers and have been proven in laboratory settings, many have not yet made their way into practical tools for realizing robustness in AI systems.

Much of the current focus of AI is on modern machine learning (ML) techniques, with a special focus on deep learning or deep neural networks. One challenge to realizing robustness in modern ML systems is underspecification [8]. Underspecification means that the algorithms training deep neural networks solve optimization problems for which there are many possible solutions with equal performance. These algorithms pick one of many possible solutions, leading to models that contain hidden biases and other flaws that are likely to result in surprising failures or unexpected behavior when deployed. Challenges exist around both algorithm selection and the range of possible solutions. Researchers in explainable AI, or *XAI*, are currently working to address the issue of opaque model selection in deep learning [9]. Explainable AI has significant implications for robustness. If model selection is more transparent, errors can be efficiently communicated and performance expectations managed. The existence of large numbers of possible solutions implies that both the problem (optimization for predictive accuracy) and the solution (deep learning models) are underspecified. The fact that modern ML is underspecified begs for new approaches to make more robust models and to include appropriate metrics and measurements of robustness in the evaluation and testing of AI systems relying on deep learning algorithms.

Related to the robustness of modern machine learning algorithms is understanding and trusting the uncertainty or confidence of ML models [10]. Dataset shift—when data in operations is drawn from different distributions than training data—can cause misrepresentation of uncertainty and ultimately overconfidence in model predictions. Results of a recent study show a 40–45% drop in performance when the context of data collection changes, implying that models trained on controlled data sets don't always work in the real world [11]. Calibration techniques can adjust the natural uncertainty scores provided by ML models to more accurately reflect the probability of a prediction being correct, resulting in models that know when they don't know. Equipped with properly calibrated uncertainty measures, design, integration, and monitoring policies can be created to make AI systems more robust.

Robustness is critical for applications intended to operate in changing or challenging environments. To understand and design robust AI systems, we need accepted and validated tools, frameworks, and practices for measuring robustness. Promising directions and best practices for robust AI include methods to "build robustness in" to systems through smart design and the use of algorithms with robustness features, such as portfolio strategies or redundancy [7]. Robustness can be built into approaches to model evaluation. It can be extended across the AI system lifecycle through testing when AI components are integrated into bigger systems and deployed, and through continuously monitoring AI systems for performance and robustness during operations. There are open questions about which testing protocols and processes are employed at each lifecycle phase. Concepts from robustness also inform AI system design principles and patterns. At the systems level, there are opportunities to include uncertainty information as signaling between AI components and other system components. Patterns like ensembles of AI components can lead to more robust AI systems. Finally, improved tools for measuring and analyzing the robustness of both AI components [12] and AI systems will support AI engineers, product managers, designers, software engineers, systems engineers, and operators in designing, building, and operating AI-enabled capabilities.

Security Challenges in Modern AI Systems

An integrated focus on security or "protection against intentional subversion or forced failure" [3] is critical to the goal of deploying robust AI systems that face dangers from a particular threat model. AI systems are software or cyber-physical systems that include AI components and likely many other software components. These AI components are built out of software and data. When considering the security of AI systems, AI engineers need to take full advantage of the vast body of knowledge and best practices for building and securing software systems as well as any security implications specific to AI components. Recent efforts to bring MITRE's ATT&CK framework to securing ML systems in production draw from established knowledge of software security [13].

Much attention has been paid over the past few years to the novel attack surfaces that modern ML techniques (specifically deep learning) present. 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 [14]. One taxonomy of adversarial machine learning organizes attacks on ML models into three categories: learn the wrong thing, do the wrong thing, and reveal the wrong thing [15]. Manipulating training data or training methods can cause ML models to *learn* the wrong behaviors while still performing well in training and model evaluation [16], [17]. Manipulating operational data can deceive ML models and cause them to *do* the wrong thing in operations. Finally, depending on how much can be known about ML models in production, attackers can use various mechanisms to extract information that was used to train models to *reveal* potentially private or confidential information. A variety of methods can be used to mitigate specific attacks and enforce security policies for ML models in the systems where they are deployed.

As adversarial AI continues to advance, defenders (system builders and operators) must make trade-offs when faced with attacks.

First are the relative trade-offs that come with information availability for both attackers and defenders. Perhaps more importantly, recent work in adversarial ML has demonstrated the potential trade-offs between enforcing the different policies of *do*, *learn*, and *reveal*. One example shows that models that are trained to *do* the right thing turn out to be more susceptible to *revealing* information about their training data [18]. Trade-offs between attacker and defender information availability and budget as well as dependencies between different defense policies are important areas for continued research and development. Furthermore, there is a large demand for tools that support AI system developers in understanding security-related considerations for their systems.

Beyond the specifics of the novel attack surfaces of ML algorithms and models, the security workforce and organizational ecosystem must also focus on the implications of increasing amounts of AI in real-world systems. Two areas of opportunity are 1) expansion of security vulnerability coordination to include new types of vulnerabilities that stem from AI technologies, and 2) enhancement of red teaming capabilities, which provide tremendous value for understanding and improving security in traditional software systems [1], [19], [20].

Processes and Tools for Testing, Evaluating, and Analyzing AI Systems

It is easy to focus attention on the robustness and security of AI systems by examining the technical, algorithmic, and even mathematical underpinnings of specific AI techniques. From an AI Engineering perspective, there is a greater need for tools, processes, design patterns, and best practices for promoting robust and secure AI system development and operations. AI engineers need tools similar to those used for software reverse engineering [21], static and dynamic code analysis, fuzz testing, and augmentations of standard approaches for unit, regression, and integration testing. In some cases, traditional software engineering tools are helpful. However, AI—and specifically ML—challenge the utility of existing testing tools. In contrast to traditional software, AI addresses problems that are often broader, less clear in purpose, and have more complex input and output spaces. In most cases, existing tools do not scale to these problems; in others, there is no clear analogy, necessitating the creation of new tools.

Additionally, AI system tools need to be incorporated into modern software development processes as well as automated development and deployment environments and frameworks. In particular, tools that support robust and secure AI should be integrated into DevOps or MLOps pipelines and systems for Continuous Integration and Continuous Delivery (CI/CD) wherever possible. AI systems require the expansion of CI/CD frameworks to include Continuous Monitoring (CM). This ensures that the robustness and security of AI systems can be assessed and assured throughout the system lifecycle and trigger necessary mitigations, incremental improvements, model retraining, or system redesign based on how the systems are performing in operation.

Robust and Secure AI

Robust and secure AI—specifically, the ability to design, develop, deploy, and operate robust and secure AI systems—is both a critical component of AI Engineering and an imperative for the DoD. Robustness and security in AI systems are key to achieving mission outcomes and can enable many other related qualities such as safety, reliability, dependability, and stability. Robust and secure systems also support policy-related concerns like privacy, fairness, and ethics.

In the DoD context, current approaches and policies for developmental test and evaluation (DT&E) and operational test and evaluation (OT&E) must evolve to include AI. In the context of AI Engineering, DT&E and OT&E have significant implications for acquisition processes and practices. They must factor in considerations for robust and secure systems, including how to generate system testing requirements, how to purchase them, and how to work within budgets when continuous monitoring is needed. A recent workshop hosted by the University of Maryland's Applied Research Lab for Intelligence and Security (ARLIS) highlighted the needs and challenges that AI introduces for OT&E specifically [22]. The workshop emphasized the disconnect between what is easy to measure and what is operationally meaningful. Test and evaluation practices need to keep pace with rapid changes in technology. This requires growing a proactive and nimble test and evaluation community across the DoD, including growing the number of AI testers that currently exist within the DoD.

Furthermore, the DoD faces a cultural challenge when it comes to instilling a mindset of experimentation across all stakeholders involved in AI system development and deployment. While experimentation and prototyping are needed for nearly all domains, the complexity of AI systems, the at-times information opacity, and the new behaviors needed to enable effective human-machine teams all necessitate early and frequent system testing. Frequently, people assume that testing is a time-consuming activity, when in fact fixing errors is what takes time, particularly in later stages of project development [23]. This challenge is not unique to the DoD: "Culture—not tools and technology— prevents companies from conducting the hundreds, even thousands, of tests they should be doing annually and then applying the results [24]."

Teams designing and developing AI systems need to engage in rigorous cycles of inquiry, learning, building, and testing to identify flaws in the structure of the system objective or data acquisition and manipulation processes."

They further need to continuously evaluate the model's robustness to unmodeled phenomena, resilience to attacks, and ability to be used for decision making. As with any system, the behaviors of component parts must be viewed in relation to each other, and teams should use experimentation to characterize interdependence within the system and assess potential unexpected behaviors that emerge when a change is implemented [25], [26].

As organizations implement AI systems in higher-stakes contexts, the robustness and security of those systems become of utmost importance: "When AI systems wield control of highway networks, power grids, financial markets, they become attractive targets [7]." National security applications, of course, fall into the same category. To enable robust and secure AI systems, the DoD must consider how to approach problems from multiple views, how to consider both known unknowns and unknown unknowns, and how to instantiate a culture of experimentation and testing to ensure that AI systems are engineered and can reach the full potential of their impact over time.

References

[1] Partnership on AI, “Incident Database,” 2021. https://incidentdatabase.ai/ (accessed Mar. 08, 2021).

[2] Lexalytics, “Stories of AI Failure and How to Avoid Similar AI Fails,” *Artificial Intelligence*, Jan. 30, 2020. https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020 (accessed Mar. 08, 2020).

[3] ISO, “Systems and software engineering — Life cycle management — Part 1: Guidelines for life cycle management,” 2018. [Online]. Available: ISO/IEC/IEEE 24748-1:2018

[4] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, “Data Validation for Machine Learning,” in *Proceedings of Machine Learning and Systems*, 2019, vol. 1, pp. 1–14.

[5] D. Jannach and C. Bauer, “Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research,” *AI Magazine*, vol. 41, no. 4, pp. 79–95, 2020.

[6] JAIC, “JAIC Spotlight: The JAIC’s Test, Evaluation, and Assessment Team Shapes Future AI Initiatives,” May 27, 2020. https://www.ai.mil/blog_05_27_20-jaic_spotlight_test_evaluation_and_assessment_team.html (accessed Mar. 08, 2021).

[7] T. G. Dietterich, “Steps Toward Robust Artificial Intelligence,” *AI Magazine*, vol. 38, no. 3, pp. 3–24, 2017.

[8] A. D’Amour *et al*., “Underspecification Presents Challenges for Credibility in Modern Machine Learning,” *ArXiv201103395 Cs Stat*, Nov. 2020, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/2011.03395

[9] H. Hagras, “Toward Human-Understandable, Explainable AI,” Computer, vol. 51, no. 9, pp. 28–36, Sep. 2018, doi: 10.1109/MC.2018.3620965.

[10] Y. Ovadia *et al.*, “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift,” *ArXiv190602530 Cs Stat*, Dec. 2019, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1906.02530

[11] A. Barbu *et al.*, “ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models,” in *33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019, p. 11.

[12] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, “Methods for comparing uncertainty quantifications for material property predictions,” *Mach. Learn. Sci. Technol.*, vol. 1, no. 2, p. 025006, May 2020, doi: 10.1088/2632-2153/ab7e1a.

[13] J. Spring, “Adversarial ML Threat Matrix: Adversarial Tactics, Techniques, and Common Knowledge of Machine Learning,” *CMU Software Engineering Institute*, Oct. 22, 2020. https://insights.sei.cmu.edu/cert/2020/10/adversarial-ml-threat-matrix-adversarial-tactics-techniques-and-common-knowledge-of-machine-learning.html (accessed Mar. 08, 2021).

[14] B. Draper, “Guaranteeing AI Robustness Against Deception (GARD),” *DARPA Research*, 2021. https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception (accessed Mar. 08, 2021).

[15] J. Beieler, “AI Assurance and AI Security: Definitions and Future Directions,” presented at the Computing Research Association, Feb. 02, 2020. Accessed: Mar. 08, 2021. [Online]. Available: https://cra.org/ccc/wp-content/uploads/sites/2/2020/02/John-Beieler_AISec_AAAS.pdf

[16] IARPA, “Trojans in Artificial Intelligence (TrojAI),” *Research Programs*, 2021. https://www.iarpa.gov/index.php/research-programs/trojai (accessed Feb. 08, 2021).

[17] P. Bajcsy, N. J. Schaub, and M. Majurski, “Designing Trojan Detectors in Neural Networks Using Interactive Simulations,” *Appl. Sci.*, vol. 11, no. 4, p. 1865, Feb. 2021, doi: 10.3390/app11041865.

[18] J. Helland and N. VanHoudnos, “On the human-recognizability phenomenon of adversarially trained deep image classifiers,” *ArXiv210105219 Cs*, Dec. 2020, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/2101.05219

[19] Software Engineering Institute, “Machine learning classifiers trained via gradient descent are vulnerable to arbitrary misclassification attack,” *Vulnerability Note VU#425163*, Jun. 04, 2020. https://www.kb.cert.org/vuls/id/425163 (accessed Mar. 08, 2021).

[20] J. M. Spring, A. Galyardt, A. D. Householder, and N. VanHoudnos, “On managing vulnerabilities in AI/ML systems,” in *New Security Paradigms Workshop 2020*, Oct. 2020, pp. 111–126. doi: 10.1145/3442167.3442177.

[21] A. Abdalla, “A Visual History of Interpretation for Image Recognition,” *The Gradient*, Jan. 16, 2021. https://thegradient.pub/a-visual-history-of-interpretation-for-image-recognition/ (accessed Mar. 08, 2021).

[22] ARLIS, ““Should you rely on that AI?” -- A Workshop to explore the multi-domain challenge of AI operational testing,” Jan. 28, 2021. https://www.arlis.umd.edu/wksp202101-rely-on-ai (accessed Mar. 08, 2021).

[23] R. Kohavi *et al*., “Online experimentation at Microsoft,” *Data Min. Case Stud.*, vol. 11, no. 2009, pp. 11–50, 2009.

[24] S. Thomke, “Building a Culture of Experimentation,” *Harv. Bus. Rev.*, vol. 98, no. 2, pp. 20–47, 2020.

[25] S. Patel and K. Mehta, “Systems, Design, and Entrepreneurial Thinking: Comparative Frameworks,” *Syst. Pract. Action Res.*, vol. 30, no. 5, pp. 515–533, Oct. 2017, doi: 10.1007/s11213-016-9404-5.

[26] R. Dzombak and S. Beckman, “Unpacking Capabilities Underlying Design (Thinking) Process,” *Int. J. Eng. Educ.*, vol. 36, no. 2, pp. 574–585.

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