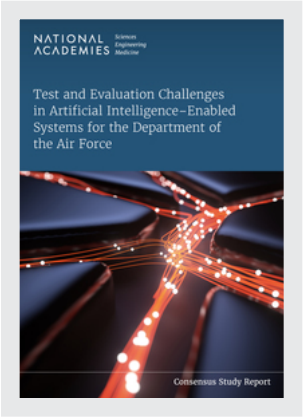

Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force (2023)

DETAILS

194 pages | 7 x 10 | PAPERBACK

ISBN 978-0-309-70439-7 | DOI 10.17226/27092

CONTRIBUTORS

Committee on Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems Under Operational Conditions for the Department of the Air Force; Air Force Studies Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine. 2023. *Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27092.

NATIONAL ACADEMIES Sciences Engineering Medicine

NATIONAL ACADEMIES PRESS
Washington, DC

Test and Evaluation Challenges in Artificial Intelligence–Enabled Systems for the Department of the Air Force

Committee on Testing, Evaluating, and Assessing Artificial Intelligence–Enabled Systems Under Operational Conditions for the Department of the Air Force

Air Force Studies Board

Division on Engineering and Physical Sciences

Consensus Study Report

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by a contract between the National Academy of Sciences and the Department of the Air Force under award number FA955016D00001 FA8651-21-F-9323. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-70439-7
International Standard Book Number-10: 0-309-70439-1
Digital Object Identifier: https://doi.org/10.17226/27092

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2023. *Test and Evaluation Challenges in Artificial Intelligence–Enabled Systems for the Department of the Air Force*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27092.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

Rapid Expert Consultations published by the National Academies of Sciences, Engineering, and Medicine are authored by subject-matter experts on narrowly focused topics that can be supported by a body of evidence. The discussions contained in rapid expert consultations are considered those of the authors and do not contain policy recommendations. Rapid expert consultations are reviewed by the institution before release.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON TESTING, EVALUATING, AND ASSESSING ARTIFICIAL INTELLIGENCE-ENABLED SYSTEMS UNDER OPERATIONAL CONDITIONS FOR THE DEPARTMENT OF THE AIR FORCE

MAY CASTERLINE, NVIDIA, *Co-Chair*
THOMAS A. LONGSTAFF, Carnegie Mellon University, *Co-Chair*
CRAIG R. BAKER, Baker Development Group, LLC
ROBERT A. BOND, Massachusetts Institute of Technology
RAMA CHELLAPPA (NAE), Johns Hopkins University
TREVOR DARRELL, University of California, Berkeley (*until December 2022*)
MELVIN GREER, Intel Corporation
TAMARA G. KOLDA (NAE), Independent Consultant, MathSci.ai
NANDI O. LESLIE, Raytheon Technologies (*until December 2022*)
ROBIN R. MURPHY, Texas A&M University
DAVID S. ROSENBLUM, George Mason University
JOHN (JACK) N.T. SHANAHAN, United States Air Force (retired)
HUMBERTO SILVA III, Sandia National Laboratories (*until December 2022*)
REBECCA WILLETT, University of Chicago

Staff

RYAN MURPHY, Program Officer
GEORGE COYLE, Senior Program Officer
EVAN ELWELL, Research Associate
CHARLES YI, Research Assistant
MARTA HERNANDEZ, Program Coordinator
AMELIA A. GREEN, Senior Program Assistant (*until July 2022*)

Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

JILL CRISMAN, Digital Safety Research Institute
MICHAEL A. FANTINI, United States Air Force (retired)
SHAUN GLEASON, Oak Ridge National Laboratory
J. MARCUS HICKS, United States Air Force (retired)
MARVIN J. LANGSTON, Independent Consultant
GARRY McGRAW, Berryville Institute of Machine Learning
YEVGENIYA "JANE" PINELIS, Johns Hopkins University Applied Physics Laboratory
AMIR SADOVNIK, Oak Ridge National Laboratory
MICHAEL SCHNEIDER, Lawrence Livermore National Laboratory
ALBERT SCIARRETTA, CNS Technologies, Inc.
JONATHAN SMITH, University of Pennsylvania
REBECCA WINSTON, Winston Strategic Management Consulting

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by STEVE BELLOVIN, Columbia University, and BOB SPROULL, University of Massachusetts Amherst. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Contents

APPENDIXES

Preface

At the request of the 96th Test Wing of the U.S. Air Force and Air Force Materiel Command, the National Academies of Sciences, Engineering, and Medicine were asked to convene a committee to conduct a consensus study to examine the Air Force Test Center's technical capabilities and capacity to conduct rigorous and objective tests, evaluations, and assessments of artificial intelligence (AI)-enabled systems under operational conditions and against realistic threats.

The National Academies of Sciences, Engineering, and Medicine appointed the Committee on Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems Under Operational Conditions for the Department of the Air Force to conduct this study, per the Statement of Task found in Appendix A and Box P-1. The committee held its initial kickoff meeting in April 2022, conducted a data-gathering workshop in June 2022 (a Proceedings of a Workshop—in Brief of which can be found in Appendix E), and held further data-gathering sessions throughout 2022 and early 2023, including a site visit to Eglin Air Force Base. Agendas for the data-gathering meetings can be found in Appendix B. Biographies of the committee members can be found in Appendix C. Appendix D contains a list of acronyms and abbreviations used in the report.

BOX P-1
Statement of Task

The National Academies of Sciences, Engineering, and Medicine will establish an ad hoc committee to (1) plan and convene a multi-day workshop and (2) conduct a consensus study to examine the Air Force Test Center's technical capabilities and capacity to conduct rigorous and objective test, evaluation, and assessments of artificial intelligence (AI)-enabled systems under operational conditions and against realistic threats. Specifically, the committee will:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyberattacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

The committee will provide workshop proceedings—in brief and in a report summarizing the results from the consensus study.

Summary

The Department of the Air Force's (DAF's) Air and Space Forces stand on the shoulders of 75 years of comprehensive, rigorous service-wide test and evaluation (T&E) policies, processes, and practices. The combination of a large cadre of designated test personnel, sustained funding, dedicated test organizations, test infrastructure, career-long T&E education and training, and a unique test culture have been instrumental in shaping the current force. Absent a highly disciplined systems engineering approach to testing and the continuous focus on T&E in every aspect of operations, today's DAF would be far less capable and safe.

In requesting this study on testing, evaluating, and assessing the performance of artificial intelligence-enabled systems under operational conditions, DAF leaders recognize both the opportunities and challenges inherent in integrating artificial intelligence (AI) at speed and at scale across the DAF. Integration of AI-enabled capabilities into the DAF has been limited, with a slow pace of adoption so far. The demand for and integration of such capabilities is expected to accelerate substantially based on current trends and expected technological developments in AI and related fields.

In its final report published in March 2021, the National Security Commission on AI (NSCAI) noted that "having justified confidence in AI systems requires assurances that they will perform as intended when interacting with humans and other systems. The T&E of traditional legacy systems is inefficient at providing these assurances. To minimize performance problems and unanticipated outcomes, an entirely new type of T&E will be needed."[1] The NSCAI recommended that all

[1] National Security Commission on Artificial Intelligence, 2021, *The National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, p. 137.

the military Services should "establish a test, evaluation, verification, and validation (TEVV) framework and culture that integrates testing as a continuous part of requirements specification, development, deployment, training, and maintenance and includes runtime monitoring of operational behavior."[2] This committee echoes this NSCAI recommendation.

DAF leaders must now address the pervasive implications of AI T&E across the entire DAF. The DAF has not yet prioritized AI T&E in a way that matches its historical investments in its other T&E capabilities. For example, DAF has not developed a cadre of DAF-wide AI experts or implemented the requisite AI T&E frameworks. Similarly, the DAF has not established enterprise-level T&E policies and infrastructures to support testing autonomous or AI-enabled autonomous systems, either in isolation or in integrated within system-of-systems architectures. Instead, T&E of current AI capabilities has largely relied on ad hoc and bespoke processes and procedures. The ad hoc nature of current DAF AI T&E and the lack of formal guidance complicated this committee's efforts to evaluate current assessment methods employed by the DAF. Much greater investments are needed in AI T&E than previous T&E resources; partially because previous T&E was notoriously under-resourced and because AI systems are so complex. However, these have been boosted over the past 2 years by the work of the AI T&E Community of Interest (CoI) established by the Office of the Secretary of Defense (OSD) Joint Artificial Intelligence Center (JAIC). As discussed in this report, the DAF cannot presently successfully incorporate AI-based solutions. Without significant improvements to the DAF's ability to test and evaluate AI, the DAF will be unable to successfully incorporate AI into DAF systems. To acknowledge the NSCAI's findings and associated recommendations for AI T&E and to enable the DAF to field AI-enabled capabilities that are highly effective, safe, and used responsibly, DAF leaders must prioritize AI T&E. They should do so in a way that, as the committee describes in more detail throughout the report, recognizes the importance of AI T&E throughout the entire AI life cycle, rather than segregated into distinct developmental test and evaluation and operational test and evaluation (OT&E) phases as with traditional weapon systems (see Section 3.2). The committee found that this prioritization includes but is not limited to:

- Fostering a unique AI T&E culture
- Establishing DAF-wide AI T&E governance with sufficient authority
- Dedicated and sustaining the resources necessary for AI T&E
- Integrating data collection and curation into the AI T&E pipeline
- Creating the virtual environments and simulations necessary to create simulated data or to use for reinforcement machine learning

[2] Ibid, p. 384.

- Emphasizing human-systems integration (HSI) such as for human-AI teaming
- Developing the AI T&E workforce

These shortcomings underscore the challenges the entire federal government faces in establishing organization- and agency-wide AI T&E processes and procedures. Unlike digital-age technology companies that rapidly embrace AI capabilities, the DAF is analogous to traditional companies that are only now beginning to adopt AI technologies across their respective industries. Therefore, it is an opportune time for the DAF to craft an AI T&E vision and commit to a long-range AI T&E strategy and implementation plan that includes specific and measurable objectives and goals. There is no time to waste: the risks to the DAF from remaining "frozen in place" regarding AI T&E are significant and will increase exponentially over time. The DAF will only gain ground through the prioritization of AI T&E and commensurate near-term commitment of resources. Rigorous and comprehensive end-to-end T&E of AI-enabled capabilities will significantly increase the DAF's ability to field systems while also allowing end-users to gain justified confidence in AI-enabled systems and tools.

As demonstrated by previous examples of AI projects carried out at scale and both DoD- and industry-wide digital modernization programs, leaders commonly underestimate the investments of time, expert human resources, and money required to implement digital modernization and establish modern AI data management best practices. Without accelerating digital modernization[3] of the DAF's underlying T&E infrastructure, to include information architecture and commitment to a DAF-wide T&E data strategy and implementation plan,[4] the DAF will struggle to assess AI-enabled solutions at the required scale. Therefore, the committee recommends that the DAF immediately update its comprehensive analysis of resource requirements immediately to ensure AI T&E digital modernization efforts

[3] This digital modernization across the AFTC, AFOTEC, and USAFWC includes but is not limited to prioritizing (and sustaining) funding for and rapidly installing AI stacks (AI tools, modern software platforms, data libraries, and providing access to the same computing environments and information technology architectures available to the nation's leading commercial technology companies). The 2022 establishment of the Autonomy, Data, and AI Experimentation (ADAX) proving ground at Eglin AFB as a joint venture between CDAO and AFWERX, supported by the Eglin AFB test ecosystem, is an encouraging first step. One of ADAX's missions is to assess the viability of commercial technologies for Air Force adoption. The ADAX team is coordinating with the DAF ABMS program office to develop initial use cases. Once a technology is determined to be suitable for integration, ADAX personnel will design an Air Force AI test plan. In July 2022, the Air Force Test Center published its own "Digital Modernization Strategy" and initiated three digital engineering efforts. The committee recommends including AI T&E as part of these efforts.

[4] Department of Defense, 2020, "DoD Data Strategy," https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DoD-Data-Strategy.pdf.

are included in the DAF's overarching digital transformation plans and take steps to sustain AI T&E resources in future DAF budgets.

The magnitude of changes this report suggests will require dedicated leadership, continuous oversight, and individual responsibility and accountability. These outcomes are best attained by formally designating a senior AI T&E official who reports to the secretary of the air force, is responsive to the chiefs of the air and space forces, and who has the necessary resources and authorities to implement DAF-wide changes. For this reason, the committee recommends that the secretary of the air force formally designate an overall DAF AI T&E champion at the general officer or senior executive service level in the DAF and grant them the necessary authorities to execute DAF-wide changes on the behalf of the secretarys and chiefs of the two Services.[5] The 2022 dual-hat designation of the 96th Operations Commander as the Chief of AI Test and Operations for the DAF Chief Data and AI Office (CDAO) is a positive and important step. The committee views the 96 OG and CC as one of the primary beneficiaries of this report. However, as currently constituted, the chief of AI test and operations for the DAF CDAO does not have the authority to make the scope and scale of changes across the DAF this committee believes necessary to enable and accelerate AI T&E. Therefore, the DAF needs a formally-designated advocate with an appropriate breadth and depth of AI and T&E experience, along with the commensurate background and expanded authorities, responsibility, and resources. This champion should establish an AI governance structure that includes delineating formally AI T&E reporting relationships and roles and responsibilities across the Tri-Center,[6] the future U.S. Space Force Operational Test Agency (OTA),[7] the DAF CDAO, and operational air, intelligence, command and control (C2), space, and cyber units. This process should include assessing what broader DAF-wide organizational and

[5] The committee uses the term "champion" as illustrative; it does not take a position on the actual title, or whether the designated official should be a general officer or civilian senior executive, or whether the position should reside within the AFTC, AFOTEC, the U.S. Air Force Warfare Center, or elsewhere. The committee notes, however, that this individual will be required to coordinate AI T&E roles and responsibilities across the three primary DAF test and evaluation commands (AFMC, ACC, and AFOTEC) and the DAF Chief Data and AI Office (CDAO). Additionally, while the committee calls for a single DAF AI champion, the committee acknowledges the potential benefits of designating separate AI T&E champions for both the air force and the space force. The committee recommends that the DAF analyze the potential benefits and drawbacks of these various options, with the goal to designate the individual(s) as soon as possible.

[6] Comprising the Air Force Test Center (AFTC) (Air Force Materiel Command, to include the AFMC Digital Transformation Office or DTO); the United States Air Force Warfare Center (USAFWC) (Air Combat Command); and the Air Force Operational Test and Evaluation Center (AFOTEC) (CSAF).

[7] If established.

governance changes are needed to reflect the differences between AI T&E and T&E for all other air force systems and capabilities.[8]

There are many similarities between the T&E of aircraft, weapons, sensors, command and control, and cyber systems and the T&E of AI-enabled systems. Most importantly, the same basic systems engineering principles that have proven instrumental in fielding all previous DAF capabilities are equally applicable to AI. Therefore, the foundational systems theory concepts that have served the DAF so well over the past 75 years provide the appropriate starting point for crafting DAF AI T&E strategies and implementation plans.[9]

In view of AI as a software-centric capability, however, major differences drive the need for a new approach to several critical aspects of AI T&E.

The major differences include:

- The lack of a clear demarcation between the developmental test (DT) and operational test (OT) or between initial operational T&E (IOT&E) and follow-on operational T&E (FOT&E) for AI capabilities.
- The importance of and reliance on iterative and incremental (agile development approaches) software development and adaptive T&E principles (AIOps or DevSecOps, see Section 3.2) instead of linear and sequential (waterfall) software development for AI systems.
- The centrality of data (including the potential for skewed, corrupted, or incomplete datasets) also necessitates the emphasis on its collection, curation, and high-end computing.
- A continuous data-based learning capability that continually changes fielded AI systems necessitates continued testing.
- The importance and challenges of domain adaptation for AI-enabled systems.
- Probabilistic or statistically predictable (i.e., non-deterministic) behavior.
- The effects and risks of adversarial attacks against AI models.
- The challenges of AI explainability and auditability.

[8] This should include evaluating the success of past integrated T&E efforts, such as combined test forces with matrixed OT and DT personnel from across the Tri-Center and assessing their utility for DAF-wide AI T&E projects. The Operational Flight Program-Combined Test Force (OFP-CTF) at Eglin AFB, with a rotating DT and OT commander with authority to direct efforts using resources from both the USAFWC and AFTC, serves as a useful reference point.

[9] See Department of Defense, 2023, "Autonomy in Weapon Systems," DoD Directive 3000.09. Office of the Under Secretary of Defense for Policy, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. This is one example of a current DoD policy that bridges the gap between hardware- and software-centric weapon systems. This directive establishes requirements for TEVV of autonomous and semi-autonomous weapon systems, to include AI-enabled capabilities.

- Continuous integration and continuous delivery (CI/CD) for fielded AI-enabled systems requires commensurate T&E.
- New T&E AI methods, tools, and processes geared toward identifying and addressing AI-related cyberattacks and their effects throughout the testing and operational life cycle of an AI system.
- The importance of adding instrumentation to fielded AI-enabled systems to monitor their performance over time, including metrics to log and analyze changes, since the performance and metrics change with continuous learning.

As the committee explains later in the report, these differences will also drive changes to existing requirements formulation processes for new AI capabilities and AI-enabled systems and how performance metrics are used and evaluated during testing (see Section 5.5).

The difficulty of defining comprehensive T&E requirements for software-centric capabilities is that the "black box" performance under operational conditions could change continually based on the ingestion of more data that generate probabilistic or statistically predictable behavior rather than deterministic results. The intersection of these two equally important considerations leads to a fundamental and persistent challenge for AI T&E today: understanding what requirements to test against when performing T&E for new or fielded AI systems.

The importance of human-system interfaces was one of the other resounding themes throughout this period of study. AI's enormous potential will never be unleashed without changing how humans and machines interact in a more digitized future. While human-system integration (HSI) has been studied extensively over the past 50 years, it is evident that the kinds of AI anticipated soon will demand a different approach to how humans learn to work with "smart" machines. User interface and user experience (UI/UX) are more important than ever, yet much more analysis is needed to understand how to optimize HSI and assess the performance of human-machine interfaces. Optimizing the integration of humans and AI-enabled machines, which in turn depends on redesigning human-machine interfaces and recalibrating human and machine roles and responsibilities, will be one of the most important and defining features of an AI-enabled future. HSI and human-AI team effectiveness must be considered during the T&E of AI-enabled systems.

As the committee's work proceeded, the committee determined that the AI T&E questions the DAF asked the committee to consider are intimately and inextricably related to larger issues of AI-based system acquisitions within the DAF. Thus, the committee realized that only by placing these questions within this larger context can they be properly understood and addressed, with resulting actionable recommendations. This report, therefore, follows and builds on this theme from chapter to chapter.

Chapter 1 reviews the current state of AI in the DAF. It finds that the DAF is in the early stages of incorporating modern AI implementations (see Section 1.3)

into its systems and operations. It has not yet acquired modern AI capability within the standard acquisition processes of a major defense acquisition program (MDAP) or major automated information system (MAIS). Chapter 1 also discusses what the committee means by AI, and several different categories of AI implementation. Chapter 1 also reports that DAF AI-related projects have been research and development initiatives, proof-of-concept demonstrations, or integrated into existing systems as upgrades or prototypes. The chapter notes that in the absence of AI-specific DoD and DAF standards, current DAF prototyping projects have adopted ad hoc acquisition and T&E processes. These ad hoc methods, by nature, do not scale, nor are they consistent. However, the projects reviewed mostly followed sound commercial practices. The chapter ends with a detailed case study of Project Maven. The lessons learned in Project Maven serve as signposts whose themes inform much of the report's findings and recommendations. In particular, the chapter highlights how Project Maven, as a pathfinder AI program within DoD, underscored the importance of rigorous T&E, adopting and adapting industry best-practices, and staying abreast of new ideas from the top AI researchers in academia. Project Maven and other examples from this chapter emphasize the need to retrain AI models to meet unanticipated and changing operational conditions.

Chapter 2 reviews AI and AI-based systems to establish definitions and introduce salient aspects of AI and AI-related technologies. The chapter points out the fundamental importance of data within the machine learning training and testing processes. The chapter presents a historical overview of AI and AI test and evaluation before discussing human-machine teaming. It then proceeds to a detailed discussion of adjusting the evolution of DAF T&E protocols in response to the rapid pace of AI technology advances. The chapter notes a higher level of trust in existing non-AI-enabled systems garnered through years of user familiarity with such systems and continual refinement in specialized T&E approaches. It observes that the DAF T&E community is especially adept at assessing and optimizing human-machine interactions for its piloted weapon systems. However, the chapter concludes that DAF T&E practices neglect important aspects of AI-based HMI (human-machine interface). In particular, it concludes that the DAF needed to refocus all of its acquisition, T&E, operation, and sustainment processes on gaining user trust for deployed and emerging AI-enabled systems. It discusses how human-AI interfaces present new challenges as responsibilities shift between humans and intelligent machines and new concepts of operations (CONOPS) emerge. The chapter notes that inexperience in a future environment characterized by the widespread fielding of AI-enabled systems, the DAF would only be able to achieve maximum performance by focusing specifically on superior human-system integration. The chapter concludes by emphasizing the importance of giving more prominence to Human Readiness Levels (HRL) and UI/UX for AI-enabled military systems and revamping how future military systems are designed for a more digital future.

Chapter 2 offers a key finding and recommendation in the area of human-system integration, or HSI.

> Finding 2-1: The DAF has not yet developed a standard and repeatable process for formulating and assessing HSI-specific measures of performance and measures of effectiveness.
>
> *Conclusion 2-1: The future success of human-AI systems depends on optimizing human-system interfaces. Measures of performance and effectiveness, to include assessments of user trust and justified confidence, must be formulated during system design and development, and assessed throughout test and evaluation and after system fielding.*
>
> **Recommendation 2-1: Department of the Air Force (DAF) leadership should prioritize human-system integration (HSI) or HSI across the DAF, with an emphasis on formulating and assessing HSI-specific measures of performance and measures of effectiveness across the design, development, testing, deployment, and sustainment life cycle.**

Chapter 3 reviews the historical, traditional approach to T&E in the Air Force and then discusses why current practices are insufficient for effective T&E of AI-based systems—particularly the lack of clean lines between developmental test and evaluation (DT&E) and operational test and evaluation (OT&E) for AI capabilities. The chapter observes the lack of formal DoD and DAF AI T&E standards and policies. It notes that seminal specifications from CDAO are emerging however, and that the committee expects OSD Director of Operational Test and Evaluation (DOT&E) will adapt CDAO's frameworks and playbooks and will promulgate the new products DoD-wide.The chapter highlights that OSD DOT&E has provided an initial roadmap for redesigning T&E for DoD AI-enabled systems to reflect the substantial differences between the T&E of traditional DoD systems and the T&E of AI capabilities. The chapter also reviews the role of AI and Development, Security, and Operations (DevSecOps)/AIOps and notes the importance of accelerating the use of agile methodologies across the DAF and designing Artificial Intelligence for IT Operations (AIOps) architectures as a critical part of the AI life cycle. The chapter notes that over the last decade, the commercial sector, particularly the autonomous vehicle industry (see Section 3.2), has employed and refined agile methods to significantly advance the design and deployment of T&E methodologies for safety-critical AI-enabled systems. It also notes that the AIOps solutions designed for commercial applications will not meet the operational requirements of the DAF. This chapter introduces the concept of *justified confidence* as a progressive measure of trustworthiness and notes that developers, testers, and users should gain justified confidence in AI-enabled systems over time as they become increasingly familiar with system

performance limits and behaviors. Next, the chapter discusses *AI assurance*, another term that, along with justified confidence and trustworthiness, replaces the binary concept of trust when referring to AI-enabled systems. The chapter ends with analyzing operationally oriented risks pertaining to integrating AI capabilities into DAF systems. It emphasizes that when fielding AI-enabled capabilities under operational conditions, DAF end-users, program offices, DevSecOps or AIOps teams, testers, and leaders must use a tailored AI Risk Management Framework (RMF), such as the National Institute of Standards and Technology (NIST) AI RMF, to address a series of risk-related questions at each stage of the AI life cycle. Chapter 3 develops a series of findings, conclusions, and recommendations:

Finding 3-1: The DAF will have similar training infrastructure requirements to support the development and maintenance of AI-enabled systems. The decentralized nature of DAF operations means training cannot be supported by standard commercial offerings. The committee knows of no commercial off-the-shelf solution presently supports these requirements.

Recommendation 3-1: The Department of the Air Force artificial intelligence testing and evaluation champion should outline and prioritize these training infrastructure requirements and coordinate with commercial providers to adapt available solutions accordingly.

Finding 3-2: The DAF has not yet developed a standard and repeatable process for integrating, testing, acquiring, developing, and sustaining AI capabilities.

Finding 3-3: OSD DOT&E has not yet published DoD-wide formal AI T&E guidance.

Finding 3-4: There is a lack of clear distinction between DT and OT phases for AI capabilities.

Conclusion 3-1: A lack of formal AI development and T&E guidance represents a considerable challenge for the DAF as AI-based systems emerge.

Recommendation 3-2: Department of the Air Force (DAF) leadership should prioritize artificial intelligence testing and evaluation (AI T&E) across the DAF with an emphasis on a radical shift to the continuous, rigorous technical integration required for holistic T&E of AI-enabled systems across the design, development, deployment, and sustainment life cycle.

Recommendation 3-3: The Department of the Air Force should track the progress of the International Organization for Standardization/International

Electrotechnical Commission TR 5469 working group report through the publication process and leverage it as a starting point for adapting their testing and evaluation processes for artificial intelligence–enabled systems.

Finding 3-5: DAF AI contributions to date have been focused on computer vision perception and natural language processing algorithms and have yet to extend to fully address system-level T&E.

Recommendation 3-4: The Department of the Air Force should adopt a definition of artificial intelligence (AI) assurance in collaboration with the Office of the Secretary of Defense Chief Digital and AI Office. This definition should consider whether the system is trustworthy and appropriately explainable; ethical in the context of its deployment, with characterizable biases in context, algorithms, and datasets; and fair to its users.

Recommendation 3-5: The Department of the Air Force should develop standardized artificial intelligence (AI) testing and evaluation protocols to assess the impact of major AI-related risk factors.

Chapter 4 proposes appointing a DAF AI T&E champion and explores the challenges in defining comprehensive T&E requirements for AI capabilities compared to traditional DAF weapon systems. The chapter discusses Project Maven as a requirements use case and recommends options for establishing AI T&E requirements and increasing interactions between system designers, developers, testers, program offices, and end-users throughout the AI life cycle. This chapter discusses the value of independent red teams as a critical component of the overarching requirements process and AI test design. Finally, in examining the role of culture and workforce development, the chapter observes the challenge of adapting a highly successful DAF test culture to the era of AI T&E. It emphasizes the immediate education, training, and certification steps that DAF leaders need to take to build and sustain an AI-ready test enterprise workforce.

Chapter 4 contains most of the committee's recommendations, as follows:

Finding 4-1: Currently, no single person below the level of the secretary or the chiefs of the Air and Space Forces has the requisite authority to implement DAF-wide changes to successfully test and evaluate AI-enabled systems.

Recommendation 4-1: The secretary of the Air Force and chiefs of the Air and Space Forces should formally designate a general officer or senior civilian executive as the Department of the Air Force (DAF) artificial intelligence (AI) testing and evaluation (T&E) champion to address the unique challenges of T&E of AI systems identified above. This AI T&E advocate should have the necessary AI and T&E credentials, and should be granted the requisite authorities, and

responsibilities, and resources to ensure that AI T&E is integrated from program inception and appropriately funded, realizing the DAF AI T&E vision.

Conclusion 4-1: Compared to traditional T&E, AI T&E requires radically deeper continuous technical integration among designers, testers, and operators or end-users.

Recommendation 4-2: The Department of the Air Force should adopt a more flexible approach for acquiring artificial intelligence (AI)-enabled capabilities that whenever possible links proposed solutions to existing joint capabilities integration and development system requirements, and that follows a development, security, and operations or AI for information technology operations/machine learning operations development methodology.

Recommendation 4-3: To the maximum extent possible and where it makes sense operationally, the Department of the Air Force (DAF) should integrate artificial intelligence (AI) requirements into programs of record, via the DAF's system program offices and program executive officers, and integrate AI testing and evaluation (T&E) into the host weapon system T&E master plan.

Recommendation 4-4: The Department of the Air Force should establish an activity focused on robust artificial intelligence–based systems red-teaming, implement testing against threats the red-teaming uncovers, and coordinate its investments to explicitly address the findings of red-team activities and to augment research in the private sector.

Recommendation 4-5: Building off the 2020 DoD Data Strategy, the Department of the Air Force should update and promulgate its data vision, strategy, and metrics-based implementation plan to explicitly recognize data as a "first-class citizen." These documents should include plans for the following:

- **Prioritizing investments in computation and storage resources and infrastructure to support artificial intelligence (AI) development**
- **Widely expanding data collection and curation for the entire range of AI planning and scoping, designing, training, evaluation, and feedback activities**
- **Investing in data simulators for AI training and testing**
- **Adapting approaches and architectures developed in private industry for AI-based systems**

Recommendation 4-6: The Department of the Air Force (DAF) should inculcate an artificial intelligence (AI) testing and evaluation (T&E) culture

espoused by DAF leaders and led by the AI T&E champion. In particular, the DAF and the DAF AI champion should:

- **Provide AI education, training, and, where applicable, certifications to all personnel, from general officers and senior civilian executives to entry-level personnel**
- **Institute career-long tracking and management of personnel with specific AI and AI T&E skills**
- **Place core AI T&E training under the Air Force Test Center**
- **Take advantage of existing AI-related education and training initiatives**

Recommendation 4-7: The Department of the Air Force (DAF) should determine the current baseline of artificial intelligence (AI) and AI test and evaluation (T&E) skills across the DAF, develop and maintain a tiered approach to AI and AI T&E-specific education and training, rebalance the test force by shifting people with needed expertise into the test enterprise, and consider placing personnel with AI T&E expertise into operational units.

Chapter 5 evaluates AI technical risks in DAF operational systems. It discusses how the employment of AI-enabled systems can have significant benefits in augmenting the capabilities of the warfighter. Still, it also notes that there are risks inherent in the use of AI-enabled systems that the DAF must address. This chapter observes that AI-enabled systems are vulnerable to several realistic performance issues, some based on adversarial AI attacks and others based on the risk of deploying the AI-enabled system in operational environments that have features or contexts that differ significantly from the representative datasets and intended contexts that were used to develop the AI capability. The chapter reviews the numerous attacks adversaries could potentially direct toward AI models within operational systems. The chapter observes that while AI models were subject to the same attacks as other software products, they were also vulnerable to AI-unique attack vectors that manipulated the training data, operational data, or the models themselves. It concludes that the DAF needed a staunch cyber defense as the first defense against such attacks. DAF T&E processes should likewise focus on detecting performance degradations and AI model susceptibility to classes of adversarial examples designed to avoid detection. Finally, it describes certain attacks, such as backdoor attacks involving adversarial triggers, that may be undetectable before they trigger with state-of-the-art test technologies.

The chapter discusses how academic research and development progress in this area has become an escalatory battle between attackers and defenders. Consequently, the chapter concludes that it would be important for the DAF to employ red-teaming of AI-based system vulnerabilities and to develop mitigations such as operational performance monitors. Furthermore, this chapter notes that DAF T&E would have an important role to play in emulating attacks identified by the red

teams and testing operational systems against these attacks. Finally, the chapter also discusses how AI models can fail in ways that are unexpected and non-intuitive. Therefore, it concludes that the DAF should focus on extensive testing to establish justified confidence in the deployed models.

Chapter 5 also makes a series of findings, conclusions, and recommendations:

Finding 5-1: Existing research on attacks on AI-enabled systems and strategies for mitigating them consider attacks that require unimpeded access to an underlying AI model. These attacks are unlikely to be practical with traditional protections and mitigations inherent in deployed DAF systems.

Finding 5-2: Ongoing research on adversarial attacks on AI-enabled systems focus on performance on benchmark datasets which are inadequate for simulating operational attacks. It appears that as robustness to adversarial attacks is improved, the performance often goes down. Even on benchmark datasets, the trade-off between potential performance reduction and improved robustness is not understood. More importantly, the defenses are designed to thwart known attacks. Such pre-trained defenses are not effective for novel attacks.

Finding 5-3: The impact of adversarial attacks on human-AI enabled systems has not been well understood.

Recommendation 5-1: The Department of the Air Force (DAF) should fund research activities that investigate the trade-offs between model resilience to adversarial attack and model performance under operational conditions. This research should account for a range of known and novel attacks whose specific effects may be unknown, but can be postulated based on state-of-the-art research. The research should explore mitigation options, up to and including direct human intervention that ensures fielded systems can continue to function even while under attack. The DAF should also simulate, evaluate, and generate defenses to known and novel adversarial attacks as well as quantitatively determine the trade-off between potential loss of performance and increased robustness of artificial intelligence–enabled systems.

Recommendation 5-2: The Department of the Air Force (DAF) should apply the DoD Zero Trust Strategy to all DAF artificial intelligence–enabled systems.

Conclusion 5-1: Promising areas of research that will improve the mitigation of adversarial AI include techniques for data sanitization, quantifiable uncertainty, and certifiable robustness.

Chapter 6 turns its attention to new and promising AI techniques and capabilities. It contends that even as the DAF addresses its current needs and opportunities, it must evaluate these emerging AI trends and their likely implications for T&E. Finally, the chapter observes that it is difficult to make precise predictions about which future AI capabilities will be most impactful for air force applications, especially given the accelerating rate at which AI technology advances. Nevertheless, it hypothesizes that five areas are particularly likely to impact DAF T&E: foundation models, informed machine learning, generative AI, trustworthy AI, and gaming AI for complex decision-making. It makes findings and recommendations accordingly.

Recommendation 6-1: The Department of the Air Force should focus on the following promising areas of science and technology that may lead to improved detection and mitigation of artificial intelligence (AI) corruption: trustworthy AI, foundation models, informed machine learning, AI-based data generators, AI gaming for complex decision-making, and a foundational understanding of AI.

Finding 6-1: Existing approaches for designing trustworthy AI-enabled systems do not consider the role of humans who interact with the AI-enabled systems.

Recommendation 6-2: The Department of the Air Force should invest in developing and testing trustworthy artificial intelligence (AI)-enabled systems. Warfighters are trained to work with reliable hardware and software-based advanced weapon systems. Such trust and justified confidence must be developed with AI-enabled systems.

Finding 6-2: Large language FMs exhibit a behavior termed "hallucination," where the model output is either non-sensical or is not consistent with the provided input or context. The effects of hallucination are task-dependent. There are no metrics to assess the impact of large FMs on the various downstream applications, they have been applied to.

Finding 6-3: Several large FMs are available for single modalies, with language being the most dominant one. DAF tasks may involve multi-modal sensing and inference. SSL-based large language models are just recently becoming available for multi-modal paired or unpaired data.

Finding 6-4: Physics-based and other knowledge-informed models have the potential to increase the robustness and computational efficiency of data-driven methods. These models can also help enforce physics or knowledge-based performance boundaries, which can increase the efficiency of the T&E process. However, to successfully deploy such models the DAF must ensure

that the parameters and assumptions upon which they are based are present during operations, which requires additional attention to operational T&E.

Recommendation 6-3: The Department of the Air Force should assess the capabilities of data generators to enhance testing and evaluation in the context of relevant applications.

Finding 6-5: Recent and anticipated advances in AI gaming technologies will enable the Air Force to build systems that are more capable than ever before and that involve AI in more sophisticated ways, but this increased system complexity will make the teaming relationship between the human and AI elements much more interrelated and complex, thereby placing additional challenges on effective T&E.

1

Introduction

1.1 A CENTRAL QUESTION: HOW TO ACHIEVE SUFFICIENT CONFIDENCE IN AI-ENABLED SYSTEMS?

The Department of the Air Force (DAF) is in the early stages of incorporating modern artificial intelligence (AI) technologies into its systems and operations. The integration of AI-enabled capabilities across the DAF will accelerate over the next few years. As demonstrated by experiences in commercial industry, the DAF will face new opportunities and challenges in integrating AI at scale.

AI is different from aircraft, missiles, and other weapons and support systems, with which the DAF has decades of experience in testing and evaluation. Existing T&E processes and procedures do not translate directly to software capabilities, especially AI's data-centric, black-box, self-learning, adaptive, and probabilistic characteristics. As a result, it is harder to gain buy-in from the DAF, DoD, public, and international communities for and sufficient confidence in AI-enabled capabilities absent the same kind of testing policies and processes for AI implementations that have guided flight testing for the past 70 years. While similarities between traditional and AI T&E mean that the DAF is not starting from scratch, the substantial differences between them make it imperative that the test community develop and promulgate AI-specific T&E policies and procedures as soon as possible.

The complexity of AI T&E is amplified by the inevitability of a future in hybrid weapons systems, including a combination of legacy non-AI systems, new non-AI systems, current or legacy systems with AI that are "bolted on," and AI that is "baked-in"—all of which may be operating together simultaneously.

Moreover, the T&E of AI-enabled systems must account for the cascading effects of multiple AI-enabled systems interacting across weapon systems, C2 architectures, and cyber networks.

One may argue that "the purpose of test, evaluation, verification, and validation (TEVV) . . . is the activity that produces the evidence that completes the needed assurance arguments."[1] Thus, while considering AI T&E requirements and the above factors, a central question becomes clear: *How to achieve sufficient confidence in AI-enabled systems?*

Similarly, what level of T&E is necessary and sufficient throughout an AI-enabled system's entire life cycle to ensure the delivery of effective, suitable, reliable, predictable, sustainable, secure, safe, trustworthy, and resilient capabilities?

As described in this report, the answer to this question will likely be considerably different for AI-enabled systems than for T&E of traditional hardware systems. It will be context-dependent, reflecting a combination of factors such as the degree of urgency; end-user requirements or operational imperatives; technology and human readiness levels (TRLs/HRLs); risks, such as threats, opportunity costs, and potential unintended consequences; scope; scale, and required levels of predictability, reliability, explainability, and transparency. While considering these factors, the DAF should be guided, though not unduly constrained by the precautionary principle—introducing a new product or process whose ultimate effects are disputed or unknown should be approached using caution, pause, and review.

Ultimately, the answer to how much testing is necessary and sufficient is defined as much by the end-user or operator as by the developers and the responsible DAF T&E organization. In all cases, end-users will assess the performance of AI-enabled capabilities relative to a given system's baseline (pre-AI) performance. This report focuses on three main tasks that the study committee was tasked with, which will help the DAF address the fundamental question of how much testing is enough.[2]

1.2 STUDY QUESTIONS TO BE ADDRESSED

The study committee was tasked with conducting this consensus study to examine the Air Force Test Center's (AFTC's) technical capabilities and capacity to conduct rigorous and objective tests, evaluations, and assessments of artificial

[1] D.M. Tate, 2021, "Trust, Trustworthiness, and Assurance of AI and Autonomy," Institute for Defense Analysis, https://apps.dtic.mil/sti/trecms/pdf/AD1150274.pdf, p. iv.

[2] As well as what kind of testing is necessary—see, for example, R. Burnell, W. Schellaert, J. Burden, et al., 2023, "Rethink Reporting of Evaluation Results in AI," *Science* 380:136–138, https://doi.org/10.1126/science.adf6369.

intelligence (AI)-enabled systems under operational conditions and against realistic threats. Specifically, the committee was asked to address three tasks:

Task 1 asks the committee to evaluate and contrast current testing and assessment methods employed by the DAF and in commercial industry. This is discussed in more detail in Chapters 3 and 4.

Task 2 asks the committee to consider examples of AI corruption under operational conditions and against malicious cyberattacks. This is discussed in more detail in Chapter 5.

Task 3 asks for recommendations promising areas of science and technology that may lead to improved detection and mitigation of AI corruption. This is discussed further in Chapter 6.

While the committee set out to maintain a narrow scope driven by the specified tasks, as the study progressed, it became clear that the impact of the findings and recommendations was more significant than anticipated. Central to the study's assessment of AFTC's technical capabilities is the capacity to deploy these new systems with the same rigor and discipline they have applied historically with traditional systems. To ensure the same level of trust that the test community has rightly earned from the space, cyber, and air forces, the processes, requirements, and culture of the test community and the DAF, in general, will need to evolve and adapt. These adjustments will be necessary to accommodate the developmental differences in AI-enabled systems.

1.3 WHAT DO WE MEAN BY "ARTIFICIAL INTELLIGENCE"?

Artificial intelligence (AI) is a broad term that means different things to different people. For example, AI can be broadly defined as a computing system that can perform tasks that are normally associated with human intelligence, such as conversing in a natural language, solving problems, and recognizing types of objects in a scene, etc. Generally, AI is defined to include all such tasks that a computing system can perform at human or near human proficiency levels. The ability to learn is also an important aspect of any intelligent system. Recently, major advances have occurred in the field of machine learning, leading to an increase in proficiency across virtually all current AI tasks. Thus, to many, AI and ML are often used synonymously today, although machine learning is just one, albeit very important subset of AI implementation.

For the purposes of discussing test and evaluation (T&E), the committee divides AI implementations into three broad categories: element, independent system, or joint cognitive system.

- The implementation may be an *element* of a program or system that uses an artificial intelligence algorithm or knowledge structure. Examples are a route planner, ground avoidance, machine learning for target detection.

- It may be an *independent system*, sometimes called a stand-alone, turnkey, or engineering system, that uses one or more intelligent elements coupled with other components (e.g., user interfaces) to produce results for a well-specified problem domain. Examples include Project Maven, logistic, and recommender systems.
- An implementation might be a *joint cognitive system*, where one or more elements or systems are coupled with additional elements supporting human-AI interaction solving a joint problem or task. Neither the human nor the engineered system is capable of individually achieving the desired outcomes. An example is the aircraft itself, where neither the pilot and nor the engineered components can with multiple heterogenous AI elements achieve the mission without the other.

In each of the three cases, the *breadth of the intelligence* is bounded by the types of problems of interest. Creating an optimal route planner is similar to a savant—someone that is extraordinarily smart, but only about some specialized field. The intelligence for an independent system is likewise specialized for the purpose, though the AI elements and the integration may be more sophisticated than a single element and constitute a systems-level form of intelligence. A joint cognitive system may be focused on only one problem or mission but solving that problem or executing that mission requires true interaction with a human operator to specify objectives, dynamically delegate and reacquire authority, supervise, coordinate, etc. A joint cognitive system must include social intelligence to support the interaction with the human. For example, the introduction of natural language chatbots to a system interface may lead the human to erroneously expect the system to behave as a joint cognitive system, further complicating T&E.

The *system complexity* of each of the three categories varies as well. An element is not a system in the traditional NASA engineering system sense and thus, while the algorithm may be quite sophisticated and pose its own T&E challenges, there are few complications for testing and evaluation due to concomitant hardware, user interfaces, other software, etc. The enabling intelligence in an independent system is much harder to discern because it is embedded in a larger engineering construct. The joint cognitive system category presents the hardest case as the system is a system-of-systems with the human operator as one of those systems.

A third dimension distinguishing the categories is the type of *user interaction*. An element would typically have minimal user interaction, as it exists as an embedded component of a system. A pilot may have a user interface for an independent system but might only be able to turn off or ignore the output from an element, assuming its contribution was obvious and accessible—for example, turning off the ground avoidance function. Joint cognitive systems differ from independent systems in that the user interaction for an independent system is generally through

user interfaces or fixed protocols; a joint cognitive system involves give-and-take interactions between the human and the computer, and introduces many of the human-machine teaming considerations discussed in Section 2.4.

These categories, summarized in Table 1-1, illustrate why testing and evaluation of AI is not a one-size-fits-all endeavor. Although the categories appear to restate decades of work in function (or unit), systems, and systems-of-systems testing, the breadth of intelligence, systems complexity, user interaction, and potential for both engineering and human error illustrates why AI imposes new demands on T&E. For example, a route planning algorithm element is straightforward to prove correctness as well as time and memory constraints but a convolutional neural

TABLE 1-1 Categories of Artificial Intelligence Implementation

	Intelligence	System Complexity	End-User Interaction	Examples	Ramifications for Testing
Element	Computation produces an exceptional, but narrow, skill or result	Little, generally producing a binary output or exists as a unit within a system	None, beyond either accepting or rejecting output	Route, ground avoidance, target detection, Bayesian models	Still involves functional or unit testing in isolation but depends on the unique vulnerabilities of the algorithms
Independent System	One or more elements coupled with other components to produce results for a well-specified problem domain	Generally, a function of the hardware, software, and application; involves combinations of different AI elements	Restricted; humans work through an interface or rigid protocols	Project Maven, logistics, recommender systems	Builds on systems testing principles but with multiple heterogenous AI elements
Joint Cognitive System	One or more systems coupled with additional social intelligence supporting human-AI interaction for a well-specified problem domain	System includes true human interaction	Fluid, the interaction mechanisms and partitioning of roles must dynamically support team behavior	Pilot-aircraft	System-of-systems, which includes human-machine cognitive testing

network element for target detection has much different structure and vulnerabilities. An independent system may involve intelligence to coordinate the intelligent components, complicating the already challenging systems testing landscape, as well as the issues such as human trust. A joint cognitive system is vulnerable to subtle mismatches between human and computer capabilities which are difficult to anticipate and simulate.

Broader applications of AI, such as artificial general intelligence, lie beyond the scope of this study. Furthermore, while some sections of the report focus on machine learning or large language models as prominent examples of AI, these are just examples—they are not, and should not be considered, the only target of this report.

Given the diversity of implementations, meanings, categories, and scope of applicability of the term AI, this study has used the general term AI without applying a series of qualifications each time the term is used. The report uses AI to refer variously to AI elements, independent systems, and joint cognitive systems. The report does not always call out the specific scope of the AI to which it refers in every instance, but the committee trusts that the context will make the meaning clear to the reader.

1.4 CURRENT STATE OF THE ART OF AI

The foundations for AI were laid through a seminal paper by Alan Turing and a 1956 summer workshop held at Dartmouth attended by some of the best-known researchers in the country. In the 1960s and 1970s, AI was focused on developing efficient search algorithms such as A* and playing games such as checkers, chess, etc. In the mid-1980s, uncertainty models were introduced as Bayesian networks and associated inference algorithms based on enumeration and variable elimination. In the 1980s, researchers pursued neural networks, which, although promising, did not lead to significant progress in AI then. Domain knowledge was the key contributor to the development of AI in the 1970s and 1980s. Starting in the 1990s, an explosion of automated data collection and computational processing power helped to seed a new age of data-driven AI. Since 2012, with the reemergence of deep learning algorithms (an expansion of the neural net concept), much of what is known as AI is based on learning from data using supervised, unsupervised, semi-supervised, or weakly supervised techniques, or reinforcement learning. More recently, generative AI models such as generative adversarial networks, diffusion models, and neural radiance fields have been used for generating synthetic data that can be used for deep learning. Over the decade, the main applications of AI have been in game playing, computer vision, natural language processing, advertising and marketing, and robotics. Figure 1-1 describes these developments.

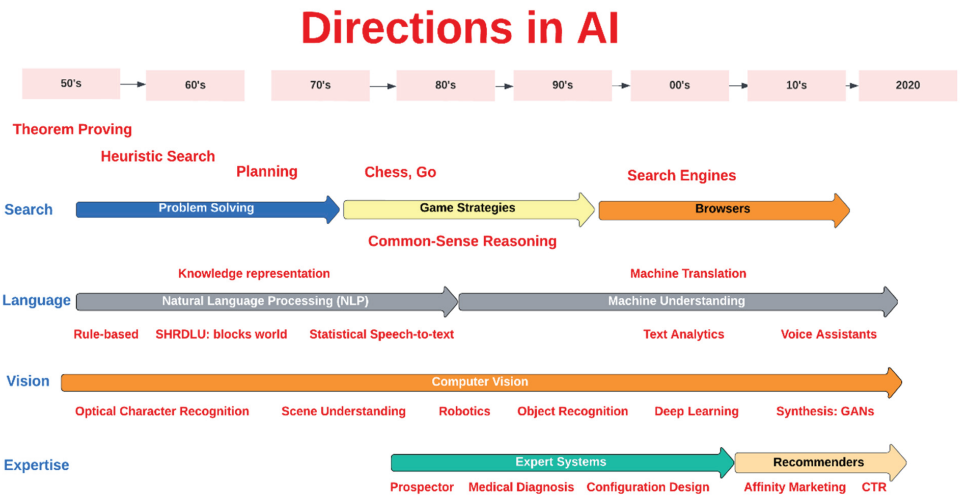

FIGURE 1-1 The history of AI.

1.5 CURRENT STATE OF THE PRACTICE OF AI IN THE DAF

The DAF is in the early stages of incorporating modern AI technology into its systems and operations. Through interviews with Air Force T&E leadership (full meeting agendas are available in Appendix C), the committee ascertained that, possibly apart from classified programs,[3] the Air Force has not yet acquired any modern AI capability within the standard acquisition processes for a major defense acquisition program (MDAP) or major automated information system (MAIS). AI-related projects to date have been research and development initiatives or proof-of-concept demonstrations or have been integrated into existing systems as upgrades or prototypes.

According to the Government Accountability Office (GAO) February 2022 report *Artificial Intelligence: Status of Developing and Acquiring Capabilities for Weapon Systems* (Table 1-2), as of April 2021, the Air Force had funded 80 projects that incorporated AI technology. Of those, research, development, test, and evaluation (RDT&E) funded 74, and 6 were acquisition procurement.

DAF AI-Based Prototypes and Demonstrations

The DAF has used proof-of-concept demonstrations and prototypes to motivate the value of AI and to increase its understanding of the processes, infrastructure needs, and specific challenges accompanying the integration of AI capability into its systems.

[3] The committee is unaware of classified programs that incorporate AI technology, but that does not mean that such activities have not taken place.

TABLE 1-2 Current DoD AI Efforts

	# of AI Projects		
DoD Component	R&D Funding	Procurement Funding	Total
DAF[a]	74	6	80
Army	209	23	232
Marine Corps	26	7	33
Navy	176	38	215
Other DoD Entities[b,c]	117	8	126

[a] DoD's methodology combined AI projects from the Air Force and Space Force.

[b] Other DoD entities include combatant commands and other unspecified DoD components included in the JAIC's methodology.

[c] DoD's initial inventory does not include classified AI projects of those funded through operations and maintenance.

SOURCES: GAO analysis of Department of Defense (DoD) information: GAO-22-104765, Appendix II.

For example, in 2020, the Air Force conducted a demonstration at Beale Air Force Base that integrated machine learning into a test aircraft. In a training flight, the AI algorithm controlled the sensor and navigation systems on a U-2 Dragon Lady spy plane. A test pilot oversaw the AI operation but did not intervene, although flight control always remained in the hands of the pilot. According to an interview[4] with Dr. William Roper, 13th Assistant Secretary of the Air Force for Acquisition, Technology, and Logistics, "Roper said the AI was trained against an opposing computer to look for oncoming missiles and missile launchers. The AI got the final vote for the initial test flight on where to direct the plane's sensors."

As another example, the Defense Advanced Research Projects Agency's (DARPA's) Air Combat Evolution (ACE) program has also been advancing the use of AI in DAF systems. As described by DARPA:[5]

> The ACE program seeks to increase trust in combat autonomy by using human-machine collaborative dogfighting as its challenge problem. This program also serves as an entry point into complex human-machine collaboration. ACE will apply existing artificial intelligence technologies to the dogfight problem in experiments of increasing realism. In parallel, ACE will implement methods to measure, calibrate, increase, and predict human trust in combat autonomy performance. Finally, the program will scale the tactical application of autonomous dogfighting to more complex, heterogeneous, multi-aircraft, operational-level

[4] A. Gregg, 2020, "In a First, Air Force Uses AI on Military Jet," *The Washington Post*, December 16, https://www.washingtonpost.com/business/2020/12/16/air-force-artificial-intelligence.

[5] R. Hefron, "Air Combat Evolution (ACE)," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/air-combat-evolution.

simulated scenarios informed by live data, laying the groundwork for future live, campaign-level Mosaic Warfare experimentation.

An early ACE success was the AlphaDogfight simulation contest. In this contest, an AI agent based on deep reinforcement learning beat a seasoned Air Force F-16 pilot 5-0 in a set of simulated one-on-one dogfights between two F-16 aircraft.[6]

AI prototype capability has also been integrated into the DAF Common Mission Control Center (CMCC).[7] The CMCC has incorporated a system called APIGEE (Automated Pipeline for Imagery Geospatial Enhancement and Enrichment), which does auto-mensuration of intelligence, surveillance, and reconnaissance (ISR) imagery to reference imagery to generate target-quality coordinates. APIGEE uses AI and deep learning to do image matching. The system currently performs electro-optical (EO)-to-EO mensuration, and additional multi-model capabilities are being tested for IR-to-EO and SAR-to-EO. CMCC has also incorporated a prototype Dynamic Electronic Order of Battle (EOB) capability that uses machine learning to develop patterns-of-life from electronic intelligence (ELINT) and identifies anomalies based on these normal behavior patterns. The project teams are migrating the EOB capability to the enterprise cloud as part of a CMCC "Common EOB" project.

Similarly, work has been ongoing with the Machine Assisted GEOint Exploitation (MAGE) program out of Air Combat Command (ACC) to incorporate AI into ISR processing, exploitation, and dissemination (PED) systems for geospatial intelligence (GeoINT) exploitation at defense geospatial service (DGS) sites. MAGE uses AI models to automate object detection workflows in various GeoINT products to support intelligence production. While actively under development, it still is not a fully deployed system and remains in development and evaluation.

The Air Force Collaborative Combat Aircraft (CCA) program pulls developed capabilities from other DAF programs such as Skyborg, AlphaDogfight, ACE, Variable In-Flight Simulator Aircraft (VISTA), and others. As the 96th Operations Group Commander briefed the committee, different aspects of each program, along with autonomy, data, and AI experimentation (ADAX),[8] will be used to help formulate AI T&E policies and processes across the DAF. Early observations include the challenges of building new or heavily modifying existing air vehicles—such as the XQ-58 Valkyrie and the Viper Experimentation and Next-Gen Operations

[6] E. Tegler, 2020, "AI Just Won a Series of Simulated Dogfights Against a Human F-16 Pilot, 5-0. What Does That Mean?" *Forbes*, August 20, https://www.forbes.com/sites/erictegler/2020/08/20/ai-just-won-a-series-of-simulated-dogfights-against-a-human-f-16-pilot-5-to-nothing-what-does-that-mean/?sh=7025a447235d; P. Tucker, 2020, "An AI Just Beat a Human F-16 Pilot in a Dogfight—Again," *Defense One*, August 20, https://www.defenseone.com/technology/2020/08/ai-just-beat-human-f-16-pilot-dogfight-again/167872.

[7] Private correspondence with Paul Metzger, MIT Lincoln Laboratory.

[8] Autonomy, Data, and AI Experimentation.

Model (VENOM) project, respectively—while incorporating extensive autonomy and, eventually, AI-enabled autonomy. At this stage, few AI-specific standalone capabilities are associated with these projects. However, the goal of Project Venom is to demonstrate autonomous capabilities on manned (and at some point unmanned) F-16s. The maturation of the CCA program will require a commensurate comprehensive AI T&E strategy and implementation plan.

Based on the committee's site visit to Eglin Air Force Base (AFB) and what various DAF representatives told the committee, it is evident from these prototype demonstrations that there is no well-established DoD-wide or DAF-wide set of standards for AI-based systems development or T&E. That is not to say that rigorous testing does not occur but that each project must develop its own T&E approaches and impose its own standards. However, what was also clear is that, due to the nature of the AI life cycle, early user involvement and continual T&E were essential elements of success.

A Case Study of Transition from Prototype to Initial Operating Capability

As the DAF seeks to transition prototypes to operational use, key aspects of AI-based systems acquisition have emerged (see Section 1.6 for a more in-depth case study). For example, Massachusetts Institute of Technology Lincoln Laboratory (MIT LL) transferred a Global Synthetic Weather Radar (GSWR) prototype to a software company called NextGen Federal Systems (NFS) for inclusion as part of the DAF weather forecasting system. The GSWR provides radar-like analyses and forecasts over regions not observed by actual weather radars by compiling lightning data, satellite imagery, and numerical weather models. The T&E process implemented to achieve the GSWR initial operating capacity (IOC) is instructive and underscores a few key points in acquiring AI-based capability. The process, in summary, was:[9]

- MIT LL developed a prototype capability based on prior work for the Federal Aviation Administration (FAA) and performed extensive testing using its access to global weather data. Several modifications to adapt the FAA model were required to tune and retrain GSWR to work in various geographic regions around the globe.
- Before officially transferring the software, MIT LL hosted a baseline implementation in the Amazon Web Services (AWS) government cloud. In addition, the contractor was given access to facilitate technology transition through a detailed assessment and refinement phase using an agile development process.

[9] From private correspondence with Dr. Mark Veillette, MIT Lincoln Laboratory.

- After this initial phase, NFS accepted the software, re-hosted the baseline as-is, and used the MIT LL prototype as a reference system to verify their implementation. In addition, MIT LL collaborated with NFS on the USAF cloud portal to update the software and make it compliant with DoD cyber security requirements. Much of this work entailed fixing code quality issues identified by scanners (e.g., SonarQube) and developing unit and integration tests to support future development operations (DevOps). During this time, NFS also integrated the baseline into Kubernetes[10] for improved cloud deployment.

The DAF is developing an in-house platform for tracking curated datasets, machine learning (ML) model training, and experimentation. Because GSWR has significant ML components, the datasets and training processes were also integrated into this platform for NFS to replicate MIT LL results.

This case study underscores a few key points that manifest in virtually all AI-based projects:

- Subject-matter experts (SMEs) in machine learning and weather forecasting were needed throughout all phases, from initial concept to IOC, and have remained involved beyond the IOC phase to facilitate a rapid and flexible DevOps process that integrates security requirements (DevSecOps).
- Extensive data was needed in the early research and development and deployment phases, and retraining for new geographic regions with new datasets was required. Curation, protection, and integration of data into the overall DevSecOps process were recognized as a necessary part of the engineering process. Retraining using operational data and the infrastructure to support this retraining were also key elements of success.

DAF AI Research and Development

The DAF, principally through the Air Force Research Laboratory (AFRL), is conducting or funding several research and development projects to advance AI implementations for the DAF. The projects span the following AI-related areas:[11]

- Basic AI research in the mathematics, information sciences, and life sciences
- AI applied to materials for structures, propulsion, and subsystems

[10] Kubernetes is an open-source container orchestration platform that automates many of the manual processes involved in deploying, managing, and scaling containerized applications.

[11] Department of Defense, 2022, *Fiscal Year (FY) 2023 Budget Estimates*, Office of the Secretary of Defense, Vol. 3 of 5 in *Defense-Wide Justification Book*, Washington, DC, https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2023/budget_justification/pdfs/03_RDT_and_E/OSD_PB2023.pdf. Some smaller RDT&E projects have been excluded.

- AI, automation, and autonomy for sensory evaluation and decision science
- AI for EO sensors and countermeasures technology
- AI applied to sensor fusion
- AI applied to C4I[12] dominance, battlespace development, and demonstration
- AI-enhanced life-cycle management
- Skyborg integrated technology demonstration

One technology demonstration of note is the AFRL-Air Force Life Cycle Management Center (AFLCMC) Skyborg project, one of the DAF's four Vanguard programs and a component of the DAF's overarching CCA project. Skyborg is an autonomous aircraft teaming architecture designed to increase the number of mission sorties while lowering costs. The program is investigating how AI-operated drones can team with human-piloted aircraft. Skyborg has established an open approach to autonomy architecture, building a scalable system designed to be portable across aircraft platforms and modular in its design to accommodate multiple software applications. Skyborg is intended to become a program of record in 2023 or 2024, depending on budget constraints. In 2022, the program executive officer (PEO) for AFLCMC's fighters and advanced aircraft directorate called Skyborg "wildly successful in terms of what we got out of it, what we continue to get out of it, and how we use that to present decision space to our leaders on how we set up programs of record."[13]

The AFRL-funded DAF-MIT AI Accelerator (AIA 1.0) is an example of a research and development project that has elements of core AI, enabling AI, and AI-enabled capability. The project's website (https://aia.mit.edu/about) provides project details and a succinct introduction to AIA 1.0. The latter is excerpted below:

In February 2019, the President of the United States signed Executive Order 13859 announcing the American AI Initiative—the nation's strategy on AI. He wrote, "Continued American leadership in Artificial Intelligence is paramount to maintaining the economic and national security of the United States."

The DAF subsequently signed a cooperative agreement with the Massachusetts Institute of Technology (MIT) to create a joint artificial intelligence Accelerator hosted at MIT. The effort, known as the DAF-MIT AI Accelerator (AIA), leverages the combined expertise and resources of MIT and the DAF. The AIA conducts fundamental research to enable rapid prototyping, scaling, and the ethical application of AI algorithms and systems to advance the DAF and society. A multidisciplinary team of embedded officers and enlisted airmen join MIT faculty, researchers, and

[12] Command, control, communications, computers, and intelligence.

[13] G. Hadley, 2022, " 'Wildly Successful' Skyborg Will Become Program of Record But Won't Stop Developing S&T," *Air and Space Forces Magazine*, August 16, https://www.airandspaceforces.com/wildly-successful-skyborg-program-of-record-developing-st.

students to tackle some of the most difficult challenges facing our nation and the air force, ranging from technical to humanitarian.

In January 2020, the AI accelerator launched ten interdisciplinary projects involving researchers from the MIT campus, MIT LL, and the DAF, as seen in Table 1-3. The 3-year projects, which encompass 15 research workstreams, advance AI research in various areas, including weather modeling and visualization, optimization of training schedules, and autonomy for augmenting and amplifying human decision-making.

While a major goal of the AIA is to develop core AI relevant to societal benefit and air force needs, the program is also developing tools, techniques, processes, and infrastructure that pioneer new DAF approaches to AI technology acquisition. Examples include the following:

- *Computational support for AI.* The "Fast AI" and "ML-Enhanced Data Collection, Integration, and Outlier Detection" projects focus on providing

TABLE 1-3 DAF-MIT AI Accelerator Projects

#	AIA Project Name	Project Type
1	Guardian Autonomy for Safe Decision Making	AI Core
2	Fast AI	AI Enabling
3	ML-Enhanced Data Collection, Integration, and Outlier Detection	AI Enabling/AI Core
4	Transferring Multi-Robot Learning Through Virtual and Augmented Reality for Rapid Disaster Response	AI Core
5	Conversational Interaction for Unstructured Information Access	AI Core
6	AI for Personalized Foreign Language Education	AI Core
7	Multimodal Vision for Synthetic Aperture Radar	AI Core
8	AI-Assisted Optimization of Training Schedules	AI Core
9	The Earth Intelligence Engine	AI Core
10	Continual and Few-Shot Learning: Transferring Knowledge to New Low Resource Domains and Tasks	AI Core/AI enabling
11	Explainable Machine Learning for Decision Support	AI Core/AI Enabling
12	AI Education Research: Know-Apply-Lead	AI Enabling
13	RAIDEN (Robust AI Development Environment)	AI Core/AI Enabling
14	Objective Performance Prediction and Optimization Using Physiological and Cognitive Metrics	AI Core
15	Robust Neural Differential Models for Navigation and Beyond	AI Core
16	AI-Enhanced Spectral Awareness and Interference Rejection	AI Core
17	Application of Coevolutionary Algorithms for DoD Complex Enterprises	AI Core
18	Space Domain Awareness	AI Core
19	Better Networks via AI-Enabled Hierarchical Connection Science	AI Core

efficient computational and data management capabilities to enable scalable AI development. The latter project also uses machine learning in its support algorithms ("AI for AI").

- *AI tools.* The "RAIDEN," "Continual and Few-Shot Learning," and "Explainable Machine Learning for Decision Support" projects focus on foundational AI advances to support T&E and responsible AI development.
- *Small, agile teams.* All projects by design include MIT researchers, MIT LL application and AI specialists, and DAF airmen and guardians working together as a team from project conception to transition. Transition partners are identified at the outset of projects and interact throughout the project life cycle. This diverse team composition encourages technology transition to the fir Force and air force user feedback to the researchers.
- *Challenge problems.* The open release of labeled datasets such as ImageNet has spurred the advancement of commercial and academic machine learning technology worldwide. Each AIA project defines a set of challenge problems and releases curated and labeled datasets to engage the broader AI research community. The datasets are unclassified but representative of key AI technology challenges in each project's research domain. Also, to facilitate challenge participation, the AIA developed a DoD-tailored data-sharing approach based on a University of California agreement that has been used for decades.[14]
- *Educational outreach.* Recognizing the importance of educating the DAF workforce, the "AI Education Research: Know-Apply-Lead" project was established to explore how to shape curricula and create course-ware to customize AI education for various learners with different needs and responsibilities.

Many AIA projects will seek to transition their technologies to air force stakeholders in the next few years. These transitions should provide opportunities to explore the efficacy of small-team AI development processes, including DevSecOps processes and continual T&E.

Summary

The DAF is only beginning to pursue AI technology for its systems and operations. To the committee's knowledge, no major DAF acquisition program (MDAP or MAIS) has incorporated modern AI technology beyond prototype capabilities and advanced concept demonstrations. In the absence of DAF and DOD AI-specific

[14] See University of California, 2022, "University of California Research Data Policy," VP-Research and Innovation, https://policy.ucop.edu/doc/2500700/ResearchData; Department of the Air Force-Massachusetts Institute of Technology Artificial Intelligence Accelerator, 2023, "Challenges Supplemental Resources," http://aia.mit.edu/challenges-supplemental.

standards, the acquisition and T&E processes adopted by these prototyping projects have been ad hoc, although they emulate sound commercial practices. Several AI projects are in the RDT&E pipeline, which motivates the need to address AI T&E for DAF as these technologies mature and become incorporated into DAF systems. Similarly, the 2022 establishment of the Autonomy, Data, and AI Experimentation proving ground at Eglin AFB as a joint venture between DAF CDAO and AFWERX is an encouraging initial step. Additionally, information is not always shared between the different pockets of AI work throughout the DAF.

This report aims to provide timely recommendations to help the Air Force establish effective T&E infrastructure and processes in anticipation of increased use of AI, especially applying AI technology to safety critical systems.

1.6 ALGORITHMIC WARFARE CROSS-FUNCTIONAL TEAM (PROJECT MAVEN) CASE STUDY

In April 2017 then-Deputy Secretary of Defense Robert Work established the Algorithmic Warfare Cross-Functional Team (AWCFT), or Project Maven, which reported to the deputy secretary through the under secretary of defense for intelligence (USDI). The AWCFT was the Department of Defense's first program to operationalize AI/ML at speed and scale. The AWCFT's primary objective was to accelerate the Department of Defense's integration of big data and machine learning and to "turn the enormous volume of data available to DoD into actionable intelligence and insights at speed."[15]

The AWCFT's first specified task was to field AI capabilities to augment, accelerate, and automate the processing, exploitation, and dissemination (PED) of full-motion video (FMV) from tactical and medium-altitude unmanned aerial systems (UAS). During the first year, Project Maven adopted and tailored commercially-developed computer vision (CV) algorithms for object detection, classification, and tracking. Its work subsequently expanded to include natural language processing (NLP) for exploitation of hard copy and digital materials collected during combat operations in the Middle East and East Africa, as well as machine translation, facial recognition, and SAR. Maven was also tasked to consolidate existing AI algorithm-based technology projects across the defense intelligence enterprise (DIE), including initiatives that developed, employed, or fielded AI, automation, machine learning (ML), deep learning (DL), and computer vision algorithms.

The initial cadre of Project Maven personnel lacked AI T&E experience, forcing them to rely extensively on T&E support from outside organizations. The primary participants were the Johns Hopkins University Applied Physics Laboratory

[15] Department of Defense, 2017, "Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)," Deputy Secretary of Defense, https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf.

(JHU APL), the Army and Air Force Research Laboratories, and several commercial companies with AI model fielding experience.

In the first few months of operations, the Project Maven team learned what has been noted elsewhere in this study, to wit, the substantial differences between AI T&E and T&E for traditional hardware systems. There was no DoD-wide AI T&E "playbook" for Maven to rely on. And because no other extant DoD-wide AI projects were dedicated to fielding AI-enabled solutions at scale, the OSD director of operational test and evaluation had not yet developed a standardized DoD AI T&E template or established AI T&E best practices.[16] Instead, individual users and organizations, primarily within the service research laboratories, had continued to develop boutique T&E processes, procedures, practices, and technical solutions tailored to their unique AI projects, the vast majority of which were research and development initiatives at relatively low technology readiness levels (TRLs).

The Maven team developed model performance benchmarks and other T&E criteria for each algorithm purchased from a commercial vendor (and subsequently trained against DoD data to become a DoD-licensed AI model). For computer vision algorithms, these included precision, recall, f-scores, intersection over union—more of a parameter than a metric in and of itself—and mean average precision. While each commercial vendor provided its own internal testing results, Maven insisted on reinforcing commercial testing results with additional, DoD-led tests and evaluation of each algorithm and trained model, using withheld test data to which the vendors were not exposed.

Since there were few examples of AI T&E within DoD apart from small-scale research laboratory projects, Maven adopted and adapted AI T&E best practices from the private sector and academia. These practices include setting aside sufficient representative, quality data for training, test, and validation or assessment; building T&E harnesses;[17] evaluating fielded models as part of ongoing operational assessments; evaluating model boundary conditions and AI failure modes; and developing T&E processes for each subsequent update to fielded models through normal CI/CD processes. The extent of T&E required for each subsequent model version depended on the breadth and depth of the changes included in each update. In most cases, later versions required a shorter T&E process than required during the first several updates to fielded models. In all cases, the Maven T&E team gained enough experience to accelerate T&E timelines. Maven also coordinated with commercial AI companies to establish contractual requirements for AI algorithm

[16] As of December 2022, OSD DOT&E had not developed AI T&E templates or promulgated AI T&E best practices.

[17] A T&E harness is "a software that tests model accuracy and other metrics." See JAIC, 2020, "JAIC Spotlight: The JAIC's Test Evaluation and Assessment Team Shapes Future AI Initiatives," CDAO blog, May 27, https://www.ai.mil/blog_05_27_20-jaic_spotlight_test_evaluation_and_assessment_team.html.

performance and detail intellectual property (IP) protections (although neither of these was entirely resolved under Maven; given the unique circumstances of each AI project, performance requirements and IP protections must be addressed separately for every AI development project—as discussed in more detail in the requirements section that follows).

In addition to developing an AI T&E "playbook," the Maven team worked closely with operational end-users, none of whom had any previous experience using AI-enabled systems and were unfamiliar with establishing requirements for or interpreting the metrics associated with AI T&E. Maven personnel "translated" T&E metrics into terms most relevant to operational end-users. Because formal requirements had not been established for AI model performance, once the Maven team had completed data quality assurance, T&E on each model, integration testing in the Maven Integration Lab, and live-fly testing; user acceptance of each trained model, and follow-on updates to those fielded models, was based primarily on an agreement between the Maven team and operational users that models had demonstrated adequate performance under operational conditions. As acknowledged elsewhere in this study, this process underscored the importance of defining future T&E requirements for all AI capabilities and AI-enabled platforms, sensors, and tools in ways that reflect consensus between developers and end-users at every stage of the AI life cycle.

After the first year of operations, Dr. Yevgeniya (Jane) Pinelis, who worked for the Institute for Defense Analyses (IDA) as a technical advisor to OSD DOT&E, moved to JHU APL to serve as their on-site representative to the AWCFT. As the Project Maven T&E team lead, Dr. Pinelis led the developmental and operational testing of AI algorithms, including computer vision, machine translation, facial recognition, natural language processing, and human-machine teaming. In addition, Dr. Pinelis relied on existing policies and standards from outside the department, particularly those established by the International Organization for Standardization (ISO), the Institute of Electrical and Electronics Engineers (IEEE), and the National Institute of Standards and Technology (NIST), to develop DoD-specific AI T&E policies, processes, procedures, and best practices in this role.[18]

Based on T&E lessons learned from Project Maven, the inaugural Director of the DoD Joint AI Center (JAIC), Lieutenant General Jack Shanahan,[19] established a test and evaluation directorate within the JAIC as part of the initial organizational structure. Dr. Pinelis served as the JAIC's inaugural chief of test and evaluation, and subsequently served as the chief of AI assurance in the OSD CDAO until her departure in early 2023. Dr. Pinelis extended her previous Maven T&E work to

[18] Once the deputy secretary of defense issued the "DoD AI Ethical Principles" in February 2020, the Maven AI T&E cadre was tasked with testing AI algorithms and models for operational effectiveness, robustness, resiliency, and alignment with those principles.

[19] Lieutenant General Shanahan (USAF, Ret.) served as a committee member for this study.

develop an AI T&E template for the JAIC, which became the accepted standard for defining T&E requirements, evaluating algorithm performance during model training and testing, and performing T&E on updates to fielded models.

In her role at the JAIC, Dr. Pinelis formed an AI T&E Community of Interest (CoI) across DoD and with academia and other government agencies, including the National AI Initiative Office (NAIIO), NIST, the Office of the Director of National Intelligence (ODNI), OSD DOT&E, the Test Resource Management Center (TRMC), the OSD Under Secretary for Research and Engineering (OUSD[R&E]), the military services, DARPA, Federally Funded Research and Development Centers (FRDC) and University-Affiliated Research Centers (UARC), and representatives from academia and industry. The CDAO has since published AI T&E playbooks and best practice guides, which are available to all government agencies and organizations, and launched a first-of-its-kind AI T&E bulk purchasing agreement that allows government components to access leading AI T&E commercial vendors.

A Project Maven vignette from 2019 to 2020 underscored the importance of rigorous and disciplined AI T&E and the need for government agencies to rely on in-house or disinterested third-party T&E to validate test results provided by commercial vendors. When evaluating the performance of a later version of a fielded AI computer vision model, T&E results indicated a decrease in performance compared to the previous model version. This was an unexpected result since all other earlier updates to the model demonstrated steadily-improving performance across all T&E metrics. Unfortunately, the results did not improve, despite repeated testing with additional test data. As a result, the team faced a decision of whether to field an update that was needed immediately by the operational end-users, under the assumption that there were unknown flaws in the T&E process rather than the model itself or delaying fielding until the unexpected results could be explained. They elected to delay fielding the updated version of the model.

After a detailed analysis of contributing factors, the team discovered that the commercial vendor responsible for the CV algorithm had lost several key data scientists over several months. Their replacements were not as familiar with the fielded model and provided an updated version of the algorithm that had been insufficiently tested. The performance of the updated model was not operationally acceptable—exactly as Maven's T&E results had indicated. The algorithm and model were improved, retrained, retested, and fielded. The Maven case study highlights how many AI T&E issues are technically feasible but organizationally challenging.

2

Definitions and Perspectives

An artificial intelligence (AI) system's capability relies on the data corpus used to create the AI. This depzndency on data presents interesting challenges to traditional test and evaluation (T&E) architectures. The following chapter's goal is to provide the reader with the fundamental definitions of these systems (Section 2.1), the role data plays in the AI life cycle (Section 2.2), and how T&E approaches have evolved in AI-enabled systems (Section 2.3). Additionally, the committee discusses the role of human-machine teaming in AI implementations and its impact on system T&E (Section 2.4). All of these distinctions warrant consideration when architecting AI-enabled systems for production deployment.

2.1 AI-ENABLED SYSTEMS

An AI-enabled system is a computer system that uses AI techniques to perform tasks that typically require advanced deductive or inductive tasks based on collected data of all types. These tasks include but are not limited to image and video analysis, predictive analytics, autonomous control and decision-making, and language and speech processing. AI-enabled systems have the potential to augment and enhance human capabilities.

There is a wide range of military applications for AI-enabled systems. A few key examples are as follows:

- Intelligence, surveillance, and reconnaissance: AI can be used to analyze large amounts of data from sensors and other sources to identify patterns and trends and detect potential threats.

- Target identification and tracking: AI can identify and track targets, such as vehicles or individuals, using sensor data and other sources.
- Cybersecurity: AI can detect and prevent or limit cyberattacks by analyzing network traffic and identifying anomalies that may indicate an attempt to breach security.
- Enhancing situational awareness: AI-enabled systems can analyze data from various sources, such as sensors and radar, to provide military pilots with real-time situational awareness and help them make better-informed decisions.
- Autonomous weapons systems: Autonomous weapons systems may use AI to make decisions and take actions without human intervention.[1]
- Training and simulation: AI can create realistic training environments and simulations for military personnel, allowing them to practice and improve their skills in a controlled setting.
- Navigation and flight planning: AI-enabled systems can be used for route planning, airspace management, and obstacle avoidance tasks.

All AI-enabled systems are produced by the same cyclical generation process, known as the AI life cycle. The three basic components of the life cycle are *training*, *inference*, and *re-training*. Each of those three components can be further broken down into constituent steps that capture the nuance in each phase, as depicted in Figure 2-1 (although, in practice, this process is not always as linear as the figure may suggest). When combined, problem framing, data processing, and model development represent the training phase of the life cycle and yield an AI model trained to perform a given task. The deployment phase represents the operationalization of the trained model into a system where it will be asked to perform its task on new, incoming data. This phase is also referred to as the inference phase. The monitoring phase informs model development and initializes any re-training that is required. Re-training is necessary if the model's performance deviates from what is expected once it is deployed and interacts with real-world data. The measurement and detection of that deviation, and the infrastructure and processes to manage it during training, inference, and re-training, are the subject of this report.

The training process for an AI-enabled system typically involves feeding the system large amounts of data and allowing it to learn from the patterns and relationships it discovers.

[1] Department of Defense, 2023, "DoD Directive 3000.09: Autonomy in Weapon Systems," Office of the Under Secretary of Defense for Policy, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. In DoD, autonomous weapon systems are governed by DoD Directive 3000.09, "Autonomy in Weapon Systems." This directive also governs the use of AI-enabled autonomous and semi-autonomous weapon systems.

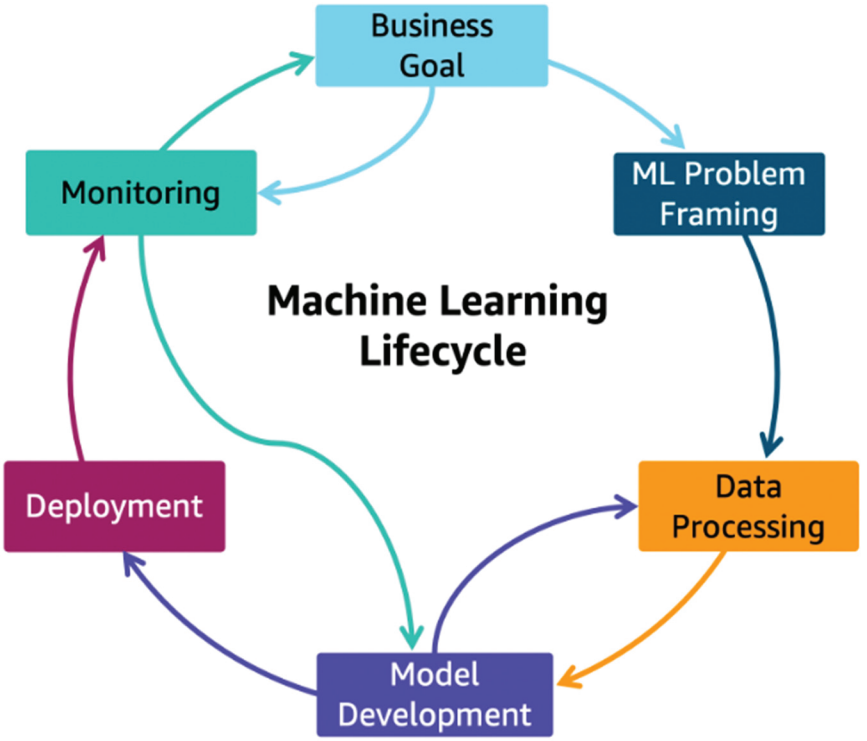

FIGURE 2-1 An example of a well-architected machine learning life cycle. SOURCE: Courtesy of Amazon Web Services.

2.2 ROLE OF DATA IN AI-ENABLED SYSTEMS

The amount of data that an AI requires can be enormous. For example, canonical computer vision models are trained on millions of images. Current state-of-the-art large language models have been trained on the content from millions of website pages. Training fleets for autonomous vehicle programs can produce petabytes a week that need to be processed. Data touch AI at every point in the AI life cycle, albeit in different ways, and is thus the most valuable part of the system. How data are collected, managed, curated (including labeling and assessing label quality), and maintained is crucial to the sustainability of AI systems.

Various statistical methods can be used to evaluate a model's performance over the training period during the training of any of these systems. These statistics can be directly related to the objective function of the learning process and help to determine if the model has converged on a reasonable solution. Or they can be independent measurements that describe some other characteristic of what the model has learned so far. These measurements indicate a model's performance relative to the data within its training corpus. For these measurements to have any

validity for operations, the training corpus must be representative of the defined operational deployment conditions. Therefore, it is a goal that the data be both sound and complete. In this context, a sound data sample is a valid sample from the input space—meaning it is a legitimate example of input data. The completeness of the data refers to how well the samples cover the input space—meaning all potential inputs have been sampled. Completeness is the distinctly more difficult characteristic to manage, yet it is incredibly important when fielding AI-enabled systems. If the completeness of a training corpus is poor, then a model will be unlikely to perform in an expected way in the field. Implicit in these requirements is that the label quality is highly emphasized.

In an AI-enabled system, operational data refers to the data collected and used by the system to perform its tasks. For example, in an autonomous vehicle system, operational data comes from various sources, such as sensors, cameras, and radar, and data from onboard systems, such as speed, location, and orientation. The AI system uses the operational data to draw conclusions about how to operate. For example, the AI system in a car may use data from the car's sensors to detect obstacles or traffic signals and make decisions about how to safely navigate around them. The system may also use data from the car's onboard systems to determine its location and track its progress toward its destination.

Unfortunately, real-world data are often complex, diverse, and constantly changing, making it impossible to capture and represent all possible scenarios in a single dataset. Furthermore, collecting data on all possible scenarios is often impractical or infeasible, as it may require extensive resources and time. Gathering data on rare or unusual events is extremely challenging, and the committee cannot gather data (except via simulations) on scenarios that have not yet occurred. As a result, it is generally necessary to make trade-offs when collecting and preparing data for AI training.

One common mitigation is to augment the data with synthetic or simulated data or use domain knowledge or expert insights to fill in gaps in the data. This approach may involve running full-scale simulations or modifying existing training data (e.g., different orientations). Usually, this augmentation is done in an iterative fashion wherein gaps in the learning are identified, and new data are added to fill in the gaps. It should be noted that identifying the gaps can be incredibly challenging in and of itself. Another common mitigation for many AI systems is learning continuously in real time, updating their underlying AI models based on new information. However, this is particularly vexing for testing since the system being tested is constantly changing.

These mitigations typically try to account for the lack of known or perceived completeness in the training corpus. However, even with all these adaptations, deployed models can still perform in unexpected ways when operational conditions change. The operational change can either be due to a domain shift or the advent of an unexpected scenario. This performance change is typically dealt with by retraining the model with the new observed data. The ability to re-train a deployed

BOX 2-1
Major Types of Training Techniques

The specific techniques and algorithms used to train the system can have a significant impact on its performance and capabilities. The general classes of learning techniques are supervised, unsupervised, self-supervised, and reinforcement.

Supervised learning is the most common type of algorithm training technique. It leverages labeled data, or ground truth, to learn what outputs align with which input values. The goal of this learning approach is to learn a function that approximates the relationship between the labeled inputs and outputs. One of the most popular supervised learning applications is in computer vision systems performing image classification or object detection.

Unsupervised learning learns without labels and instead consumes a corpus of data to learn the data structure. Typical examples of unsupervised learning include types of clustering or auto-encoders. These approaches can be useful for data exploration and dimensionality reduction.

Recent advancements in language models have been made using *self-supervised techniques*. Self-supervised learning attempts to correctly predict part of the input data by hiding components of the data and trying to guess the missing pieces from context. For example, a corpus of sentences or paragraphs may be used to train a large language model. Within that corpus, sentences are modified by hiding words during the training process. The algorithm's goal is to be able to predict the missing word. The prediction capability of the model improves by the model being exposed to more examples of similar sentence structure and word usage.

Reinforcement learning (RL) is a technique for training AI models at more complex, action-based tasks using learning agents within virtual environments. The virtual environments can be entirely virtual (i.e., a video game) or representative of the real world (i.e., real-world physical simulation). The RL agents are tasked with learning the ideal set of actions within their environment to be successful at some defined objective. It is similar to supervised learning in that the RL agents are trained using feedback based on the desired relationship between inputs to outputs. However, the feedback isn't simply the correct course of action when an agent guesses wrong. Instead, the agents decide on a set of actions within the virtual environment during the training process. The agents are given either rewards or punishments to encourage or correct the behavior given the defined objective. Over time the RL agents learn the optimal actions to achieve the objective.

model implies instrumentation in place at the deployment location to collect organized samples of the new scenario and either re-train in place or transmit the new data back to a training environment. Either way, the result of the retraining process is a newer version of the model that must be evaluated and redeployed. This process should continue for the lifetime of the model. Box 2-1 discusses some major types of training techniques.

2.3 HISTORY OF T&E IN AI-ENABLED SYSTEMS

Major applications of AI include computer vision, natural language processing, robotics, etc. Of these, computer vision applications have dominated largely due to decades of investment by DARPA and other agencies. Computer vision methods have found applications in many DoD and Intelligence Community applications

such as, automatic target recognition (ATR), image exploitation, 3D modeling and rendering, face recognition and identification, action detection and recognition, geolocation, navigation, etc., with the most recent example being Project Maven, described in Section 2.7. A recent survey[2] of DARPA's investment in computer vision and robotics summarizes the various DARPA programs since 1976 to date. The seminal paper on AlexNet, published in 2012,[3] is considered a major catalyst for the re-emergence of deep learning methodologies with a tremendous impact on computer vision.

The field of AI has gone through many phases. In early years, game playing, search algorithms, rule-based systems and constraint satisfaction problems drew the attention of researchers. As an example, the famous Waltz algorithm developed in the 1970s exemplifies a constraint satisfaction problem. In the 1980s, Bayesian graphical models were developed for addressing a wide variety of decision and inference problems. In 1982, the introduction of Hopfield networks created new enthusiasm for neural networks.[4] Traditional computer vision researchers were somewhat unappreciative of the three-layered neural networks that behaved like "black boxes." Since 2012, given the performance of AlexNet on the ImageNet challenge, the tide has turned. Deep convolutional neural networks developed by LeCun and subsequently adopted in Alexnet for addressing the ImageNet challenge have become dominant in computer vision. However, some of the doubts expressed by CV researchers in the 1980s and 1990s have not gone away! AI models based on deep learning are not eminently interpretable (the black-box structure still lives on). They are as fragile as traditional CV methods, if not more so, when data are intentionally corrupted by adversarial attacks. The question of bias and fairness has emerged as a serious concern, especially in applications such as face recognition. Many concepts in AI have not transferred to data-driven AI. Only recently, efforts to integrate symbolic processing into neural computations, leading to neuro-symbolic processing,[5] have emerged. The big issue is the inability of data-driven AI to incorporate rich domain knowledge. This is critical in domains such as medicine and, surely, in some DoD applications.

[2] T.M. Strat, R. Chellappa, and V.M. Patel, 2020, "Vision and Robotics," *AI Magazine* 41:49–65, https://doi.org/10.1609/aimag.v41i2.5299.

[3] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," paper presented at Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV.

[4] J.J. Hopfield, 1982, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences* 79(8):2554–2558, https://doi.org/10.1073/pnas.79.8.2554.

[5] G. Sir, 2022, "What Is Neuro-Symbolic Integration," *Towards Data Science*, February 14, https://towardsdatascience.com/what-is-neural-symbolic-integration-d5c6267dfdb0.

While CV has existed for many decades, T&E started in earnest in the early 1980s. There are many reasons for the lag between algorithm development and T&E in computer vision: Insufficient data, lack of agreement on appropriate metrics, and the notion of end-to-end CV systems was in its infancy. One of the standard image databases available then was from the University of Southern California (USC), which had images of textures, aerial images of buildings, and some other images that were more appropriate for evaluating image compression algorithms. In the 1980s, the DARPA Image Understanding Program, in collaboration with the defense mapping agency, ventured to develop performance metrics for aerial image analysis. With the emergence of more applied programs, such as Research and Development for Image Understanding Systems (RADIUS), Uncrewed Ground Vehicle (UGV), Moving and Stationary Target Acquisition and Recognition (MSTAR), Dynamic Database (DDB), etc., from DARPA and the Face Recognition Technology (FERET) program from the Army, more data became available and standardized metrics and T&E protocols were put in place. Some examples of large-scale evaluations are the evaluation of stereo algorithms,[6] face recognition algorithms,[7] and SAR target recognition algorithms.[8] The face recognition evaluations started in the early 1990s and morphed into a long-standing series of Face Recognition Vendor Tests[9] conducted by NIST. FRVT evaluations are ongoing, involving over 300 companies and other entities, and have kept pace with changing technologies.

In computer vision, T&E efforts are often packaged as challenges that go on for some years with active participation from companies and academic groups. Some examples are given below:

PASCAL VOC Challenge

The PASCAL Visual Object Classes (VOC) challenge[10] is a visual object category recognition and detection benchmark. The PASCAL VOC challenge includes three principal challenge tasks on classification, detection, and segmentation and

[6] J. Kogler, H. Hemetsberger, B. Alefs, et al., "Embedded Stereo Vision System for Intelligent Autonomous Vehicles," Pp. 64–69 in *2006 IEEE Intelligent Vehicles Symposium*, Meguro-ku, Japan, https://doi.org/10.1109/IVS.2006.1689606.

[7] DARPA, "AI Next Campaign (Archived)," https://www.darpa.mil/work-with-us/ai-next-campaign, accessed April 27, 2023.

[8] SAR, 2020, "DARPA Looking for SAR Algorithms," *SAR Journal*, August 3, http://syntheticaperturerадar.com/darpa-looking-for-sar-algorithms.

[9] National Institute of Standards and Technology, 2020, "Face Recognition Vendor Test (FRVT)," https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt.

[10] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, 2010, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision (IJCV)* 88:303–338, https://doi.org/10.1007/s11263-009-0275-4.

two subsidiary tasks on action classification and person layout. The challenge consisted of two components: (1) a publicly available dataset of annotated images with standardized evaluation tools and (2) an annual competition and corresponding workshop. The PASCAL VOC challenge was held annually from 2005 to 2012 and developed over the years. In 2005 when the PASCAL VOC challenge was first held, the dataset had only 4 classes and 1,578 images with 2,209 annotated objects. The number of images and annotations continued to grow, and in 2012 when the PASCAL VOC challenge was last held, the training and validation data had 11,530 images containing 27,450 ROI annotated objects and 6,929 segmentations. The PASCAL VOC challenge also witnessed a steady increase in performance over the years it was in operation. The PASCAL VOC challenge also contributed to establishing the importance of benchmarks in computer vision and provided valuable insights about organizing future challenges, which has led to a new generation of challenges, such as ImageNet.

ImageNet Challenge

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[11] is arguably the most influential challenge in the computer vision community. It was run annually from 2010 to 2017 and has become the standard benchmark for large-scale object recognition. ILSVRC consists of a publicly available dataset, an annual competition, and a corresponding workshop. ILSVRC also followed the practice of PASCAL VOC, where the annotations of the test set were withheld from the public, and the results were submitted to the evaluation server. The backbone of ILSVRC is the ImageNet dataset. ILSVRC uses a subset of ImageNet images for training the algorithms and some of ImageNet's image collection protocols for annotating additional images for testing the algorithms. ILSVRC scaled up to 1,461,406 images and 1,000 object classes in ILSVRC 2010. The ILSVRC challenge witnessed the success of deep learning and convolutional neural networks (CNNs). The breakthrough came in 2012 when a deep CNN called AlexNet achieved a top-5 error of 15.3 percent, more than 10.8 percentage points lower than the runner-up. Since then, the best-performing algorithms in ILSVRC have been dominated by deep CNNs. For example, in 2015, ResNet[12] achieved a 4.94 percent top-5 error rate on the ImageNet 2012 classification dataset, the first algorithm that surpasses human-level performance (5.1 percent) on the dataset.

[11] O. Russakovsky, J. Deng, Hao Su, et al., 2015, "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)," *International Journal of Computer Vision (IJCV)* 115(3):211–252, https://www.image-net.org/challenges/LSVRC/index.php.

[12] C. Shorten, 2019, "Introduction to ResNets," *Towards Data Science*, January 24, https://towardsdatascience.com/introduction-to-resnets-c0a830a288a4.

AICity Challenge

NVIDIA Corporation, in collaboration with research groups across academia, launched an annual challenge known as AICity in 2017 to invite teams globally to compete for the development of high-performing systems for a few transportation-related tasks, including monocular speed estimation, city-scale multi-camera re-identification and tracking, anomaly detection, and natural language-based vehicle retrieval. These tasks address important and long-lasting problems in transportation which currently require many transportation analysts and significant amounts of time to solve. However, automating these tasks can provide actionable insights promptly. Traffic data for these tasks have been mainly collected from traffic cameras in the state of Iowa, curated and open-sourced accordingly. Since 2017, these datasets have motivated thousands of researchers to participate in the challenge, significantly contributed to the landscape of AI-powered transportation algorithms and extended the boundary to which computer vision can help the transportation industry.

2.4 HUMAN-MACHINE TEAMING

The modeling of human-AI (HAI) and human-machine team (HMT) interactions and how these interactions enhance or diminish human effectiveness and efficiency evolved from decades of research and development in human-systems, human-computer, and human-robot interactions. The Department of the Air Force (DAF) has over 70 years of experience integrating hardware and software of ever-increasing complexity into aircraft and spacecraft cockpits, back-end mission compartments, command and control and intelligence systems, and ground control stations.[13] Aircrews' comfort level in interacting with hardware and software systems in airplanes and spacecraft has derived primarily from extensive T&E and user feedback focused on issues such as ease of use, systems integration, reliability, failure modes, resilience, and upward compatibility. In the pre-HAI era, the combination of well-defined and well-understood developmental, operational, and follow-on T&E processes and extensive operational use led to a broad understanding of potential hardware and software failure modes, hardware and software fault or failure indications, and corrective actions. This approach generated a level of confidence that led to greater trust in aircraft systems. At the same time, lessons learned from performance limitations linked explicitly to poor cockpit design during the Vietnam War led to extensive attention on human-systems

[13] The Air Force Research Laboratory (AFRL) 711 Human Performance Wing, for example, serves as a DAF and Joint DoD Center of Excellence for human performance sustainment, readiness, and optimization. The 711th's areas of research include biological and cognitive research, warfighter training and readiness programs, and systems integration.

teaming and aircraft cockpit and back-end ergonomics in the following decades. This culminated in developing DAF aircraft and space systems designed explicitly with human-system integration as a primary consideration.[14] So far, DoD has been focused on AI that augments humans, rather than AI systems that might serve as teammates.

As one of the briefers to this committee noted however (see Appendix E), AI is fundamentally different: it is useful "to consider AI as a teammate of a different species."[15] This briefer suggested considering interactions between humans and AI-enabled machines as similar to how humans interact with their pets—a considerable difference from all previous human-machine interactions that acknowledges the differences inherent in rapidly evolving AI capabilities. AI's incredible potential will never be unleashed without changing how humans and machines interact in a more digitized future. Optimizing the integration of humans and AI-enabled machines, which in turn depends on redesigning human-machine interfaces and recalibrating human and machine roles and responsibilities, will be one of the most important and defining features of an AI-enabled future.[16] When used operationally, a human-AI system is evaluated by how well the human team using that system performs their given tasks. One fundamental measure of success is when the human-AI system performs better than either the human alone, the AI system alone, or the previous version of the human-AI system.[17]

As learned during Project Maven and with other AI projects at the DoD JAIC/CDAO, there are inherent limitations in adding an AI implementation as a "sidecar" to legacy systems instead of baking in AI while developing new systems. However impressive, the performance of AI-enabled legacy systems inevitably reaches a plateau. System design and development must be completely revamped to get the most out of a human-smart machine team—similar to, but much more extensive than, the approach taken to cockpit redesign in the aftermath of the Vietnam War. This applies to *all* AI-enabled systems, not only those embedded in aircraft or spacecraft. It demands a different approach to training humans to work with "smart"

[14] The Boeing 737 Max aircraft crashes serve as a recent glaring example of poor human-system design.

[15] N. Cooke, 2022, "Effective Human-Artificial Intelligence Teaming," presentation to the committee, June 28, Washington, DC: National Academies of Sciences, Engineering, and Medicine.

[16] This includes establishing the interdependencies between humans and machines. On the point of human-machine interdependence, see, for example, M. Johnson and J.M. Bradshaw, 2021, "How Interdependence Explains the World of Teamwork," Pp. 122–146 in *Engineering Artificially Intelligent Systems*, W.F. Lawless, et al., eds., Vol. 13000 of *Lecture Notes in Computer Science*, https://doi.org/10.1007/978-3-030-89385-9_8.

[17] The committee acknowledges the difficulty of formulating and assessing HSI-related measures of performance and measures of effectiveness, especially during the fielding of early versions of AI models that often initially perform worse than the non-AI system they are designated to replace.

machines, unlike any previous systems, to better understand how humans and machines learn and improve through repeated interactions and interventions.[18] The committee underscores the importance of taking human-system integration and human-AI team effectiveness into account during the T&E of AI-enabled systems, including considering human needs and limitations as early as possible in the design and development process.[19]

Integrating AI elements into DAF weapon systems, decision support systems, and back-end office systems raises new challenges in the design, build, deployment, and employment stages, as well as in T&E. In addition to the human-systems integration issues noted above, T&E programs must address how to deal both with the best-case—and potentially unexpected—outcomes and unpredictable failures of AI-enabled systems. For example, AI-enabled systems are sensitive to domain shift, subject to adversarial attacks, and generally lack explainability or transparency (as discussed in Chapter 5). Additionally, a smart machine may assume more responsibilities from the human over time as human confidence in the system grows. As a result, the effective workload distribution between the end-user and the AI systems will change continuously. As systems become sufficiently advanced, the roles, responsibilities, and interdependencies between human and machine could constantly and seamlessly shift back and forth based on what humans and machines do best in each situation. In some cases, an AI-enabled machine may operate at speeds greater than the human operator is accustomed to or is comfortable with. These characteristics must all be dealt with during the AI T&E life cycle, but there are presently no clear standards to address them.

The concept of human-centered design, such as world-class user interface and user experience (UI/UX), is at the heart of every successful modern commercial software product. Too often, however, user-focused, intuitive UI/UX has not been prioritized while developing government systems. UI/UX must be one of the primary criteria considered during the design phase of all future AI-enabled systems. In a future environment characterized by the widespread fielding of AI-enabled systems, maximum performance can only be achieved by focusing on superior human-system integration, or what the U.S. Special Competitive Studies

[18] See, for example, this story on the new roles of AI "prompt engineers": D. Harwell, 2023, "Tech's Hottest New Job: AI Whisperer. No Coding Required," *The Washington Post*, February 25, https://www.washingtonpost.com/technology/2023/02/25/prompt-engineers-techs-next-big-job.

[19] See, for example, National Academies of Sciences, Engineering, and Medicine, 2022, *Human-AI Teaming: State-of-the-Art and Research Needs*, Washington, DC: The National Academies Press. This report was written as a result of the Air Force Research Laboratory (AFRL) 711th Human Performance Wing's request to the National Academies to examine the requirements for appropriate use of AI in future operations.

Project (SCSP) refers to as human-machine cognitive collaboration (HMC).[20] In an AI-enabled digital future, users of AI-enabled systems analysts should be able to train with smart machines so that those systems adapt to an individual's preferences, the pace of their cognitive development, and even their past behaviors. As technology advances rapidly, highly-tailored human-machine interaction and interdependence are achievable.[21] Human-machine testing and training processes will be vital to better understanding human-machine team composition, optimal assignment of human and machine roles and responsibilities, and effective and efficient workflow integration.[22]

In addition, these efforts should consider both technology readiness levels (TRLs) and human readiness levels (HRLs) and incorporate continuous assessments of human-machine team performance.[23] The consideration of HRL, while always important, becomes critical for AI-enabled systems that depend on continuous human interaction, as opposed to traditional pre-AI systems that primarily report results for human consideration. Substantial and sustained human intervention has compensated for poor HRLs in earlier and current fielded military systems. In future AI-enabled systems, human-machine integration must accord equal consideration to HRL and TRL. Otherwise, the DAF can expect sub-optimal

[20] This report makes the distinction between human-machine cognitive collaboration (HMC) and human-machine combat teaming (HMT). HMC focuses primarily on cognitive tasks, while HMT "will be essential for more effective execution of complex tasks, especially higher-risk missions at lower human costs" (p. 25). See Special Competitive Studies Project, 2022, *Defense Interim Panel Report: The Future of Conflict and the New Requirements of Defense*, Arlington, VA, https://www.scsp.ai/wp-content/uploads/2022/10/Defense-Panel-IPR-Final.pdf.

[21] Johnson and Vera examine the importance of human-machine "teaming intelligence." They argue that "no AI is an island" (p. 18), meaning that there is no such thing as a completely autonomous system. Humans are always involved at some level, from design and development through oversight of fielded systems. M. Johnson and A. Vera, 2019, "No AI Is an Island: The Case for Teaming Intelligence," *AI Magazine* 40(1):16–28, https://doi.org/10.1609/aimag.v40i1.2842.

[22] Guillory and Carrola refer to the concept of "Cognitive Mission Support," which they define as systems designed to help humans deal more effectively with the inevitability of information and cognitive overload (p. 3). S.A. Guillory and J.T. Carrola, 2022, "What Online-Offline (O-O) Convergence Means for the Future of Conflict," Information Professionals Association, August 3, https://information-professionals.org/what-online-offline-convergence-means-for-the-future-of-conflict.

[23] Technology Readiness Levels (TRLs) are a method for estimating the maturity of technologies during the acquisition phase of a program. TRLs are measured from TRL 1 (basic principles observed) to TRL 9 (actual system proven in its operational environment). The Human Factors and Ergonomics Society notes that "many system development programs have been deficient in applying established and scientifically-based human systems integration (HSI) processes, tools, guidance, and standards, resulting in suboptimal systems that degrade mission performance" (p. 1). This draft paper includes a useful table assessing program risk due to HRL-Technology Readiness Level (TRL) misalignment (p. 6). See Human Factors and Ergonomics Society, 2021, *Human Readiness Level Scale in the Human Development Process*, ANSI/HFES 400-2021, draft version, Washington, DC, https://www.hfes.org/Portals/0/Documents/DRAFT%20HFES%20ANSI%20HRL%20Standard%201_2_2021.pdf.

results from both humans and machines. During testing, training, and fielding, overall human-system team performance can be optimized through continuous human feedback to the system as it returns its results.

Human-AI interactions are more intriguing and interesting as all such interactions are based on trustworthiness. A recent review of the literature has identified the following six grand challenges that need to be addressed before efficient, resilient, and trustworthy HAI systems can be deployed.The six challenges are "developing AI that (1) is human well-being oriented, (2) is responsible, (3) respects privacy, (4) incorporates human-centered design and evaluation frameworks, (5) is governance and oversight enabled, and (6) respects human cognitive processes at the human-AI interaction frontier."[24] The committee recommends that the DAF adopt and adapt similar principles (although this list does not reflect a prioritization scheme) when designing, developing, testing, fielding, and sustaining AI-enabled systems (in Chapter 3, the committee discusses in more detail the core concepts of trust, justified confidence, AI assurance, and trustworthiness).

The pace at which humans and AI make decisions is a challenge for HAI systems. Whether humans will trust smart machines is a major concern. It is especially acute when considering AI's role in supporting combat operations. However, since trust is typically perceived as a binary yes-or-no concept, rather than considering whether end-users will "trust" their AI-enabled machines, we should consider instead how users gain *justified confidence* in smart systems over time. The process of building confidence in any AI-enabled system is continuous and cumulative. It never ends. A user's confidence in a smart system will depend on context: the nature and complexity of the task, the system's previous performance record, the user's familiarity with the system, and so on. In general, continued successful performance in lower-risk, lower-consequence tasks will give users more confidence in using AI when facing higher-risk, higher-consequence tasks. Until users gain more experience teaming with smart machines, they will face the dilemma of placing too much or too little confidence in their AI-enabled systems. The committee develops these concepts further in Chapter 3.

Finding 2-1: The DAF has not yet developed a standard and repeatable process for formulating and assessing HSI-specific measures of performance and measures of effectiveness.

Conclusion 2-1: The future success of human-AI systems depends on optimizing human-system interfaces. Measures of performance and effectiveness, to include

[24] O.O. Garibay, B. Winslow, S. Andolina, et al., 2023, "Six Human-Centered Artificial Intelligence Grand Challenges," *International Journal of Human-Computer Interaction* 39(3):391–437, https://doi.org/10.1080/10447318.2022.2153320.

assessments of user trust and justified confidence, must be formulated during system design and development, and assessed throughout test and evaluation and after system fielding.

Recommendation 2-1: Department of the Air Force (DAF) leadership should prioritize human-system integration (HSI) or HSI across the DAF, with an emphasis on formulating and assessing HSI-specific measures of performance and measures of effectiveness across the design, development, testing, deployment, and sustainment life cycle.

3

Test and Evaluation of DAF AI-Enabled Systems

The previous two chapters summarized the history of artificial intelligence (AI), ongoing Department of the Air Force (DAF) AI projects and defined key AI and AI test and evaluation (T&E)-related terms and definitions. This chapter begins with a synopsis of the air force's historical approach to traditional flight T&E. Section 3.2 discusses OSD and DAF T&E policies for AI-enabled systems (noting, as applicable, where there are still gaps in the formulation of AI T&E-specific policies). Section 3.3 addresses the importance of DevSecOps or artificial intelligence operations (AIOps)/machine learning operations (MLOps) to the design, development, testing, fielding, and sustainment of national security and commercial sector AI-enabled systems. The speed of AI advances in the commercial sector over the past decade has included the commensurate design and deployment of T&E methodologies for AI-enabled commercial systems (autonomous vehicles, large language models, chatbots, recommendation engines, and so on), although DoD systems are more complex, consequential, and subject to more regulation than commercial systems. Section 3.4 presents a detailed discussion of these developments. In Section 3.5, the committee examines the core concepts of trust, justified confidence, AI assurance, and trustworthiness and how together they play an instrumental role in gaining end-user buy-in for fielded AI-enabled systems. Finally, Section 3.6 closes the chapter with a consideration of the critical importance of risk management throughout the entire AI life cycle, including risk awareness, analysis, acceptance, accountability, and responsibility.

3.1 HISTORICAL APPROACH TO AIR FORCE TEST AND EVALUATION

The Air Force Test Center (AFTC) was established in 1951 to consolidate aircraft, missiles, and other systems' testing and evaluation functions under a single organization; to standardize and streamline test processes; ensure consistency in T&E practices; deal with the rapid growth in numbers and types of air force aircraft entering fielding; and reduce unacceptable aircraft mishap rates. Today, the AFTC conducts developmental and follow-on T&E of manned and unmanned aircraft and related avionics, flight control, munitions, and weapon systems. The AFTC comprises the Arnold Engineering Development Complex (AEDC) at Arnold AFB, the 96th Test Wing (TW) at Eglin AFB, the 412th TW at Edwards AFB, and the Test Pilot School (TPS) at Edwards AFB. The 96th TW is the T&E center for air-delivered weapons, navigation, and guidance systems; command and control systems, and AF Special Operations Command systems. It is the principal AF organization for command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR) developmental testing, often in coordination with the Air Combat Command's 505th Command and Control Wing (a subordinate unit of the U.S. Air Force Warfare Center). The 412 TW plans, conducts, analyzes, and reports on all flight and ground testing of aircraft, weapons systems, software, components, and modeling & simulation (M&S). The 412 TW flies an average of 90 aircraft and performs over 7,400 missions (over 1,900 test missions) annually. The USAF TPS at Edwards AFB trains pilots, navigators, and engineers on how to conduct flight tests.[1]

Understanding the historical context of AF T&E is important to conceptualize the changes needed to effectively and efficiently test and evaluate AI and autonomous systems. The AF test process has always focused on data collection, while evaluation emphasizes data analysis and comparing expected-to-actual performance to support decision-making. T&E is accomplished through a T&E master plan (TEMP), which contains thresholds and objectives, evaluation criteria, and milestone decision points. The TEMP is developed by the designated program management office (PMO). Traditionally, AF T&E has been divided into two primary components: developmental (D) and operational (O) T&E. At the basic level, DT&E centers on safety of flight concerns, while OT&E focuses on tactics and operating concepts. DT&E is conducted throughout the acquisition process to assist in engineering design and development and to verify that technical performance specifications are achieved. It includes the T&E of components, subsystems, hardware and software integration, and production qualification testing. DT&E examines the system's compliance with contractual requirements

[1] Air Force Test Center, 2021, "Fact Sheet: Air Force Materiel Command," https://www.aftc.af.mil/About-Us/Fact-Sheets/Article/2382275/air-force-materiel-command.

and the ability to achieve key performance parameters (KPP) and key system attributes (KSA).

OT&E, on the other hand, measures the overall ability of a system to accomplish a mission when used by representative personnel in the environment planned for the operational employment of the system. It conducts independent evaluations, operational assessments, and the ability to satisfy KPPs and KSAs. OT&E is conducted under realistic operational conditions, as close as possible to those expected in combat operations. The objective of OT&E is to determine a system's operational effectiveness, operational suitability, survivability, and lethality for combat. It is a mission capability assessment.

For aircraft and aircraft systems, DT&E and OT&E have traditionally been treated as two distinct phases of T&E that do not overlap. If a system under test fails DT&E, the engineering design and development process must be addressed before testing. Once a system passes DT&E, it transitions to OT&E. If it fails OT&E, it reverts to DT&E to re-evaluate its technical performance specifications and ability to comply with contractual requirements. Once a system passes OT&E, it is cleared for operational fielding. After initial fielding, it will be declared to have achieved initial operating capability (IOC), a formal milestone noting that an operational (non-test) unit can employ the system effectively. Once IOC is declared, the system may require further development and testing to achieve its full capabilities. Once that occurs, the system will be declared fully operational capability (FOC). The FOC milestone is achieved when a system has demonstrated it can perform all its intended missions and functions in various operational environments and is fully integrated into the overall operational structure—the operational unit can employ and maintain the system. It is not unusual for FOC to be declared for several years after the IOC milestone, especially for more complex weapon systems. The FOC milestone represents completion of a system's T&E and development efforts.

As discussed in more detail in the following sections, extant T&E processes that have worked so well for most DAF weapon systems over the past 70 years, with a clear delineation between DT&E and OT&E, were not designed to be applied to T&E of AI implementations and software and consequently fail.

3.2 AI AND DEVSECOPS/AIOPS IN THE DAF AND COMMERCIAL SECTOR

The DAF has been transitioning from waterfall to agile development methodology, albeit at its traditional pace. The transition initially only encompassed basic development and deployment processes but has expanded to incorporate security evaluations earlier in the development process (DevSecOps). The migration from waterfall to DevSecOps-based processes is largely instigated by the

software's increasingly wide and deep footprint in the complex systems the DAF deploys. While modern software has been a large catalyst for this evolution, the development and deployment of AI capabilities will be a true forcing function. AI will introduce speed into decision systems due to the simple automation of traditionally human-driven tasks and the ability to process previously insurmountable amounts of data. Additionally, the AI life cycle is inherently iterative and requires infrastructure to enable the continuous maintenance and improvement required on deployed models. The increase in pace will be an additional stress on traditional development and evaluation infrastructure.

It is a certainty that deployed AI models will encounter operational conditions not represented in the original training corpus and behave in unanticipated ways. A simple example of an unanticipated behavior could be an AI model labeling an object in an image incorrectly because it has never seen the object in training. This trite example is appropriate for illustration but can easily be extended to higher-risk and higher-consequence scenarios. There are significant implications of an AI-enabled system mislabeling an object that subsequently informs a high-risk, high-consequence targeting decision. Because of the guarantee of encountering unknown scenarios, adopting agile development improves the T&E processes for AI-enabled systems.

AI implementations are developed cyclically, often referred to as either AIOps or MLOps, and require continuous training, evaluation, and retraining as operational conditions change. Figure 3-1 shows a generic architecture and the cyclical feedback required to enable AI deployment to the edge. No organization can manage this production cycle and develop high-performing AI systems without using agile development methodologies that integrate T&E across the AI life cycle. Deployed models require "maintenance" that addresses shifts in operational conditions not represented in training data. This architecture is not a substitute for the safety systems and processes encompassing deployed AI systems; however, you cannot safely and effectively deploy AI without this iterative approach. For AI-enabled

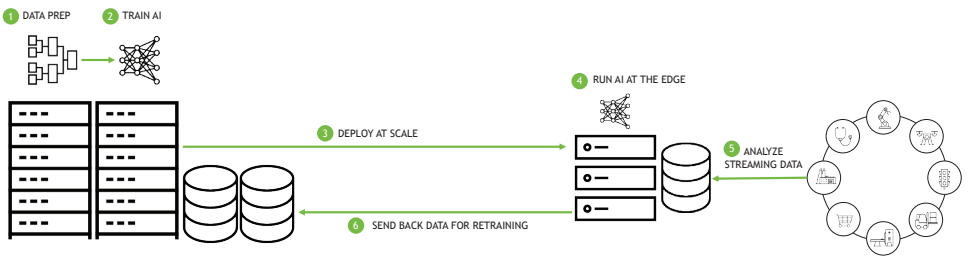

FIGURE 3-1 A generic architecture and the cyclical feedback required to enable AI deployment at the edge. SOURCE: Courtesy of NVIDIA.

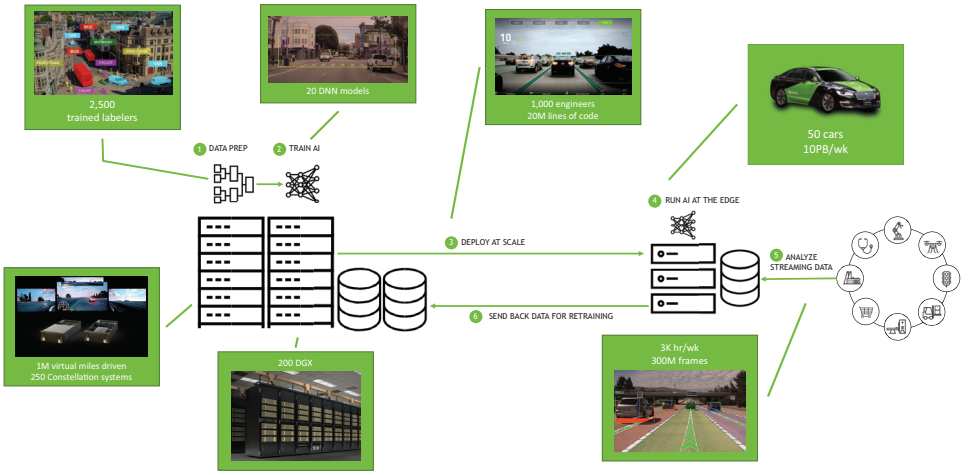

FIGURE 3-2 Connections between the development, testing, and deployment of the AI capabilities required to deploy an autonomous vehicle. SOURCES: Courtesy of NVIDIA; DGX image courtesy of NVIDIA and Oak Ridge National Laboratory, Department of Energy.

systems, the DAF is currently not prepared for this level of continuous integration and continuous deployment or delivery.

For example, Figure 3-2 illustrates the connection between the development, testing, and deployment of the AI capabilities required to deploy an autonomous vehicle. Of significance in this example is not only the scale of infrastructure and tooling to create the original models but also the supporting fleet of cars that continually collect more operational data and refine the deployed models. Some autonomous vehicle systems will selectively record data correlated with an AI-driver disagreement to reduce the rate of data to curate and improve model performance. In addition, model creation and refinement are supported by a robust simulation architecture for handling edge cases and domain shifts known a priori, as well as those observed operationally. This simulation environment supports both the creation of synthetic data as well as hardware-in-the-loop training.

Key components of this architecture include:

- *Trained labelers:* Labelers are trained on tooling and the data they are labeling.
- *Continuous monitoring, retraining, and redeployment of AI models:* Model performance is constantly monitored. Models are regularly retrained and redeployed.

- *Instrumented deployment platforms to capture ML-ready data:* Both the deployed models and the data streams they consume must be instrumented to capture the behavior deviation and the observations that manifested the performance shift.
- *Synthetic data engines and supporting digital twins:* Enable faster incorporation of emergent threats, observed domain shifts, or previously unknown edge cases. These components must be built for the appropriate domains and modalities.[2]

These three components have distinct implications for traditional T&E processes in the DAF. Methodology and infrastructure are required to detect when model behavior has deviated from expected performance during operations, to retrain the model with the new, associated observations, and then to evaluate how the new model performs under both previous and newly-observed conditions. Integral to the retraining of models in operation is the ability to retrieve these observations. For the DAF, this implies platforms that adopt AI-enabled systems or components require the capabilities to record ML-ready data from their sensors and associated actuators and then send that data back to a training environment in an easily consumable form. These requirements have significant impacts beyond test processes. They must be accounted for in platform and sensor operational requirements, up to and including at the PMO or system program office level.

Similarly, synthetic data engines and digital twins are key to supplementing datasets with training examples for situations where there is either insufficient real data or data are too difficult to collect. Synthetic data engines and digital twins must be relevant, adaptable, and considered to be part of the AI life cycle. For the DAF, sensor models and situational constructions of interest are represented in modern modeling and simulation environments that can keep pace with the cadence required for maintaining a collection of supporting AI models.

There are MLOps solutions on the commercial market today that facilitate this life cycle. The solutions are varied and support myriad deployment scenarios. In combination with commercial T&E vendors, some of these solutions can supplement an organization with the infrastructure and processes required to maintain an enterprise deployment of AI-enabled systems. Larger institutions that were early adopters of AI integration have evolved large internal systems and tooling to support their requirements. However, the DAF is neither one of the early adopters

[2] For a private industry example of unprecedented use of a synthetic environment to virtually "build" and evaluate an entire automobile factory and production line 2 years before physical production begins. See BMW Group, 2023, "BMW Group at NVIDIA GTC: Virtual Production Under Way in Future Plant Debrecen," *PressClub Global*, March 21, https://www.press.bmwgroup.com/global/article/detail/T0411467EN/bmw-group-at-NVIDIA-gtc:-virtual-production-under-way-in-future-plant-debrecen?language=en.

nor is it feasible for the DAF to be a direct consumer of unmodified commercial solutions. There are fundamental requirements rooted in operational requirements and constraints of DAF systems that demand a different approach. Figure 3-3 highlights the areas where the DAF's requirements can be met with gaps in current commercial architectures. These gaps represent the areas where the DAF needs to invest in modifying commercial solutions to meet service needs.

Real-time operational testing implies the need to continuously maintain and test AI-enabled solutions once they are operational. This represents a fundamental departure from the traditional waterfall approach that characterized historical DAF T&E efforts and is a critical change from current approaches. This change in approach is necessary to handle domain shifts and edge cases. Commercial solutions will certainly incorporate methodology for monitoring and retraining models, but it is unlikely they will incorporate processes that capture the complex system integration and risk frameworks that apply to DAF systems, especially safety-critical systems in the foreseeable future. The DAF should invest in synthetic data engines, live virtual constructive environments, data repositories, and support for digital twins representative of their modalities and platforms of interest to facilitate rapid model retraining and maintenance. Data standards must be extended to the platforms to support this retraining and enable fast capture of AI-ready data to facilitate retraining around model failure events.

Many commercial MLOps solutions assume constant, high-bandwidth connectivity to the AI-enabled systems they support, with many of their deployment patterns dependent on commercial cloud infrastructure. This assumption breaks down in most DAF operational environments, especially during crises or conflicts. Many forward-deployed organizations will not have the luxury of high-bandwidth

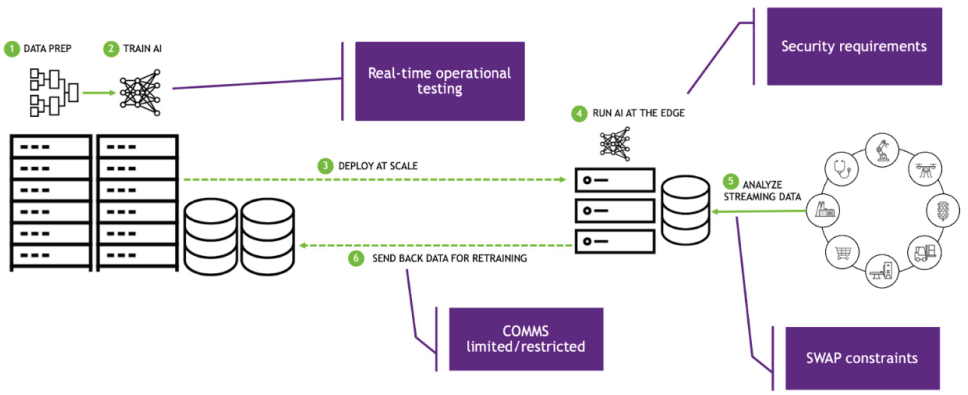

FIGURE 3-3 Areas where the DAF's requirements can be met with gaps in current commercial architectures. SOURCE: Courtesy of NVIDIA.

data connections back to large MLOps factories to retrain and retest model updates. The decentralized nature of forward-deployed operations likely requires some edge-based computing for model maintenance and testing while in the field, along with trained personnel capable of retraining and retesting models under suboptimal conditions. Model updates produced at any edge node would also need to eventually be transmitted back to some centralized management system, implying a federated learning model. The DAF AI T&E champion should outline and prioritize these requirements and coordinate with commercial providers to adapt available solutions accordingly.

Finding 3-1: The DAF will have similar training infrastructure requirements to support the development and maintenance of AI-enabled systems. The decentralized nature of DAF operations means training cannot be supported by standard commercial offerings. The committee knows of no commercial off-the-shelf solution presently supports these requirements.

Recommendation 3-1: The Department of the Air Force artificial intelligence testing and evaluation champion should outline and prioritize these training infrastructure requirements and coordinate with commercial providers to adapt available solutions accordingly.

Developing and deploying AI-enabled systems implies that there are companion deployment systems designed to receive and run the trained models in operations. These systems would comprise the sensors and reasoning systems that leverage the deployed models to extract information or make decisions. Any AI-enabled system that needs to operate at the edge—whether originating from commercial or military sources—will have unique size, weight, and power (SWAP) challenges. The DAF deploys systems and platforms that typically have bespoke security and SWAP requirements. These requirements will likely be constraints to the edge computing architectures that complement and integrate with commercial MLOps solutions. AI-enabled systems require high-performance computing solutions—typically graphics processing units (GPUs) or field programmable gate arrays (FPGAs)—to run AI models. Deployment configurations vary dramatically across DAF platforms, making the integration of these devices challenging and time-consuming. This fractured and bespoke approach to computing requirements limits the DAF's ability to drive features into these commercial products and results in costly customizations per platform that are not repeatable or cost-effective.

The DAF should invest in standards that enable the consolidation of computing requirements into fewer modular configurations designed to meet the needs of AI and autonomous systems. Through consolidation, the addressable

market for these solutions becomes larger and more feasible for commercial vendors to tackle at a scale that would accelerate time to market and reduce cost. It would also reduce the test footprint by simplifying test configurations for DT&E and OT&E.

3.3 OSD AND DAF T&E POLICIES FOR AI-ENABLED SYSTEMS

The DAF's current requirements formulation and acquisition processes continue a tradition of directly testing various capabilities against functional requirements under expected operational deployment conditions. As noted in previous chapters, the rapid introduction of AI-enabled capabilities across the DAF over the next several years requires an assessment of the applicability of established DAF-wide T&E approaches to AI and revising current policies or developing new ones that apply specifically to AI T&E. This is equally true whether AI or hML is integrated into a program of record, or added separately after a system has already been fielded. Based on presentations by DAF test enterprise leaders to this committee, the committee concludes that the DAF has not yet developed a standard and repeatable process for integrating, testing, and sustaining AI capabilities in DAF major acquisition programs. The few examples the committee knows of, such as Project Maven, consist of capabilities added to major programs (such as Air Force Distributed Common Ground Station (AF DCGS)) after fielding, outside of the traditional program of record acquisition processes. As one speaker commented, "We are not classically trained to do this [type of] T&E."

> Finding 3-2: The DAF has not yet developed a standard and repeatable process for integrating, testing, acquiring, developing, and sustaining AI capabilities.

Much like the advances in DAF-wide T&E for C2, cyber, and ISR systems over the past decade, the committee expects the DAF to make up ground with AI T&E relatively quickly. This assumes that DAF leaders prioritize AI T&E accordingly, applying sufficient resources in funding, infrastructure, policies, and personnel management. Despite its current shortcomings, the DAF is no further behind in AI T&E than most other government organizations and agencies. The DAF can take advantage of the extensive work already carried out by OSD CDAO developing AI T&E policies, processes, and frameworks, as well as applying lessons learned from commercial companies that have substantially advanced their internal AI T&E processes over the past several years (the committee includes some examples later in this chapter). This is an opportune time for the DAF to craft an AI T&E vision and commit to a long-range AI T&E strategy and implementation plan that includes specific and measurable objectives and goals.

At the OSD level, myriad instructions, directives, and policies referenced throughout this report exist to guide T&E. However, most of these are not AI-specific, and OSD DOT&E has not yet published DoD-wide formal AI T&E guidance.[3] Moreover, as noted elsewhere in the report, there has been limited direction addressing the lack of a clear distinction between developmental test (DT) and operational test (OT), or between initial operational T&E (IOT&E) and follow-on operational T&E (FOT&E), for AI capabilities. This represents a considerable challenge for the department.[4]

Finding 3-3: OSD DOT&E has not yet published DoD-wide formal AI T&E guidance.

Finding 3-4: There is a lack of clear distinction between DT and OT phases for AI capabilities.

Conclusion 3-1: A lack of formal AI development and T&E guidance represents a considerable challenge for the DAF as AI-based systems emerge.

As noted in the Project Maven case study, the JAIC T&E division refined Maven's T&E processes, procedures, and practices, and under their new organizational structure, the OSD CDAO is publishing AI T&E playbooks and providing AI T&E frameworks to OSD DOT&E. These include frameworks for testing AI-enabled systems, human-system integration (HSI), operational test,

[3] OSD director of operational test and evaluation (DOT&E) is an independent entity whose director does not report to the secretary of defense, but to Congress. The DOT&E director is the principal staff assistant and senior advisor to the Secretary of Defense on operational test and evaluation in DoD. The DOT&E mission is to issue DoD OT&E policy and procedures; review and analyze the results of OT&E conducted for each major DoD acquisition program; provide independent assessments to the Secretary of Defense, the under secretary of defense for acquisition and sustainment (USD(A&S)), and Congress; make budgetary and financial recommendations to the secretary regarding OT&E; and oversee major DoD acquisition programs to ensure OT&E is adequate to confirm operational effectiveness and suitability of the defense system in combat use. DOT&E is tasked to assess operational effectiveness, suitability, survivability, and sustainability. The organization currently relies on red teams for evaluation of DoD cyber capabilities but does not presently manage any AI-specific red teams.

[4] While the committee realizes this is an OSD-level concern, it recommends that the DAF AI T&E champion coordinate with OSD (especially OSD CDAO, OSD DOT&E, the DASD [DT&E], and the test resource management center or TRMC), the joint staff, and the military services to explore organizational solutions that address the lack of clear lines and lanes between AI developmental and operational test and evaluation. Also, as noted in the report summary, the DAF AI T&E champion should assess what broader DAF-wide organizational changes are called for to reflect the differences between AI T&E, and T&E for all other air force systems and capabilities.

and operationalizing responsible AI. Also, as noted elsewhere, the 96th Operations Group is developing AI T&E academic materials and curricula, and the DAF-MIT AIA is developing an AI T&E Guidebook (which will not be official policy). Finally, in 2020 the AFTC's 412th Test Wing published *Test and Evaluation of Autonomy for Air Platforms*, a technical information handbook.[5] While it deliberately does not address AI-enabled autonomous systems, it could be modified to address the T&E of AI-enabled autonomous systems and promulgated DAF-wide.

Until DOT&E and the DAF publish and promulgate formal AI T&E guidance, the committee recommends that the DAF consider adopting the OSD CDAO's AI T&E playbooks and frameworks. The tri-center should adapt these documents based on Air and Space Force AI T&E requirements, modifying them as necessary once OSD DOT&E promulgates official department-wide AI T&E directives, policies, and instructions. It is worthwhile to integrate the appropriate commercial best practices, which are documented in Chapter 3.

DOT&E is in the early stages of formulating AI T&E guidance. As a DOT&E official[6] told the committee, DOT&E recognizes that the department is "not where we need to be . . . with respect to even machine learning, never mind AI."[7] He noted that AI T&E is a young field, with very few, if any, operational use cases across the department, almost no DoD-wide AI T&E best practices,[8] and almost no historical military AI T&E studies or reports to fall back on. The official echoed a critical question from this report's summary, namely, for AI-enabled learning systems, "How much testing is enough?" He also emphasized addressing where, how, and when AI testing is accomplished. He acknowledged not only that DOT&E's "tried-and-true" test designs of the past were insufficient for fully testing AI-enabled systems, but that DOT&E did not yet possess the same kind of tried-and-true test designs or processes for AI. He noted that agile principles (see Section 3.3) were critical in developmental test and postulated that they would be equally important

[5] R.A. Livermore and A.W. Leonard, 2020, *Test and Evaluation of Autonomy for Air Platforms*, Edwards, CA: 412 Test Wing, Edwards Air Force Base, https://apps.dtic.mil/sti/pdfs/AD1105535.pdf.

[6] M. Crosswait, 2023, presentation to the committee, September 28, Washington, DC, National Academies of Sciences, Engineering, and Medicine.

[7] The same official urged the committee to make available to the entire department this report's findings and recommendations, with the goal of accelerating the development and promulgation of AI T&E best practices DoD-wide.

[8] With the exception of the OSD/CDAO's development of AI T&E playbooks and frameworks, which CDAO has provided to DOT&E. The committee expects DOT&E will publish their own frameworks, modeled after CDAO's products, after concluding their ongoing comprehensive review of all DOT&E test guidance.

in operational test (while acknowledging that DOT&E had not yet determined what exactly this would entail for AI-enabled systems).[9]

In addition to underscoring the importance of developing an AI T&E culture and supporting the development of a more operationally relevant AI T&E risk management framework (RMF), DOT&E is analyzing how to test AI-enabled systems for unexpected outcomes (to include testing boundary conditions and system behavior under varying conditions);[10] how training, validation, and test data should be selected and evaluated as part of the overall AI T&E process[11] (to include assessing security vulnerabilities and susceptibility to adversarial attack); how to account for the black-box nature of AI models; how to evaluate user trust and justified confidence in AI-enabled systems (under both expected and unanticipated operational conditions); and how to assess the ability of AI models to adapt to different missions and in different domains. Also, similar to DOT&E's extensive use of cyber red teams during operational test and evaluation (to include when integrated into combatant command exercises), it intends to evaluate the effects of adversarial attacks on AI-enabled systems, especially those systems designated as mission- and safety-critical.[12]

For DoD systems performing mission- or safety-critical missions, especially those capable of generating lethal effects or that can lead directly to generating lethal effects, the committee agrees with DOT&E's recommendation that, before fielding a substantive update to an AI capability, such systems must be operationally tested before the new version is fielded operationally. This type of "mini-OT" can be accomplished in the future by testing the updated capability in a digital twin or with an equivalent modeling and simulation architecture under operationally-realistic conditions (simulated or actual). Additionally, some processing architecture is required for system testing including data acquisition, cleaning, and labeling. The goal is to test updated capabilities as rapidly as possible (to include validation, verification, and accreditation) based on operational requirements, test conditions,

[9] The official emphasized that for non-AI-enabled programs under his purview, adherence to agile principles has led to high-performance fielded systems, especially several Missile Defense Agency (MDA) missile defense projects. He suggested that a large part of the success derives from the principle, described in the Requirements section of this report, of early and frequent interaction between users, developers, program managers, and testers throughout the entire life cycle of a program.

[10] The DOT&E official explained that one of the organization's primary objectives is to ensure that, for AI-enabled systems in DoD, the "intolerable outcomes" do not occur (while acknowledging that the definition of intolerable outcomes had to be determined for each AI-enabled system and integrated system-of-systems).

[11] The DOT&E official mentioned the possibility of maintaining the equivalent of a "war reserve mode" (WRM) of AI training data, that could be used to continue to develop and sustain AI models in the aftermath of adversarial attacks against existing datasets or fielded models.

[12] OSD DOT&E has relied on DOT&E-sponsored and service-led cyber ted Teams for the past several years. See for example, DOT&E, 2022, "Cyber Assessment Program," FY 2021 Annual Report, https://www.dote.osd.mil/Portals/97/pub/reports/FY2021/other/2021cap.pdf.

and the use of test personnel with experience testing the original fielded model and system.[13] Understanding the limitations of data drift, domain adaptation, and AI model boundary conditions (see Section 4.3) will help to increase the level of confidence that a certified system will operate as expected once deployed. Based on lessons from Project Maven and the JAIC, the committee expects that the extent of T&E required for each subsequent AI model version will depend on the scope of the changes included in each update (this is also the state of industry best practices; see Section 3.5). In most cases, later versions of fielded models will require a shorter T&E process than during earlier updates. Backed by industry best practices (see Section 3.5), updates to AI-enabled mission- or safety-critical systems require full transparency between testers, developers, and end-users to ensure all stakeholders have a common understanding of how much additional T&E is required and acceptable before fielding each update.

In summary, OSD DOT&E has provided an initial roadmap for how to redesign T&E for DoD AI-enabled systems to reflect the substantial differences between the T&E of traditional DoD systems and the T&E of AI capabilities. It does not, however, currently have the resources or the expertise, nor is the necessary foundational knowledge available, to make the changes needed to move beyond vision to immediate DoD-wide implementation. While DOT&E provides further guidance in the form of official policies, directives, instructions, templates, and frameworks, the committee recommends that in the near term, the DAF continue to work closely with DOT&E, the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)), and the CDAO AI T&E community of interest while adopting or adapting T&E best practices from across the government (for example, OSD CDAO's AI T&E playbooks and frameworks), the private sector, and academia. The committee recommends that the DAF AI T&E champion focus on new test designs for AI-enabled systems that incorporate the core systems engineering principles of non-AI-enabled systems[14] while adding new elements that reflect the best AI T&E practices from academia, commercial industry, and other government test organizations.

[13] As noted elsewhere in the report, the Missile Defense Agency (MDA) makes extensive use of modeling and simulation (M&S) during missile defense system design and testing, to include integrating actual hardware and M&S as part of an overall design, development, and testing architecture (through the command and control, battle management, and communications [C2BMC] program).

[14] To include, for example, principles that straddle the traditional and AI T&E worlds such as MIL-STD-882F, the replacement to MIL-STD-882E, *DoD Standard Practice: System Safety*, 11 May 2012. This system safety standard practice is a key element of systems engineering (SE) that provides a standard, generic method for the identification, classification, and control of hazards. The revised document will include a section on AI and ML, to include the AI criticality index (AICI), which will be used to determine the level of rigor (LOR) of software assurance activities to be imposed on the software. Department of Defense, 2012, *DoD Standard Practice: System Safety*, MIL-STD-882F, Washington, DC, https://cdn.ymaws.com/system-safety.org/resource/resmgr/documents/Draft_MIL-STD-882F.pdf.

The introduction of AI-enabled capabilities into the Air Force and Space Force has been limited and has proceeded slowly. The DAF has not addressed the pervasive implications of AI throughout the DAF or how T&E has to be integrated throughout the entire AI life cycle, from design through sustainment. The DAF has not yet committed to making immediate, sustained investments in AI governance, workforce development, AI research and development, AI development and T&E infrastructure, AI standards and practices, and targeted experimentation. The DAF has not developed the digital infrastructure needed to support AI development and T&E, and the requisite investments have not been programmed into the DAF budget. The lack of a designated AI T&E champion at the senior executive or general officer level, with commensurate SECAF-delegated authorities and resources at their disposal, has contributed to the low priority accorded to AI T&E across the DAF.

To ensure that the future AI-enabled Air Force and Space Force remain the most capable, responsible, and safe defense forces in the world, the committee recommends that DAF leaders prioritize AI development and T&E and address the implications across the entire DAF, including committing the necessary level of resources—people and funding. As a key initial step, the DAF should update its AI T&E vision and commit to a long-range AI T&E strategy and implementation plan that includes specific and measurable objectives and goals. The DAF, in coordination with OSD CDAO, should update its analysis of the resources required for digital modernization across the Air and Space Forces to reflect AI T&E-specific requirements, and sustain those resources in future DAF budgets.[15] The DAF should leverage investments from OSD CDAO, OSD DOT&E, and OSD DASD(DT&E) and make or sustain AI-specific modernization investments in the Test Resource Management Center (TRMC),[16] DAF CDAO, and Air Force Materiel Command's (AFMC's) Digital Transformation Office (DTO), and work closely with TRMC to identify AI T&E needs that will be addressed with TRMC funding and use DAF AI-specific modernization investments to address AI T&E gaps not being pursued by TRMC. These investments should include major and near-term investments in modern AI stacks across AFTC, Air Force Operational Test and Evaluation Center (AFOTEC), and the United States Air Force Warfare Center (USAFWC) (to include access to enterprise cloud-as-a-service and platform-as-a-service [PaaS] capabilities); modeling and simulation; the Virtual Test and Training Center (VTTC) at

[15] See, for example, National Academies of Sciences, Engineering, and Medicine, 2022, *Digital Strategy for the Department of the Air Force: Proceedings of a Workshop Series*, Washington, DC: The National Academies Press, https://doi.org/10.17226/26531.

[16] TRMC has several efforts under way to develop tools for testing AI. TRMC's T&E and S&T program, a 6.3 advanced technology development effort, has 10 test technology areas (TTAs). Autonomy and AI test (AAIT) is one of the TTAs. DAF T&E representatives participate in the AAIT working group (WG). AFRL and Edwards AFB have been the two USAF organizations represented in the AAIT WG.

Nellis AFB and the joint simulation environment (JSE); digital synthetic range environments at Edwards AFB and Eglin AFB; digital twins; and live-virtual-constructive (LVC) integration. The DAF AI T&E champion should work closely with the DAF's representatives on the TRMC AAIT (Autonomy and Artificial Intelligence Test) WG to identify AI T&E projects for TRMC's T&E and S&T program, while the DAF should also increase its representation on the AAIT WG.

Recommendation 3-2: The Department of the Air Force (DAF) leadership should prioritize artificial intelligence (AI) testing and evaluation (T&E) across the DAF with an emphasis on a radical shift to the continuous, rigorous technical integration required for holistic T&E of AI-enabled systems across the design, development, deployment, and sustainment life cycle.

3.4 AI T&E IN THE COMMERCIAL SECTOR

The committee was briefed by representatives from current defense industrial base companies actively developing T&E capabilities for the department,[17] the autonomous vehicle safety group at NVIDIA, and an ISO working group developing a consensus report on functional safety for AI-enabled systems. It is important to note that while commercial industry is more sophisticated than the DAF in implementing and scaling up T&E for large scale AI deployments, it is still very much a field under development.

The ISO/IEC (International Electrotechnical Commission) TR 5469 working group[18] is currently drafting a consensus report with representative members across several stakeholder industries, including avionics, robotics, healthcare, and autonomous vehicles. While in draft form, the standard is potentially subject to change from the briefed version, the TR 5469 report has the potential to provide a well-informed framework for thinking about risk, mitigation, and verification and validation (V&V) for AI-enabled systems. The main goal of the TR5469 report is "to enable the developer of safety-related systems to appropriately apply AI technologies as part of safety functions by fostering awareness of the properties, functional safety risk factors, available functional safety methods, and potential constraints of AI technologies." Many of the main points proposed in the draft align with what the committee found from the commercial sector, summarized in Table 3-1. Therefore, it is the committee's recommendation that the DAF track the progress of this report through the publication process and leverage it as a starting point for adapting their T&E processes for AI-enabled systems.

[17] For example, Morse Corporation and Calypso AI.

[18] This is a working group under the auspices of the International Organization for Standardization tasked with establishing standards on functional safety and AI systems.

Recommendation 3-3: The Department of the Air Force should track the progress of the International Organization for Standardization/ International Electrotechnical Commission TR 5469 working group report through the publication process and leverage it as a starting point for adapting their testing and evaluation processes for artificial intelligence–enabled systems.

Work presented by the industry executing on DoD T&E requirements presented a rich set of T&E tooling that has been iterated on through various pilot AI efforts and informed by engagements with non-DoD commercial customers. These techniques have begun to either be packaged in, or at least inform, various government off-the-shelf (GOTS) developer libraries being released to the broader community. When integrating these toolchains into larger systems, these developer kits codify the pilot projects' best practices for statistical analysis and Application Program Interface (API) designs. Specifically, through work on Project Maven, one developer could develop and implement T&E systems that achieved significant reductions in model evaluation time for model vendors (from months to hours). This evaluation system enables the fast iteration of model development against withheld test datasets for model comparison. While these techniques have become more sophisticated over time, they still only codify specific mathematical approaches for validating models' accuracy in isolation. To date, the committee found these contributions are very biased toward computer vision perception algorithms and have yet to extend their capabilities to fully address system-level T&E and the impact integration has on system-wide verification and validation.

Finding 3-5: DAF AI contributions to date have been focused on computer vision perception and natural language processing algorithms and have yet to extend to fully address system-level T&E.

Autonomous vehicle development was selected as a valid case study for the committee to investigate due to its similarity to some of the autonomy goals of Air Force programs. It is the modern example of AI being integrated into a safety-critical system that requires complex system-level integration. Commercial industry is increasingly investing in the technology fundamental to making autonomous vehicles a reality for consumers and participating in standards creation that govern their deployment. A presentation by a representative of NVIDIA's autonomous vehicles safety team gave an overview of their system-wide approach to managing the T&E of developed AI models within an extension of a systems engineering risk modeling framework (shown in Figure 3-4).

Within this framework, the development of the AI-enabled system begins with defining the product specification (e.g., what does the system need to do?). The product specification drives the risk model creation that, in turn, generates the

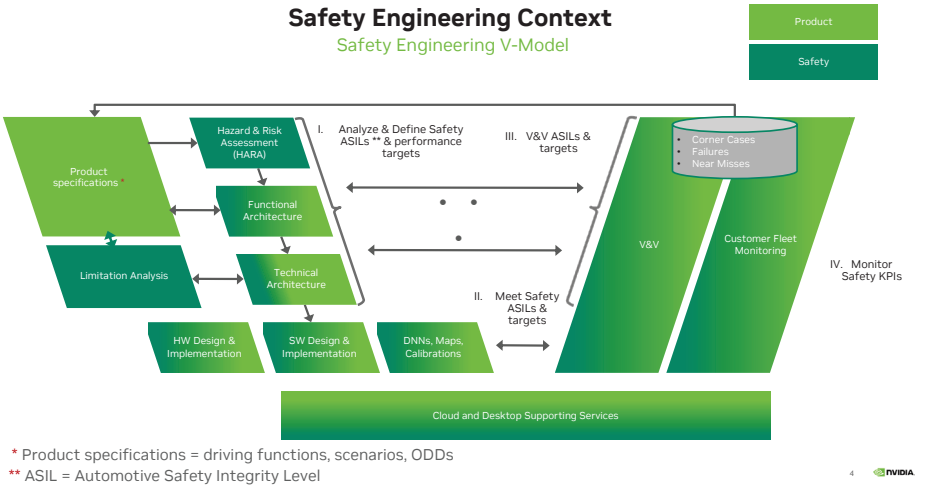

FIGURE 3-4 NVIDIA's system-wide approach to managing the T&E of AI models. SOURCE: Courtesy of NVIDIA.

functional requirements to achieve the goals of the system. The product specifications and the risk model are continuously updated through cyclical review. Two cornerstone concepts of the architecture were the assertion that AI implementations will always have a failure mode and that there are no known formal methods to demonstrate the "correctness" of AI. To manage the risks associated with these assertions, NVIDIA implements a methodology for decomposing and reducing requirements into the minimal components required to make validation and verification tenable. While the decomposition of requirements into fundamental components simplifies testing, it has a limitation with deep neural networks. Multi-model DNNs in isolation can induce common cause failures (CCFs), where multiple failures occur due to the same cause, that become impossible to capture in testing (see Figure 3-5). Because there are no known technologies to analyze the CCFs of DNNs, the failure rate of DNNs is hard to quantify, even in the presence of diverse inputs. An alternative design pattern pairs DNNs with rule-based software blocks and empowers an arbiter module to decide the best decision given the risk (see Figure 3-6). Through analysis and due diligence, a software arbiter "inherits" the argumentation so that it achieves a failure rate of 10^{-2n}.

To safely and effectively integrate AI capabilities into safety-critical system processes, continual test and refinement approaches must be implemented to manage against accepted and residual risks. Validation and verification are accomplished in both complex simulation environments and real-world test fleet deployments. Both capabilities feed refinements back into product specification. Additionally, the processes and tools that manage and implement the T&E must themselves be

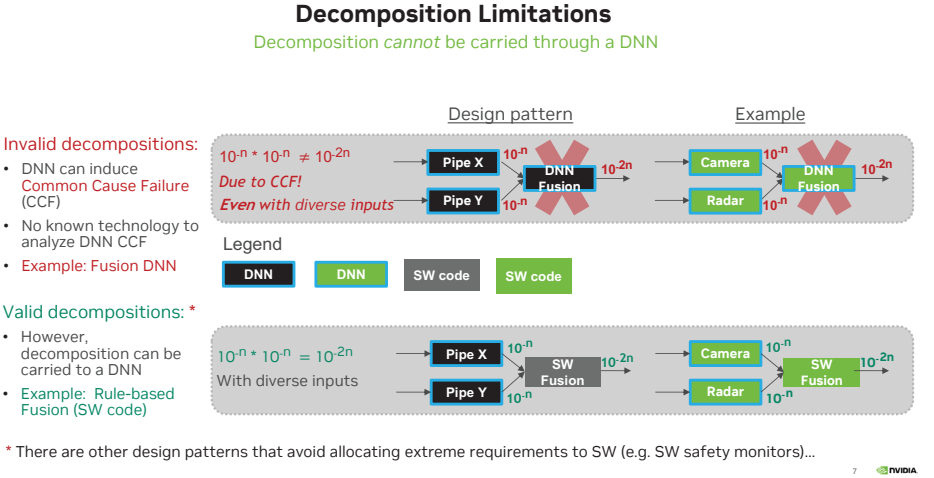

FIGURE 3-5 Common cause failures (CCFs) induced by multi-model DNNs in isolation that become impossible to capture in testing. Failure rates are represented in values 10^{-n}. Because CCF modes in a solution such as DNN fusion, the overall failure rate cannot be represented as the product of the two incoming failure rates. SOURCE: Courtesy of NVIDIA.

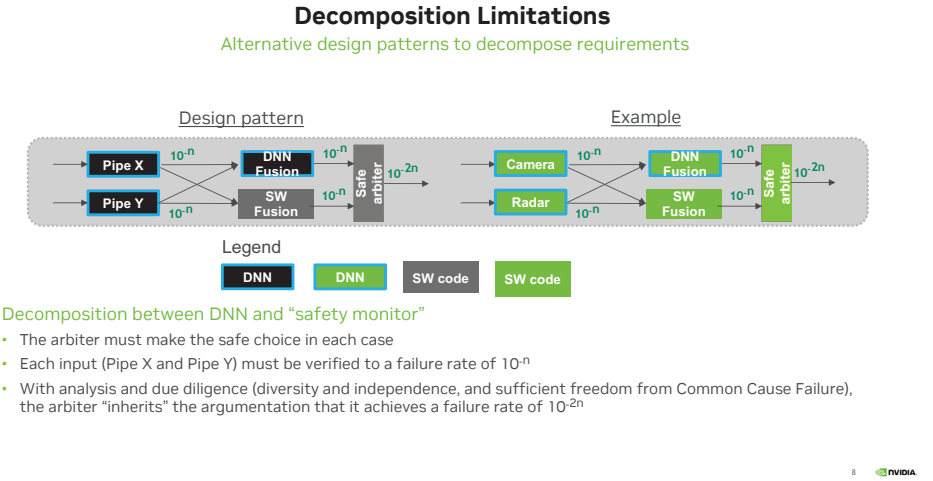

FIGURE 3-6 An alternative design pattern that pairs DNNs with rule-based software blocks. With analysis and due diligence (diversity and independence, and sufficient freedom from common cause failure), the arbiter "inherits" the argumentation that it achieves a failure rate of 10^{-2n}. SOURCE: Courtesy of NVIDIA.

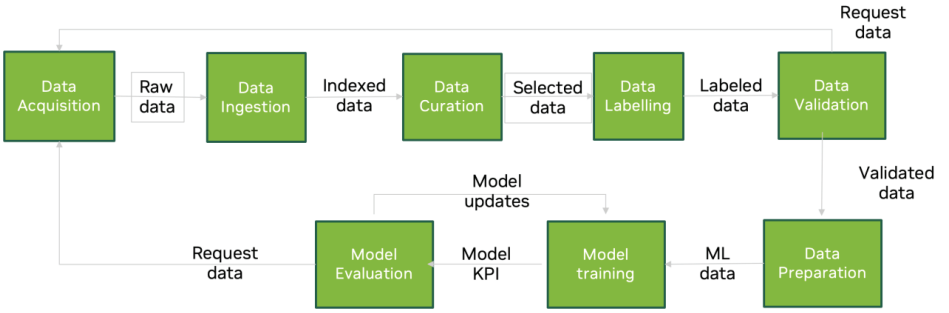

FIGURE 3-7 NVIDIA's "AI factory." SOURCE: Courtesy of NVIDIA.

secure and safe. To that end, NVIDIA has built suites of cloud-native toolchains to support the scale and latency requirements to support the iterative cycle required. Each step of the process, shown in Figure 3-7, is analyzed to identify and eliminate errors that could lead to safety-critical DNN results. Every software tool is evaluated for safety-critical bugs and user errors. NVIDIA asserted that they treat cloud-based DNN generation as a manufacturing process and view the infrastructure as an "AI factory."

It is equally important and critical to point out that the same V&V rigor applied to the creation and testing of the AI models themselves must be extended to the data used to create the models, further emphasizing the critical importance of data within the AI life cycle. Two main questions asked about all data used to train AI models are: Is the sample being considered sound? And: Is it complete? Sound data implies the sample is valid and a true member of the input space for a model. A dataset is complete when one can say that all samples that can affect safety have been identified. Demonstrating the soundness of data can be addressed with various "levels of difficulty" of a few approaches (e.g., simulation, replay of collected data, replay of augmented data, labeling of ground truth, and A/B testing). Demonstrating the soundness of data remains a significant challenge in the AI field and is currently managed through detailed limitation analysis of operational design domains (ODDs or "the scenarios") and test escapes via the test fleet or deployed fleet. Further discussion of these significant challenges can be found in Section 3.6.

As shown in Figure 3-8, there are several valid test methodologies that can be leveraged within a T&E framework for autonomous vehicles. Each methodology has its place, but also presents unique challenges. The methodologies are the following:

- *Replay of the collected data:* Sensor and meta-data collected in (large) field data campaigns are replayed as input to the system under test (SUT).

Test methodology	Challenge		
	Labeling GT	Demonstrating soundness	Demonstrating completeness
Replay of collected data	Difficult	Easy	Very difficult
Replay of augmented data	Augmentation labels easy	Moderately difficult	Very difficult
Simulation	Easy	Moderately difficult	Very difficult
Track & road testing	--	Easy	Very difficult

FIGURE 3-8 Comparison of test methodologies for autonomous vehicle systems. SOURCE: Courtesy of NVIDIA.

- *Replay of augmented data:* Collected data are augmented with 3D modeling to create input data that would otherwise be very difficult to collect on public roads.
- *Simulation:* Simulation of all input to the ego-vehicle (which contains the sensors) and closed-loop response to all output from the ego-vehicle.
- *Track and road testing:* System-level behavior testing on track or public roads.

3.5 CONTRAST OF COMMERCIAL AND DoD APPROACHES TO AI T&E

Large structural and organizational limitations within the DAF T&E ecosystem will affect the DAF's ability to meet the T&E requirements to operationalize AI implementations. A major source of these challenges has to do with the discrete differences between traditional waterfall approaches and the cyclical nature of actions required to support the AI life cycle. To highlight these limitations, it is worthwhile to walk through a conceptual example of how an AI capability would progress from development through to deployment using the current systems in place for T&E within the DAF. The point of walking through this example is to clearly point out where the current accepted T&E approach for the DAF will not support the AI life cycle. For simplification purposes, one can make several assumptions:

- All the integration requirements for a deployment platform are satisfied.
- All data collection and labeling requirements are satisfied.
- Reasonable requirements can be constructed that describe developmental and operational requirements.

- The developmental test community has the infrastructure needed to verify the delivered capability meets those requirements.
- The operational test community can iteratively test the capability and has the AI infrastructure in place to retrain as needed or can reach back to the contractor to facilitate the modifications.

With all these assumptions in place, the current process would produce a capability that needs to be handed off to the operational test community. At some point in the process, the test community will certify this new AI capability, and it will be handed over to an operational unit to employ and maintain. This part of the deployment process essentially amounts to the "operations and maintenance" (O&M) of the AI model, yet these operational units have no capacity, requirements, or infrastructure to monitor, retrain, or re-certify models as the AI life-cycle demands. Furthermore, there are no personnel in these units whose training would enable them to facilitate this type of O&M. The current processes fail to meet the AI life-cycle requirements.

The gaps become more obvious when contrasting the DAF's current approach to AI T&E against what approaches successful AI-ready commercial organizations are employing. Table 3-1 presents what the committee observed as the major differences between commercial approaches and the DAF's current approach and is not intended to be comprehensive. An AI-ready organization in this context means a group can safely, reliably, and continuously create and deploy AI-enabled systems into operational environments.

TABLE 3-1 Comparison of AI T&E Approaches Between Commercial Industry and the DAF

Commercial Approach	DOD/DAF Approach
• Significant up-front investments in data centralization, processing capability, and tooling for making data accessible, discoverable, and organized. Data are easily formed into datasets for training purposes. • Treat the creation of AI implementations as a manufacturing process. Assumes secure, scalable infrastructure to support the continuous development and test of AI components. • Employ a methodical use-case-based development of AI requirements that focuses on AI integration, not bolt-on design patterns. • Decomposes requirements based on test and evaluation requirements. • Continuous development and monitoring are supported by an operational deployment fleet and large-scale simulation environments.	• Large-scale investments in promoting data as a first-class citizen by improving accessibility and discoverability via data feeds and APIs. Lacking any rigor or tooling around dataset creation, tracking, and improvement. • AI infrastructure (compute, AIOps services, and data management services) investments are ad hoc and lack consistency. • AI requirements are functional and not developed without considering test and evaluation. • Simulation and digital twin capabilities are ad hoc and not scalable. • Development to deployment process is not well aligned with the AI life cycle.

3.6 TRUST, JUSTIFIED CONFIDENCE, AI ASSURANCE, TRUSTWORTHINESS, AND BUY-IN

Trust has been at the heart of the relationship between the operational community and the air force test community for decades. When operational units accept a new aircraft, new hardware that is integrated into an aircraft, or new or updated embedded aircraft software, they start from a position of explicit trust. It is a level of trust earned over the past 70 years by working with an air force test community characterized by its credibility, expertise, professionalism, discipline, and track record. Operational buy-in is also gained through the air force test community's well-understood standardized sequence of flight testing—DT, OT, IOT&E, DOT&E, live-fire test and evaluation (LFT&E), and follow-on testing. And when a fielded system fails for any reason, operational crews trust that the test community will identify and fix the problem before returning the system to the field.

Once the test community approves a system to be fielded, line organizations rely on testing results (to include explicit warnings and cautions about performance envelopes), academic instruction, simulators, and flights to gain confidence in the system's performance. Academic training focuses on normal performance parameters, expected critical failure modes, and how to respond to cockpit indications of degraded system performance. Even for highly complex integrated systems such as an aircraft terrain-following radar, crews adapt relatively quickly through dedicated training—academics, simulators, and flights—and trust and confidence in the original equipment manufacturer and air force test enterprise software and hardware testing processes.

The air force test community's reputation and track record have been instrumental in allowing end-users to gain and maintain deep confidence in traditional aircraft and other hardware systems. With hardware, trust is typically perceived as a binary yes-or-no concept. AI, however, is fundamentally different. Existing T&E procedures and standards do not work well for nascent and immature software capabilities, especially the black-box, self-learning, adaptive, data-centric nature of AI. Furthermore, it is hard to gain buy-in for AI-enabled capabilities when the DAF test community has not yet established the same kind of testing policies, processes, and procedures that have guided flight testing for the past 70 years. This lack of an established baseline for AI T&E makes it difficult to establish the same level of trust between the testing and operational communities that has been instrumental in fielding traditional hardware systems.

The general concept of *justified confidence* has gained traction in the AI community over the past several years. This term recognizes the challenges with using the concept of trust that has worked for other legacy hardware systems. It refers to the level of certainty or reliability achieved through direct evidence collected during design and operational test events that can be assigned to the outputs or

decisions made by AI-enabled systems. It is a term that describes how well a system can be expected to justify its decisions or predictions, considering the data used for training, the algorithms used, and any potential biases or limitations in the system. Justified confidence helps to provide evidence, transparency, and accountability in AI-enabled systems and helps establish trust in their output.

Instead of looking at trust in AI-enabled systems as a binary concept, users of AI-enabled systems will seek to gain justified confidence in that system over time. Justified confidence will also have different meanings at different levels. For example, the test community will establish internal conditions determining when an AI-enabled system can be released to the field. At the operational level, users will be less interested in tests performed in controlled or curated environments than in whether the system performs as expected under operational conditions and what could happen if the system degrades or fails. At the policy level, for higher-consequence, higher-risk systems such as AI-enabled weapons, decision-makers will seek to gain sufficient confidence in a system before approving operational deployment, measured in ways such as expected behavior, boundary conditions, potential failure modes, and possible consequences or consequence sets. Calibrating confidence in any AI-enabled system will be continuous and cumulative for end-users. It will never end. A user's confidence—and it is important to make a distinction between "trust" and functional acceptance—in a smart system will depend on context: the nature of the task, the complexity of the question to be answered, the system's previous performance record, the user's familiarity with the system, and so on, and may vary over time. In general, continued high performance in lower-risk, lower-consequence tasks will give users more confidence when facing higher-risk, higher-consequence tasks. Until users gain more experience teaming with smart machines, they will face the dilemma of placing too much or too little confidence in AI-enabled systems. Justified confidence applies any time a human and machine interact—not just when they are working together as a team.

When referring to AI-enabled systems, justified confidence is increasingly joined with the concepts of *assured systems* and *trustworthiness*. David Tate of IDA provides a framework for determining whether a system is assured and defining whether an AI system is trustworthy. He proposes that a system is assured: "when the relevant authorities have sufficient justified confidence in the trustworthiness of the system to authorize its employment in specified contexts."[19] He also defines a system to be trustworthy to the extent that "(1) when employed correctly, it will

[19] Institute for Defense Analysis (IDA), 2021, "Trust, Trustworthiness, and Assurance of AI and Autonomy," Alexandria, VA, https://apps.dtic.mil/sti/trecms/pdf/AD1150274.pdf. Tate also argues that three key features determine the level of assurance: whose trust is needed (i.e., a regulating authority); the level of confidence required (given potential benefits and risks); and the (context-dependent) level of confidence justified by the available evidence (p. 5).

dependably do well what it is intended to do; (2) when employed correctly, it will dependably not do undesirable things; (3) when paired with humans it is intended to work with, it will dependably be employed correctly."[20] The committee agrees with his assertion that "the purpose of T&E becomes clear: it is the activity that produces the evidence that completes the needed assurance arguments."[21] The committee recommends that the DAF adopt this framework as part of its AI T&E practices.

AI assurance is another term that, along with justified confidence and trustworthiness, replaces the binary concept of trust for AI-enabled systems. It refers to the process of evaluating, monitoring and ensuring the reliability, effectiveness, robustness, and safety of AI systems. AI assurance comprises a set of practices and methodologies for assessing the quality of AI models and systems, including verifying their accuracy and performance, detecting and mitigating potential biases, and evaluating their ethical and societal implications. The goal of AI assurance is to provide confidence in the decision-making processes of AI systems and to promote the responsible and trustworthy deployment of AI technologies. For DoD, AI assurance combines AI T&E and the tenets of responsible AI (RAI).[22]

RAI helps promote the safe, lawful, and ethical use of AI. AI T&E should be designed to test system performance across the RAI attributes of fairness, interpretability, reliability, and robustness. The NSCAI final report includes a detailed framework to guide the responsible development and fielding of AI implementations, which includes key considerations for policymakers and technical practitioners across the entire AI life cycle.[23] The DAF should consider using this framework and the NIST AI RMF[24] in establishing AI Assurance best practices. The committee concluded that DAF does not need to sacrifice speed to ensure adherence to the principles of RAI: it is possible to move at the speed of operational relevance while accounting for the importance of fielding AI implementations that are reliable,

[20] IDA, 2021, p. iii.

[21] IDA, 2021, p. 9.

[22] Department of Defense, 2022, *Responsible Artificial Intelligence Strategy and Implementation Pathway*, Washington, DC, https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

[23] National Security Commission on Artificial Intelligence, 2021, *The National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, p. 384.

[24] National Institute of Standards and Technology, Department of Commerce, 2023, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Washington, DC, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. See also R. Elluru, C. Howell, and M. Garris, 2023, *National Security Addition to the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework Playbook (NIST AI RMF)*, Special Competitive Studies Project, https://www.scsp.ai/wp-content/uploads/2023/04/National-Security-Addition-to-NIST-AI-RFM.docx-1.pdf.

safe, lawful, and ethical.[25] The NIST AI RMF concludes that the safe operation of AI systems is improved through[26] the following:

- Clear information to deployers on the responsible use of the system
- Responsible decision-making by deployers and end-users
- Explanations and documentation of risks based on empirical evidence of incidents

The DAF should work with OSD CDAO to adopt a definition of AI assurance. One definition to consider is "a process that is applied at all stages of the AI engineering life cycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy, and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users."[27] The committee also recommends that the DAF adopt and promulgate DoD's RAI principles and implementation plan.

Recommendation 3-4: The Department of the Air Force should adopt a definition of artificial intelligence (AI) assurance in collaboration with Office of the Secretary of Defense Chief Digital and AI Office. This definition should consider whether the system is trustworthy and appropriately explainable; ethical in the context of its deployment, with characterizable biases in context, algorithms, and datasets; and fair to its users.

Until AI is fielded widely across the DAF, the air and space force test communities gain DAF-wide agreement on AI TEVV definitions, and the test community establishes DAF-wide AI testing policies, processes, and procedures, the committee recommends that the DAF—through the AI T&E champion—codify the concepts of justified confidence, trustworthiness, and AI assurance for all AI-enabled systems. The committee expects that operational buy-in of AI-enabled systems will be neither instantaneous nor permanent. Instead, the test community and end-users

[25] See, for example, M. Ekelhof, 2022, "Responsible AI Symposium—Translating AI Ethical Principles into Practice: The U.S. DoD Approach to Responsible AI," West Point: The Lieber Institute, November 23, https://lieber.westpoint.edu/translating-ai-ethical-principles-into-practice-us-dod-approach.

[26] National Institute of Standards and Technology, Department of Commerce, 2023, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Washington, DC, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. See also R. Elluru, C. Howell, and M. Garris, 2023, *National Security Addition to the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework Playbook (NIST AI RMF)*, Special Competitive Studies Project, https://www.scsp.ai/wp-content/uploads/2023/04/National-Security-Addition-to-NIST-AI-RFM.docx-1.pdf.

[27] This definition of AI Assurance was proposed by F.A. Bararseh, L. Freeman, and C.-H. Huang, 2021, "A Survey on Artificial Intelligence," *Journal of Big Data* 8(60), https://doi.org/10.1186/s40537-021-00445-7.

will have to work closely together over the next several years in an iterative process to gain a better understanding of AI TEVV, gather more insights on AI-enabled system performance under all conditions, and establish AI testing roles, responsibilities, and authorities at all levels across the DAF, to include at the unit level.

3.7 RISK-BASED APPROACH TO AI T&E

The formulation of T&E requirements across the AI life cycle is linked inextricably to the concept of risk management. One cannot be considered in isolation from the other. In this section, the committee considers operationally-oriented risks pertaining to the integration of AI capabilities into DAF systems and the fielding decisions associated with those systems. In Chapter 5, the committee examines a broader and more detailed set of technical risks, particularly corruption and adversarial attacks, throughout the AI life cycle.

As with the T&E of all other DAF systems, risk management will play a vital role in testing AI-enabled systems. Risks are increasing as AI moves beyond specific-purpose systems to more general-purpose AI systems that are expected to become vastly more capable in different operational settings and across multiple domains.

Risks will also increase significantly as different AI-enabled systems are integrated into and begin to interact across system-of-systems architectures in complex, highly dynamic multi-domain environments and demonstrate online learning and even emergent behavior.[28] Therefore, the DAF should incorporate an AI risk management framework (RMF), such as the National Institute of Standards and Technology (NIST) AI RMF,[29] in all AI-related design, development, fielding, and sustainment. Any AI RMF includes assessing and understanding the potential risks of fielding AI-enabled systems based on different levels of dedicated T&E,

[28] See, for example, J. Harvey, 2018, "The Blessing and Curse of Emergence in Swarm Intelligence Systems," Chapter 6 in *Foundations of Trusted Autonomy: Studies in Systems, Decision and Control*, H.A. Abbas, ed., Vol. 117, https://doi.org/10.1007/978-3-319-64816-3_6. Harvey defines emergence as behavior "at the global level that was not programmed in at the individual level and cannot be readily explained based on behaviour at the individual level," p. 117.

[29] National Institute of Standards and Technology, Department of Commerce, 2023, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Washington, DC, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. "The AI RMF refers to an AI system as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments" (p. 1). The NIST AI RMF defines trustworthy AI as AI that "is valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed" (pp. 2–3). See also R. Elluru, C. Howell, and M. Garris, 2023, *National Security Addition to the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework Playbook (NIST AI RMF)*, Special Competitive Studies Project, https://www.scsp.ai/wp-content/uploads/2023/04/National-Security-Addition-to-NIST-AI-RFM.docx-1.pdf.

communicating risks to decision-makers and end-users, and determining responsibility and accountability for system failure or unanticipated performance problems.

The NIST AI Risk Management Framework (RMF) states that "AI risk management offers a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights while providing opportunities to maximize positive impacts. Furthermore, addressing, documenting, and managing AI risks and potential negative impacts can lead to more trustworthy AI systems."[30] It also notes that risk management "should be continuous, timely, and performed throughout the AI system life-cycle dimensions." Since this study is directed primarily toward AI T&E under operational conditions, the committee does not address the kinds of broad societal-level risks described in the NIST AI RMF. The committee recommends, however, that the DAF adopt the NIST's AI RMF Core, comprising the four major functions of Govern, Map, Measure, and Manage.[31]

Major risk factors commonly associated with the design and operation of AI-enabled systems are potential drop in performance due to domain shift (discussed in Section 3.2), vulnerability due to adversarial attacks (discussed in Chapter 5), perception of bias, privacy concerns, and a lack of explainability. Therefore, T&E protocols should assess the impact of each of these factors on the operational viability of AI-enabled systems and take the needed corrective measures.

Every AI capability, like every hardware system, introduces operational risks. AI shares the combination of safety and security risks with all other extant hardware and software systems.[32] Applying the NIST AI RMF categories can be a useful decomposition of some of the risks inherent in AI-enabled systems.[33] The DAF T&E enterprise has a distinguished performance record of assessing and mitigating the

[30] National Institute of Standards and Technology, Department of Commerce, 2023, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Washington, DC, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[31] The NIST AI RMF describes these major functions as follows: "Govern: A culture of risk management is cultivated and present; Map: context is recognized and risks related to context are identified; Measure: Identified risks are assessed, analyzed, or tracked; Manage: Risks are prioritized and acted upon based on a projected impact."

[32] MIL-STD-88F, "DoD Standard Practice System Safety," will replace the May 11, 2012, version (MIL-STD-88E) and will include a section on AI and ML. It will also include the AI criticality index (AICI), which will be used to determine the level of rigor (LOR) of software safety assurance activities to be imposed on the software. See Department of Defense, 2012, *Department of Defense Standard Practice: System Safety*, MIL-STD-882F, Washington, DC, https://cdn.ymaws.com/system-safety.org/resource/resmgr/documents/Draft_MIL-STD-882F.pdf. For an insightful examination of an integrated approach to safety and security, see, for example, W. Young and N.G. Leveson, 2014, "An Integrated Approach to Safety and Security Based on Systems Theory," *Communications of the ACM* 57(2):31–35.

[33] For current organizations using NIST RMF companies to help manage risk, it is important to realize that the NIST AI RMF requires a different set of expertise and would likely require a separate organization to perform the AI risk analysis.

risks inherent in hardware systems, especially flight and space weapon systems. As a form of self-learning software, however, AI presents novel sources of risk in the operational environment that are not presently well understood by the DAF test community due to a lack of familiarity with how AI systems operate, lack of operational experience with AI-enabled capabilities, and the inherent characteristics of advanced AI models. This problem will become especially acute when AI is integrated into a system-of-systems or network-of-networks architecture, leading to unknown or unanticipated cumulative and aggregate risks.

Potential risks must be considered at every stage of the AI life cycle, beginning with the formulation of AI capability requirements and associated T&E metrics and performance measures through operational fielding and sustainment via CI/CD processes.[34] AI-enabled capabilities should be fielded using a "measured risk" approach (see Section 4.3) as rapidly as operational requirements dictate while taking steps to prevent the emergence of unnecessary risks resulting from fielding capabilities that are immature, insufficiently tested, unproven, or unsafe. As one speaker argued, in some cases, the performance of an AI-enabled capability may be so compelling that leaders will have to make a risk-based decision to field even in the absence of full trust or a completely explainable system.

The committee acknowledges the challenges inherent in finding and maintaining the right balance between speed-to-field and the rigors of comprehensive T&E. As opposed to processes used for traditional hardware fielding decisions, DAF leaders should embrace the concept of "field to learn," putting capabilities in the hands of users after sufficiently rigorous "back bench" T&E by a certified AI T&E team and incorporating end-user feedback to make iterative improvements to fielded systems via accepted CI/CD processes (with the commensurate amount of T&E for all model updates).[35] Until the DAF test, program office, and operational communities gain more experience developing, testing, and fielding AI-enabled systems, the committee recommends biasing toward a more cautious—but not inherently lethargic—approach to ensure sufficient testing before any AI technology is fielded. Precaution should guide but not unduly constrain the DAF from introducing a new product or process whose ultimate effects are disputed or unknown.

One speaker noted that AI model complexity is currently doubling every 2 months. This is a staggering rate of change. Unfortunately, as presently structured, the committee expects that the DAF T&E enterprise is not capable of adapting to this rapid evolution.

[34] As one example, the Chief Architect from a commercial company briefed the committee on their use of traditional safety engineering V-models that had been adapted to reflect the entire AI life cycle, up to and including the impact of data feedback loops and CI/CD on overall system safety.

[35] S. Moore, 2023, "Right Hands, Right Place: Why We Must Push Military Technology Experimentation to the Edge," *Defense One*, January 19, https://www.defenseone.com/ideas/2023/01/right-hands-right-place-why-we-must-push-military-technology-experimentation-edge/382000.

Operational risks will increase as AI implementations (see Section 1.3) expand beyond narrow, single-task, and single-domain computer vision and natural language processing (NLP) capabilities to more advanced AI, such as reinforcement learning (RL); reinforcement learning with human feedback (RLHF); transfer learning (TL); semi-supervised, self-supervised and unsupervised learning; and foundational models and generative AI that will be vastly more capable in different operational settings and across multiple domains. Risks will also increase significantly as different AI-enabled systems are integrated into and begin to interact across system-of-systems architectures and demonstrate emergent behavior.[36] Therefore, as discussed above, the DAF should incorporate an AI risk management framework in all AI-related design, development, fielding, and sustainment; the committee recommends incorporating key elements of the NIST AI RMF, "Special Competitive Studies Project (SCSP)" in the *National Security Addition to the NIST AI RMF Playbook*,[37] and ISO/IEC standards and frameworks,[38] along with any DAF-specific additions. Any AI RMF includes assessing the potential risks of fielding AI-enabled systems based on different levels of dedicated T&E, communicating risks to decision-makers and end-users, and determining responsibility and accountability for system failure or unanticipated performance problems. Risk assessments should also address the risks presented by user unfamiliarity with AI-enabled systems (risks expected to decrease but not disappear with increasing user familiarity with such systems).

[36] See, for example, Richard Danzig's 2018 monograph, "Technology Roulette." Danzig offers a compelling caution that "Experience with nuclear weapons, aviation, and digital information systems should inform discussion about current efforts to control artificial intelligence (AI), synthetic biology, and autonomous systems. In this light, the most reasonable expectation is that the introduction of complex, opaque, novel, and interactive technologies will produce accidents, emergent effects, and sabotage. In sum, on a number of occasions and in a number of ways, the American national security establishment will lose control of what it creates" and that "twenty-first technologies are global not just in their distribution, but also in their consequences." R. Danzig, 2018, "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority," Washington, DC: Center for New American Security, https://s3.us-east-1.amazonaws.com/files.cnas.org/hero/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101&focal=none.

[37] R. Elluru, C. Howell, and M. Garris, 2023, *National Security Addition to the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework Playbook (NIST AI RMF)*, Special Competitive Studies Project, https://www.scsp.ai/wp-content/uploads/2023/04/National-Security-Addition-to-NIST-AI-RFM.docx-1.pdf?utm_source=substack&utm_medium=email.

[38] See, for example, ISO/IEC SC 42. SC 42 is a joint committee between the IEC and ISO. It serves as the focus and proponent for the ISO/IEC joint technical committee (JTC 1) international standardization program on AI and provides guidance to JTC, IEC, and ISO committees developing AI applications. Draft ISO/IEC TR 5469, "Functional Safety and AI Systems," is expected to be published in 2023. Also, see, for example, SAE AS 6983, "Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI."

For all AI-enabled capabilities, the DAF should clearly distinguish between mission- and safety-critical systems and all other AI-enabled systems. Mission- and safety-critical systems demand a much higher level of rigor and scrutiny throughout the entire T&E process, from design and development through sustainment under operational conditions. This includes an examination of reliability, repeatability, predictability, directability, safety, and security. When individual AI-enabled systems are integrated into network-centric architectures, this analysis also requires individual platform-centric assessments as well as aggregated assessments.[39] As noted earlier in this chapter, the committee heard examples from the private sector of an integrated, iterative, and comprehensive approach to AI T&E for safety-critical systems such as autonomous vehicles. These represent a good example of a complex system employed in a safety-critical operation requiring perception, decision-making, and other autonomous characteristics.

In summary, when fielding AI-enabled capabilities under operational conditions, DAF end-users, program offices, DevSecOps/AIOps teams, testers, and leaders must use a tailored AI RMF to address a series of risk-related questions at each stage of the AI life cycle.[40] These include, though are not limited to: what are the risks at each stage of the AI life cycle (including when AI systems are fielded, the potential risk to mission, and risk to force)? How are those risks determined and measured (including red teams' roles and responsibilities in assessing adversarial AI attacks against AI models)? Who assesses each risk? How are risks briefed to decision-makers at each level, and who has the authority to accept each risk or, if the risk is deemed unacceptable, to pause further development or fielding? What are the risks of catastrophic failure, either in isolation or when integrated across multiple architectures (i.e., worst-case failure modes)? How are risks managed and mitigated where necessary (including adjusting AI T&E requirements as necessary)? Finally, who is held responsible and accountable for system failure?

Recommendation 3-5: The Department of the Air Force should develop standardized artificial intelligence (AI) testing and evaluation protocols to assess the impact of major AI-related risk factors.

[39] Aggregated risk assessments of complex network-centric architectures should be completed by a multidisciplinary group that has broader visibility into all the components of the network and system-of-systems architecture.

[40] See, for example, Appendixes A–C for a proposed comprehensive AI RMF for the U.S. Intelligence Community, In C.R. Stone, 2021, "The Integration of Artificial Intelligence in the Intelligence Community: Necessary Steps to Scale Efforts and Speed Progress," *Joint PIJIP/TLS Research Paper Series*, 73, https://digitalcommons.wcl.american.edu/research/73.

4

Evolution of Test and Evaluation in Future AI-Based DAF Systems

4.1 INTRODUCTION

When the committee first set out to answer the questions driving this report, there was a healthy discussion about the study's scope. At first, the questions asked appeared constrained, and the boundaries for areas of investigation seemed clear. However, after investigating each question, it became obvious to the committee that these questions could not be viewed in isolation. The areas being explored were as entangled with the complexity of the Department of the Air Force (DAF) bureaucracy as they were with the complexity of the technology. A common refrain in several data-gathering sessions was that the "DAF has a tiger by the tail"—a euphemism for the unexpected and unintended consequences that come with bold moves. These unexpected and unintended consequences are not necessarily unwanted or unneeded, but potentially more impactful than the DAF has anticipated. The evolution required to effectively operationalize artificial intelligence (AI) will affect a significantly larger part of the DAF than seems obvious at first glance, as the committee expects AI to be embedded throughout the entire DAF over the next decade. This chapter discusses the actual scope of the impact of these advancements—not only on the test community but also on the requirements processes and DAF culture. The chapter also reviews trends in AI technology that illustrate how quickly the field is changing, and hence how important it will be to maintain a firm yet flexible grip on this tiger's tail as AI-based systems emerge ever more rapidly across the DAF.

4.2 APPOINTING A DAF AI T&E CHAMPION

The magnitude of change this report suggests will require dedicated leadership, continuous oversight, and individual responsibility and accountability. This is best accomplished by formally designating a senior AI test and evaluation (T&E) official who reports to the Secretary of the Air Force, is responsive to the Chiefs of the Air and Space Forces, and who has the necessary resources and authorities to implement DAF-wide changes.

The 2022 dual-hat designation of the 96th Operations Commander as the chief of AI test and operations for the DAF Chief Data and AI Office (CDAO) is a positive and important step, and the report committee views the 96 OG/CC as one of the primary beneficiaries of this report. However, as currently constituted, the chief of AI test and operations for the DAF CDAO does not have the authority to make the magnitude of changes across the DAF this committee believes necessary to enable AI T&E.

> Finding 4-1: Currently, no single person below the level of the Secretary or the Chiefs of the Air and Space Forces has the requisite authority to implement DAF-wide changes to successfully test and evaluate AI-enabled systems.

For this reason, the committee recommends that the Secretary of the Air Force formally designate an overall DAF AI T&E champion at the general officer or senior executive service level in the DAF, and delegate to them the necessary authorities to make changes on behalf of the Secretary and Service Chiefs. This advocate should have breadth and depth of experience in both AI and T&E, to include extensive experience with human-systems integration and agile software T&E. This advocate should establish an AI governance structure that includes formally delineating AI T&E reporting relationships and roles and responsibilities across the cri-Center, the future U.S. Space Force Operational Test Agency (OTA), the DAF CDAO, and operational air, intelligence, C2, space, and cyber units.[1] This process should include assessing what broader DAF-wide organizational and governance changes are needed to reflect the differences between AI T&E and T&E for all other Air Force systems and capabilities.

The AI T&E champion should be charged with implementing the DAF AI T&E vision, granted the requisite authorities and resources (to include personnel) and fully empowered to help realize that vision for the DAF. The DAF AI T&E champion should focus on new test designs for AI-enabled systems that incorporate the core systems engineering principles of non-AI-enabled systems while adding new

[1] Because of the unique T&E expertise required, the committee does not propose dual-hatting the DAF CDAO as the DAF AI T&E champion. Given the centrality of data to AI testing, however, the offices of the AI T&E champion and CDAO will be inextricably linked.

elements that reflect the best AI T&E practices from academia, the private sector, and other government test organizations.

Recommendation 4-1: The Secretary of the Air Force and chiefs of the Air and Space Forces should formally designate a general officer or senior civilian executive as the Department of the Air Force (DAF) artificial intelligence (AI) testing and evaluation (T&E) champion to address the unique challenges of T&E of AI systems identified above. This AI T&E advocate should have the requisite AI and T&E credentials, and should be granted the requisite authorities, and responsibilities, and resources to ensure that AI T&E is integrated from program inception and appropriately funded, realizing the DAF AI T&E vision.

A successful model for appointing and empowering the AI T&E champion can be found with the response of the DAF to a previous National Academies study. In 2015, a study on the role of experimentation in the Air Force innovation life cycle[2] recommended as its highest priority that a single individual at the top of the organization be responsible for "catalyzing" their desired outcome. That report emphasized the need for a singular authority responsible for "owning" the problem—and articulated that successful innovative organizations ensured that a "clearly identified individual was assigned responsibility for leading this work, was evaluated on their success in doing so, and woke up every workday focused on how to get it done better." The DAF adopted this recommendation with great success.

General Mark Welsh, the then-Air Force Chief of Staff (CSAF), designated General Ellen Pawlikowski, then AFMC Commander, to spearhead the innovation and experimentation effort. Gen Pawlikowski instituted the strategic development planning and experimentation group (SDPE) to execute this responsibility. This group reported directly to Gen. Pawlikowski, and a new capability development council (CDC) reported to Gen. Welsh. Significantly, both of these institutions were chartered by the CSAF before the conclusion of the experimentation study. Notably, the SDPE continues to stimulate innovation across the Air Force—the next generation air dominance (NGAD) group is a salient example. Gen. Duke Richardson (the current AFMC commander) also recently established a digital transformation office (DTO) within AFMC; the AFMC commander used this similar approach to rectify the shortfall in implementing an effective digital strategy in the Air Force.

This model is just one successful demonstration that the DAF has of identifying and empowering a champion who is able to effectively implement the necessary changes.

[2] National Academies of Sciences, Engineering, and Medicine, 2016, *The Role of Experimentation Campaigns in the Air Force Innovation Life Cycle*, Washington, DC: The National Academies Press, https://doi.org/10.17226/23676.

4.3 ESTABLISHING AI T&E REQUIREMENTS

Throughout this study, one of the constant refrains this committee heard from speakers was the importance of formulating T&E requirements for AI capabilities that reflected the needs of end-users or operators, not only developers or testers.[3] Yet the same speakers acknowledged the difficulty of defining comprehensive T&E requirements for software-centric capabilities whose "black box" performance under operational conditions could change continually based on the ingestion of more and more data and that generate probabilistic or statistically predictable behavior rather than deterministic results, and whose performance could change significantly with every update to a fielded model.

Most current AI models do not learn by themselves in the field. They are trained and tested a priori and deployed. They may be re-trained under operational conditions in an operational environment, in which case regression testing is required. Most AI models are per se deterministic in that, for example, a neural network has weights and thresholds and a method for combining the operations that is deterministic (i.e., the model is based on mathematical functions that operate in a predictable way). However, the data they ingest under operational conditions is stochastic, subject to environmental noise, sensor noise, data dropouts, faulty equipment and data collection, and environmental conditions. This "probabilistic" behavior is intrinsic to all sensing systems. What is unique to many AI models is that their behavior under these data corruption and stochastic behavior scenarios are not well understood at the theoretical level and often exhibits what is today seen as non-intuitive and brittle failure modes. At the same time, while overall model performance is expected to improve over time as more operational data are ingested, absorption of more data could also lead to significant reductions in performance if the new data are corrupted or poisoned or the AI model is subject to other forms of adversarial attacks (see Chapter 5). This would be particularly problematic if such attacks are undetected.

The intersection of these two equally important considerations sets AI T&E apart from all previous DAF T&E. It leads to a fundamental and persistent challenge for AI T&E today: understanding what requirements to test against when evaluating standalone AI models, and what requirements to test against once one or more AI capabilities are integrated into a DAF system. As the NSCAI noted and as Project Maven demonstrated, the former is challenging enough; the latter introduces formidable new complexities that will require entirely new approaches to performing T&E of AI-enabled weapon systems or decision support systems—not only for AI

[3] One speaker noted that it was essential for AI developers to talk to operators or end-users at the beginning of a system's design phase. This would not only allow developers to gain better insights into how a given AI-enabled capability would be used operationally, it would also help end-users gain a better understanding of the AI T&E process. The committee returns to this point later in this section.

added to fielded systems, but also for AI that is baked-into new systems beginning with the design phase.[4] With the current state of technology, AI T&E does not align conveniently with either T&E for traditional hardware weapon systems or T&E associated with DoD's software acquisition pathway (although, generally, it is a closer fit to the latter than the former).[5]

This dilemma is a manifestation of the major differences between AI T&E and traditional T&E of hardware systems, which assesses and evaluates well-defined

[4] In its recommendations for AI T&E future actions, the national security commission on AI final report notes that "Progress on a common understanding of TEVV concepts and requirements is critical for progress in widely used metrics for performance. Significant work is needed to establish what appropriate metrics should be used to assess system performance across attributes for responsible AI according to applications/context profiles. (Such attributes, for example, include fairness, interpretability, reliability, and robustness.) Future work is needed to develop: (1) definitions, taxonomy, and metrics needed to enable agencies to better assess AI performance and vulnerabilities; and (2) metrics and benchmarks to assess reliability and intelligibility of produced model explanations. In the near term, guidance is needed on: (1) standards for testing intentional and unintentional failure modes; (2) exemplar datasets for benchmarking and evaluation, including robustness testing and red teaming; and (3) defining characteristics of AI data quality and training environment fidelity (to support adequate performance and governance)," p. 645.

The committee encourages the DAF to adopt these recommendations. See National Security Commission on Artificial Intelligence (NSCAI), 2021, *National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, p. 137.

[5] DoD Instruction 5000.89 describes DoD-wide test and evaluation policies, processes, and procedures for urgent capability acquisition, middle tier of acquisition (MTA), major capability acquisition, software acquisition, and defense business systems (DBS). See U.S. Office of the Under Secretary of Defense for Research and Engineering, 2020, "DoD Instruction 5000.89: Test and Evaluation," Washington, DC: Department of Defense, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500089p.pdf. Defense acquisition of services does not require T&E policy and procedures. DoDI 5000.89 states that "For non-major defense acquisition programs (MDAPs) and for programs not on T&E oversight, these guiding principles should be used as a best practice for an integrated and effective T&E strategy," p. 4. AI T&E is not discussed in DoDD 5000.89; accordingly, as currently written this directive provides "guiding principles" for AI T&E, not definitive guidance. See U.S. Office of the Under Secretary of Defense for Acquisition and Sustainment, 2020, "DoD Instruction 5000.02: Operation of the Adaptive Acquisition Framework," Washington, DC: Department of Defense, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf. This Instruction addresses the use of the adaptive acquisition framework (AAF) in software acquisition. DoDI 5000.02 states explicitly that "Programs executing the software acquisition pathway are not subject to the Joint Capabilities Integration and Development System (JCIDS), and will be handled as specifically provided for by the Vice Chairman of the Joint Chiefs of Staff, in consultation with Under Secretary of Defense for Acquisition and Sustainment (USD(A&S)) and each service acquisition executive," p. 3. It also notes that "Programs executing the software acquisition pathway will not be treated as major defense acquisition programs," p. 3. See U.S. Office of the Under Secretary of Defense for Acquisition and Sustainment, 2020, "DoD Instruction 5000.87: Operation of the Software Acquisition Pathway," Washington, DC: Department of Defense, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500087p.pdf.

key performance parameters (KPPs) or, for information systems, net-ready KPPs (NR-KPP), and other largely static metrics[6] driven by the joint capabilities integration and development system (JCIDS)[7] and joint requirements oversight council (JROC) and established during system design and development. One of the dilemmas the AI test community must grapple with is to understand when more traditional KPPs or NR-KPPs should apply, and when more flexibility is required to avoid placing undue constraints on AI systems that are designed to meet end-user needs under operationally-relevant timelines. In other words, for AI capabilities the sponsoring organization must find the appropriate balance between overly broad and unnecessarily restrictive performance specifications, as the committee discusses in more detail below.

In general, AI requires greater integration between designers, testers, and operators or end-users to enable transparency of approach and outcome, as is common in the application of a DevSecOps process (see Recommendation 4-2). The differences between the two approaches need to be acknowledged in the near term. Still, all short-term solutions will continue to evolve over time through an iterative, interactive process as air force end-users and personnel within responsible test organizations gain more experience with writing AI-centric T&E requirements and with AI T&E processes and practices and as AI T&E becomes more automated and test results become more explainable. The committee echoes the NSCAI's recommendation to the military services to "establish a TEVV framework and culture that integrates testing as a continuous part of requirements specification, development, deployment, training, and maintenance and includes run-time monitoring of operational behavior."[8] Section 255 of the FY2020 National Defense Authorization Act (NDAA) established a "shift left" for software that requires T&E be incorporated into the development life cycle of the software, at minimum. This policy would naturally extend to AI T&E, which will then need to go further to include the continuous T&E necessary for AI.

Conclusion 4-1: Compared to traditional T&E, AI T&E requires radically deeper continuous technical integration among designers, testers, and operators or end-users.

[6] Such as critical technical parameters (CTP), critical intelligence parameters (CIP), key system attributes (KSA), interoperability requirements, and cybersecurity requirements. KPPs/NR-KPPs will still exist for AI-enabled systems, particularly in areas such as the safety and security of AI-enabled safety-critical systems.

[7] For traditional hardware systems, the sponsoring service or agency enters the JCIDS process with a capabilities-based analysis (CBA), Doctrine, Organization, Training, materiel, Leadership, Personnel, Facilities, Policy (DOTmLPF-P) analysis, other studies or analyses, or transition of rapidly fielded capability solutions.

[8] NSCAI, 2021, *National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, p. 384.

Treating requirements for AI capabilities in the same manner as those for traditional hardware systems is likely to lead to unnecessary delays in development, acquisition, fielding, and sustainment.[9] As AI is a software capability, it is essential for developers to be as flexible and agile as possible to allow fielding models and model updates on operationally-relevant timelines.[10] Rather than applying the extreme rigor of and adhering to the extended timelines associated with the JCIDS requirements process, the preferred approach for AI-enabled capabilities is to link proposed solutions—whether provided by commercial vendors or DoD organizations—to existing JCIDS requirements while being sure to follow a DevSecOps or AIOps/MLOps development methodology. This will shorten development and fielding timelines considerably.[11] One of AI's most distinguishing features is the importance of relying on real and near real time feedback from operational users, and ingesting operational data, to make rapid iterative improvements in fielded AI models via the agile methodology and CI/CD processes.

Recommendation 4-2: The Department of the Air Force should adopt a more flexible approach for acquiring artificial intelligence (AI)-enabled capabilities that whenever possible links proposed solutions to existing joint capabilities integration and development system requirements, and that follows a development, security, and operations or AI for information technology operations/machine learning operations development methodology.

The DoD Algorithmic Warfare Cross-Functional Team (Project Maven) used this approach when soliciting computer vision solutions to meet standing operational needs: members of the Maven team performed an exhaustive search of JCIDS databases to find existing requirements that had identified operational limitations and requested solutions that could augment, accelerate, and automate processing, exploitation, and dissemination of tactical and medium-altitude UAS full-motion video. Once a commercial CV algorithm solution could be linked to an existing

[9] Another risk that has not been sufficiently considered when "testing to requirements" in accordance with the JCIDS process, is that AI systems that return better-than-expected testing results could be discarded for not meeting specific narrowly defined JCIDS-dictated requirements.

[10] See for example, W. McHenry and M. Brown, 2022, "The 1960s Had Their Day: Changing DoD's Acquisition Processes and Structures," *Real Clear Defense*, December 5, https://www.realcleardefense.com/articles/2022/12/05/the_1960s_had_their_day_changing_dods_acquisition_processes_and_structures_868279.html. The authors emphasize the difference between DoD's linear acquisition processes and successful commercial technology programs that rely on cross-functional teams and continual user feedback during design, development, fielding, and sustainment.

[11] DoDI 5000.89 requires a test strategy when using the software acquisition pathway, and notes that this pathway "focuses on modern iterative software development techniques such as agile, lean, and development security operations, which promise faster delivery of working code to the user. The goal of this software acquisition pathway is to achieve continuous integration and continuous delivery to the maximum extent possible" (p. 24).

formal DoD requirement and translated into a request for proposal (RFP), the Maven T&E team established testable and verifiable performance measures for that algorithm, as described previously in this report.

Members of Project Maven "translated" esoteric T&E metrics into terms that were most relevant to operational end-users. Because formal requirements had not been established for AI model performance, once the Maven team had completed data quality assurance, T&E on each model, integration testing in the Maven Integration Lab, and live-fly testing, user acceptance of each trained model and follow-on updates to those fielded models, was based primarily on an agreement between the Maven team and operational users that the models had demonstrated adequate performance under operational conditions (as compared to the baseline performance achieved with existing, non-AI systems). Once a minimum viable product (MVP) model was fielded, user feedback was instrumental in refining model performance through continuous integration and continuous delivery (CI/CD). This entire process, which was considerably less rigid than the T&E of major acquisition program hardware systems, underscored the importance of defining future T&E requirements for all AI capabilities and AI-enabled platforms, sensors, and tools in ways that reflect consensus between developers and end-users at every stage of the AI life cycle. The JAIC T&E division (now under the OSD CDAO) refined Maven's processes, procedures, and practices and is publishing CDAO AI T&E playbooks and providing AI T&E frameworks to OSD DOT&E that the DAF should consider adopting.[12]

This less constrained approach to AI requirements formulation introduces risks. It creates the potential for overly broad performance specifications and disparities between contract language and end-user requirements. However, such risks can be mitigated substantially through a continuous dialogue between developers (DevSecOps or AIOps/MLOps teams), end-users, designated acquisition officials, and the responsible DAF test organization. Such a dialogue will help developers and testers formulate T&E metrics and performance measures that best match the end-users' operational needs. While end-user involvement and feedback are valuable during the T&E of all systems, it is especially important during every stage of the AI life cycle due to general unfamiliarity with AI capabilities, as well as AI's unique self-learning characteristics compared to all other traditional DAF hardware systems and software capabilities.

The National Institute of Standards and Technology's (NIST's) AI Risk Management Framework (RMF) lists representative AI actors across the AI life cycle.[13]

[12] These include frameworks for T&E of AI-enabled systems (AEIS); operational testing of AEIS; human-system integration (HSI); system integration; responsible AI (RAI); and AI assurance.

[13] National Institute of Standards and Technology, Department of Commerce, 2023, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Washington, DC, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. Also, see the accompanying (draft) NIST AI RMF Playbook, available at https://pages.nist.gov/AIRMF.

This list of activities and representative AI actors in each stage of the AI pipeline underscores the role that operators or end-users should play in the AI life cycle, most importantly beginning with plan and design. Which, in coordination with testers, domain experts, AI designers, product managers, and others, is intended to lead to the formulation of AI T&E metrics and performance measures. The level and frequency of end-user or operator involvement in this process is another feature that distinguishes AI T&E from most traditional DAF hardware testing practices.

Addressing this foundational "how much?" question should be one of the DAF AI T&E champion's initial top priorities, guided by discussions with the OSD CDAO, DOT&E, DASD(DT&E), AF CDAO, AFMC DTO, and other relevant DAF and joint AI test organizations and agencies. The answer to this question will always be context dependent, reflecting a combination of myriad factors such as end-user requirements, degree of urgency, technology and human readiness levels (TRLs/HRLs), assessed risks of action and inaction, scope, scale, and differences between an original fielded model and subsequent model version updates. It will also depend on the level of risks that end-users are willing to accept based on their operational imperatives. Yet this requires the test-responsible organization to communicate as transparently as possible to end-users measured and expected performance capabilities, system limitations, and possible failure modes of AI-enabled systems that users intend to accept for fielding.[14]

As the NSCAI recommended in its final report, one of the DAF's critical first steps, led by the AI T&E champion in coordination with the OSD CDAO, OSD DOT&E, DASD(DT&E), and DAF CDAO, should be to establish "a process for writing testable and verifiable AI requirement specifications that characterize realistic operational performance," and to provide "testing methodologies and metrics that enable evaluation of these requirements—including principles of ethical and responsible AI, trustworthiness, robustness, and adversarial resilience."[15]

As noted above, the iterative and interactive dialogue between end-users, testers, and the broader AI community will help operators and testers agree on request for proposal/request for information (RFP/RFI) and contract language, help end-users understand how AI performance will be assessed by testers, and help testers develop appropriate test metrics and performance measures. As noted in the Project Maven case study, other AI T&E best practices include setting aside sufficient representative data for training, validation, or assessment, and test; building T&E

[14] See, for example, M.A. Flournoy, A. Haines, and G. Chefitz, 2020, *Building Trust Through Testing: Adapting DOE's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems*, Washington, DC: Center for Security and Emerging Technology (CSET). https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

[15] NSCAI, 2021, *National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, p. 384.

harnesses; evaluating fielded models as part of ongoing operational assessments; defining model boundary conditions and assessing AI failure modes; and developing T&E processes for each subsequent update to fielded models through normal CI/CD processes. One of AI's most distinguishing features is the importance of relying on real- and near-real-time feedback from operational users and ingesting operational data to make rapid iterative improvements in fielded AI models via agile principles and CI/CD processes. This includes integrating test personnel with operational units when feasible. The DAF should consider training a subset of the DAF-wide test cadre to be integrated into operational units to assist with onsite AI T&E.

When the committee refers to requirements, it includes the need for DAF-wide investments in enabling capabilities that support AI T&E at enterprise scale. Because of AI's unique characteristics and the uncertainties associated with AI performance in operational environments, the committee recommends that the DAF prioritize investments for digital modernization of the DAF test enterprise and for implementing an enterprise-level T&E architecture, enabled to the maximum extent possible by the OSD CDAO,[16] OSD DOT&E, OSD DASD(DT&E), the test resource management center (TRMC), and DAF CDAO. This should include major and near-term investments in modern AI stacks across AFTC, AFOTEC, and USAFWC (to include access to enterprise cloud-as-a-service and platform-as-a-service [PaaS] capabilities); modeling and simulation; the Virtual Test and Training Center (VTTC) at Nellis AFB; the joint simulation environment (JSE);[17] the Air Force Digital Test Environment; the 96th Operations Group's new initiative to establish

[16] Through its National AI T&E Infrastructure Capability (NAITIC) study, OSD CDAO is coordinating with TRMC, DTE&A, and DOT&E to answer the following basic question: is DoD properly resourced to adequately test and evaluate AI-enabled capabilities? This study is designed to systematically explore supply and demand for T&E of AI capabilities and identify gaps in DoD infrastructure. The study's primary conclusion is that *there is no evidence-based analysis of DoD AI T&E infrastructure gaps tied to demand (programs with AI capabilities) or supply (extant T&E infrastructure)*. In the near term, the DAF can take advantage of CDAO's test harnesses (available through the CDAO "test and evaluation factory"), T&E bulk purchase agreements (BPA), and the red teaming handbook. The CDAO's joint AI test infrastructure capability (JATIC) is an interoperable set of state-of-the-art software for rigorous AI model and algorithm test and evaluation. The CDAO AI Assurance division also makes available actual test products such as test and evaluation master plans (TEMP), to include one for an autonomous system; red team assessments; algorithmic integrity assessments; and human-system integration type assessments.

[17] The joint simulation environment or JSE is a scalable, expandable, high-fidelity, government-owned non-proprietary modeling and simulation environment. While designed originally for testing fifth-generation aircraft in a simulation environment, its use is expanding to fulfill other integrated testing requirements.

a digital synthetic version of the air and ground ranges in and around Eglin AFB; digital twins;[18] and live-virtual-constructive (LVC) integration.

As more AI-enabled weapon systems, especially AI-enabled autonomous weapon systems, are fielded across the Air Force and Space Force, there will be tremendous value in providing dedicated T&E "sandbox environments." Such environments will be vital in supporting T&E for systems in more operationally-realistic settings and in providing more insights into potential AI system limitations and failure modes while also allowing the appropriate assessment of individual system risks and the risks associated with integration into a system-of-systems.

Digital modernization includes building and sustaining data management pipelines (DMP) for all AI projects. Every DAF AI project requires building a project-specific information architecture and establishing processes and procedures to generate training-quality data (TQD) essential to building and testing high-performance AI models. Additionally, the DAF AI T&E champion, in coordination with DAF system program offices (SPOs) and PEOs, should provide standardized contract options to address the need for TQD and machine-readable data, along with options for intellectual property (IP) protections and ownership of data rights and licenses for both commercial vendors and government entities.[19] Finally, one of the primary duties of the DAF AI T&E champion would be to formally adopt and promulgate DAF-wide guidance, such as the February 2022 DAF-MIT AI Accelerator *Artificial Intelligence Acquisition Guidebook*.[20]

As noted previously, the committee expects DAF leaders to substantially underestimate the level of investments required to implement digital modernization of

[18] While there are many definitions, a digital twin is generally defined as a digital representation of a physical object or system that can be used to simulate its real-world behavior and characteristics. Cortez et al. (2022) define a digital twin as a "digital representation of a Single Board Computer (SBC) and/or components representing a functionally correct, predictable and reproducible representation of that board or system at the appropriate level of fidelity to perform software verification, performance analysis and software validation tasks." N.F. Cortez, E. Williams, A. House, and J. Ramirez, 2022, "Virtualization: Unlocking Software Modularity of Embedded Systems," 2022 DoD Weapon Systems Software Summit, Orlando, FL: Orange County Convention Center, December 13, https://repo1.dso.mil/dsawg-devsecops/team-8/team8_artifacts/-/blob/master/Virtualization_-_Unlocking_Software_Modularity_of_Embedded_Systems_v2.pdf.

[19] See, for example, A. Bowne and R. Holte, 2022, "Acquiring Machine-Readable Data for an AI-Ready Department of the Air Force," *The JAG Reporter*, November 29, https://www.jagreporter.af.mil/Post/Article-View-Post/Article/3216144/acquiring-machine-readable-data. In addition to describing the importance of TQD and machine-readable data, the authors also address IP and data rights as part of the contracting and acquisition process for AI projects. See also Department of Defense, 2020, *DoD Data Strategy*, September 20, Washington, DC, https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DoD-Data-Strategy.pdf.

[20] Department of the Air Force, 2022, *Artificial Intelligence Acquisition Guidebook*, Cambridge, MA: MIT, https://aia.mit.edu/wp-content/uploads/2022/02/AI-Acquisition-Guidebook_CAO-14-Feb-2022.pdf.

the DAF test enterprise and establish modern AI data management best practices. Therefore, in coordination with OSD CDAO, the committee recommends that the DAF immediately initiate a comprehensive analysis of the resources required to carry out digital modernization across the Air and Space Forces and resource those requirements appropriately in future DAF budgets.

Over the longer term, when feasible and it makes sense operationally, the DAF should strive to integrate AI into programs of record, via the DAF's SPOs and PMOs, and program executive officers (PEO), rather than "bolting on" AI to a system after fielding, as is the case today.[21] In these cases, AI T&E can be integrated into the host weapon system test and evaluation master plan (TEMP). However, DAF responsible test organizations should be wary of allowing AI T&E to be "held hostage" when there are excessive delays in the parent weapon system test schedule during DT, OT, IOT&E, FOT&E, or live-fly test and evaluation (LFT&E). One speaker provided an example of a delay in flight testing that caused an undue delay in the planned rapid T&E of an AI capability integrated into the system under test. Another speaker cited an example of overly-restrictive conditions directed by a program of record owner on the ability to update an integrated AI capability hosted on a hardware platform. As a result, large portions of T&E can be accomplished on AI capabilities before they need to be tested as part of a fully hardware-software integrated weapon system. This departure from established hardware test practices suggests the need for a DAF-wide test enterprise cultural shift, which in turn depends on providing more education and training on AI T&E and agile principles. The committee addresses this in more detail in the following Culture Change and Workforce Development section.

Recommendation 4-3: To the maximum extent possible and where it makes sense operationally, the Department of the Air Force (DAF) should integrate artificial intelligence (AI) requirements into programs of record, via the DAF's system program offices and program executive officers, and integrate AI testing and evaluation (T&E) into the host weapon system T&E master plan.

Even with the rapid development of new AI capabilities and the maturation of earlier AI-enabled systems that provide opportunities for rapid updates to fielded models, many AI systems today remain brittle. Apart from the difficulties of fielding AI-enabled capabilities that perform as well in the operational environment as they do on the laboratory bench, AI will be subject to corruption and adversarial attacks in the form of model or algorithm denial and deception, data poisoning, evasion attacks, and cyberattacks among others. Adversarial attacks will occur at

[21] One notable exception is the Air Force Ground-Based Strategic Deterrence "Sentinel" program, which has incorporated digital modernization principles, to include the use of digital twins, since program inception.

the model level, system level, during operational deployment, and throughout the entire AI life cycle and data management pipeline. As such, the DAF must establish dedicated independent AI red teams that are considered as fully integrated elements of AI T&E. These teams can help develop and update defenses against adversarial attacks while also supporting the development of offensive adversarial attack techniques—much in the same way that "red air" has played an indispensable role in improving the effectiveness of DAF over the past 40 years, or in how cyber "white hat" and red teams have been honing the skills of the DAF cyber defenders and attackers over the past decade.

Red teams represent a critical component of AI test design and the overarching requirements process. These teams must be capable of emulating current and future peer competitor capabilities and performance and should be integrated into the entire AI life cycle. Furthermore, the committee underscores the importance of not viewing AI red teams as entities that are completely separate from the AI T&E enterprise.[22] Instead, they should be integral to AI T&E, *although independent*, and focused on *operational performance* and *mission resilience* in the face of known and unknown—but expected—adversarial attacks, beginning with the presumption of attack at every stage of the AI life cycle, including cyberattacks, data manipulation, and data corruption and poisoning (as discussed in the next chapter). This includes the importance of instrumenting fielded systems to inform end-users of a potential adversarial attack or unexpected degradation in model performance (which may indicate an adversarial attack). Similar to OSD DOT&E's use of cyber red teams, the committee recommends that DAF AI red teams fall under the direction of the DAF AI champion (or designated AI T&E lead).

Establishing a DAF activity focused on AI-based systems red-teaming would provide trust and justified confidence in the face of potential adversarial attacks that present unique challenges for which the DAF is currently unprepared. To ensure that AI-enabled systems are resilient during development, training, deployment, and retraining or updating, the committee recommends the DAF develop T&E approaches integrated with red-team findings that reflect the range of adversarial activity anticipated during all phases of the AI life cycle.

Recommendation 4-4: The Department of the Air Force should establish an activity focused on robust artificial intelligence–based systems red-teaming, implement testing against threats the red-teaming uncovers, and coordinate its investments to explicitly address the findings of red-team activities and to augment research in the private sector.

[22] OSD DOT&E has relied on DOT&E-sponsored and service-led cyber red teams for the past several years. See for example, DOT&E, 2022, "Cyber Assessment Program," FY 2021 Annual Report, https://www.dote.osd.mil/Portals/97/pub/reports/FY2021/other/2021cap.pdf?ver=597qqovFSFg_PajZvaLu_w%3D%3D.

Finally, the DAF AI T&E champion must address how to respond to requests for changes to fielded AI models beyond the process described earlier, which accounts for regular, periodic updates through CI/CD processes. For all designated DoD-wide weapon systems, existing urgent operational needs (UON)/joint urgent operational needs (JUON)/joint emergent operational needs (JEON) processes are used for capability requirements identified as impacting ongoing or anticipated contingency operations. For example, for air force aircraft and related systems, the normal peacetime change process begins with unit-level requests (such as an operational change request [OCR], or Form 1067, which is used to document the submission, review, and approval of requirements for modifications). For fielded electronic warfare (EW) systems, units can seek emergency reprogramming updates through the EW integrated reprogramming (EWIR) process. For cyber systems, the AFCYBER incident response plan can trigger requirements for changes to fielded capabilities. DAF leaders should consider the advantages and limitations of all these different processes—as well as those in other DoD organizations and private sector companies—when establishing new processes and procedures that govern requests for urgent updates to fielded AI models. These processes and procedures must account for data requirements, model retraining, and the extent of additional T&E required.[23]

Data Management Requirements

Despite the focus on digital transformation and data over the past 5 years, the DAF is not yet an AI-ready force. The DAF does not yet treat its huge capacity for data collection in its internal business operations and its external missions in ways optimized for AI-based processing and exploitation. With few exceptions, data are not treated as a "first-class citizen." It is not sufficiently tracked, managed, curated, protected, or stored in formats that make it readily accessible by AI developers and testers, and AI models. The DAF has not established policies and practices for building and sustaining the data-management pipelines crucial to modern AI development. The DAF does not have the modeling and simulation architectures, synthetic environments, digital twins, or computational power needed to support developing, testing and sustaining advanced AI-enabled systems. These deficiencies,

[23] As noted by the 96 OG/CC, because Eglin AFB is a designated Major Range and Test Facility Base (MRTFB) and is funded through a "pay to play" model (as directed by the NDAA), DAF leaders must address the disconnect between the timelines inherent in this type of funding model, and the certainty of needing immediate funding for high-priority emerging AI T&E requirements. The DAF AI T&E champion will also need to assess the impacts of traditional contractually mandated response timelines when responding to urgent and emerging AI T&E requirements.

if not redressed, will adversely affect all aspects of AI-based systems development, including T&E activities.

Building off the 2020 DoD Data Strategy, the DAF should update its data vision and strategy to explicitly recognize data as a "first-class citizen." This strategy and accompanying implementation plan should include policies and establishing processes to track, manage, curate, protect, and store data in ways optimized for AI developers and testers and that account for possible sources of bias in data. The DAF needs to provide guidance on building and sustaining DMPs, to include highlighting government and private sector best practices for collecting and generating AI-ready data. Data at all levels of classification should be stored in the purview of a zero-trust network architecture, particularly accounting for data privacy when systems are trained on sensitive data.

The committee recommends storing and protecting data at all levels of classification within the purview of zero-trust network architecture and accounting for data privacy when systems are trained on sensitive data.

Recommendation 4-5: Building off the 2020 DoD Data Strategy, the Department of the Air Force should update and promulgate its data vision, strategy, and metrics-based implementation plan to explicitly recognize data as a "first-class citizen." These documents should include plans for the following:

- **Prioritizing investments in computation and storage resources and infrastructure to support artificial intelligence (AI) development**
- **Widely expanding data collection and curation for the entire range of AI planning and scoping, designing, training, evaluation, and feedback activities**
- **Investing in data simulators for AI training and testing**
- **Adapting approaches and architectures developed in private industry for AI-based systems**

4.4 CULTURE CHANGE AND WORKFORCE DEVELOPMENT

The concept of culture is much easier to experience viscerally than it is to define or even describe adequately. Yet culture is very real. It materially affects an organization or community's health and long-term performance. In general, culture refers to a set of shared behaviors, beliefs, and values. It is formed over time, resulting from the combined actions and words of all the people within an organization or community. While an organization or community's leaders play a paramount role in establishing a particular culture through the promulgation of their vision, mission, and value statements; their leadership philosophy and style; the way they treat members of the organization; the norms they establish and enforce; and how they incentivize good

and correct bad behaviors, organizational culture can only be formed, sustained, or changed by the collective behaviors of the entire organization or community over time.

While establishing and sustaining a particular culture is difficult, it is even harder to change an ingrained culture formed over many decades that is viewed as unique, elite, and highly successful. Those three qualities describe the culture of today's DAF test enterprise.

It is hard to argue with past successes. While the committee cannot offer a "recipe" for culture change, it nonetheless believes that culture changes are necessary to ensure AI T&E's best practices, processes, and procedures are adopted as rapidly as possible across the DAF. As noted throughout this report, despite many commonalities between traditional T&E and AI T&E, there are also notable differences. In particular, these include the lack of a clear delineation between DT and OT for AI capabilities; the importance of and reliance on agile principles and adaptive T&E principles (AIOps, MLOps, or DevSecOps) instead of waterfall development for AI systems; the centrality of data and high-end computing; the potential for a continuous data-based self-learning capability; the importance and challenges of mission- and domain-specific adaptation for AI-enabled systems; probabilistic or statistically predictable behavior rather than deterministic results; the effects and risks of dedicated adversarial attacks against AI models, at every phase from initial algorithm training through model deployment and sustainment; the desire for AI explainability and auditability; and the need for continuous integration and continuous delivery (CI and CD) for fielded AI-enabled systems.

The committee asserts that the magnitude of these differences warrants developing a new culture, one that combines the best of the extant test culture with a new and more risk-tolerant, agile, and adaptive mindset and approach to AI T&E. This sort of culture change will be instrumental in accelerating the adoption and integration of AI across the DAF at speed and at scale.

There are inherent dangers in rushing to change the legacy DAF test culture. Attempting to drive systemic changes across the test community without fully understanding the nature and magnitude of the change required or failing to communicate the rationale for change throughout the entire community can cause irreversible harm to the existing culture while simultaneously preventing leaders and organizations from forging and sustaining a culture that can endure for the foreseeable future. For these reasons, it is critical to identify specific aspects of the DAF test enterprise culture that need to be changed and why. Likewise, it is equally important to understand what elements of the existing DAF test culture should be preserved and how. These are not trivial steps. They will require active participation and buy-in from stakeholders and experts across the DAF test enterprise. Initial problem-framing must also include the participation of experts in AI and other emerging technologies from across the DAF, the federal government, and industry and academia—especially those with extensive AI and software T&E

experience, to include recent experiences with leading-edge T&E techniques and adversarial attacks. In essence, DAF leaders should seek to maintain the "best of both worlds": combining elements of today's test culture with new elements that the test community agrees will most likely lead to T&E success in a future environment characterized by software-defined warfare.

Culture change begins at the top. Changing any culture depends on setting and adhering to a coherent vision that aligns strategies, actions, incentives, and metrics. The committee recommends that DAF leaders communicate immediately to the Air and Space Forces both the importance of AI T&E and their commitment to establishing a culture unique to AI T&E through the right combination of people, processes, and technology. At the same time, they should emphasize the value of preserving the successful elements of the current DAF test community culture. The designated DAF AI T&E champion should be equally committed to long-term culture change and should be responsible for recommending changes to DAF leaders that are designed to help forge a new AI T&E culture. The champion should also be accountable for following through on the decisions of DAF leaders.

Workforce development is a critical component of the DAF-wide plan to introduce and sustain new AI T&E capabilities culture. In broad terms, workforce development comprises training, education, certification, and talent management. Because AI remains relatively new, these elements will, of necessity, include both general and test-specific AI training, education, and certification. Similarly, current DAF initiatives and programs that provide AI education and training—led primarily at present by the Department of the Air Force-MIT AI Accelerator (AIA) in coordination with Air University and OSD CDAO—should ensure that all levels of personnel have the appropriate training, from general officers and senior civilian executives to entry-level personnel.[24] CDAO now also has AI education initiatives with JHUAPL and Naval Postgraduate School (NPS)/Stanford. This includes establishing requirements for continuing education and training (CET) on AI and AI T&E-specific topics. It will be equally important for the DAF AI T&E champion to advocate for centralized career-long tracking and management of personnel with specific AI and AI T&E skills, similar to other DAF efforts to manage myriad career fields (appropriate analogies include the cyber, space, and intelligence career fields, which recognize baseline training and certification along with additional identification of specialized training and certifications for specific positions held throughout a career).

The committee recommends that as opposed to general AI training, which can be accomplished by various DoD organizations, core AI T&E training should fall under the AFTC. Since few DAF organizations and agencies presently have the

[24] For example, the DAF-MIT AI Accelerator and the MIT Sloan School of Management host a 3-day AI for National Security Leaders (AI4NSL) education program in Cambridge, Massachusetts.

requisite level of AI and T&E expertise, the committee recommends that the DAF rely on UARCs and federally funded research and development centers (FFRDCs) to run AI T&E training, under the oversight of the DAF AI T&E champion and supported by AFOTEC, the USAFWC, Air Force Institute of Technology (AFIT), AFRL, and AIA.[25] Furthermore, the AFTC AI T&E curriculum should be developed by personnel with substantial AI and AI T&E experience, not only from within the DAF but also, as appropriate, from industry and academia. The committee expects the test community will achieve better results this way rather than relying primarily on retraining AFTC, AFOTEC, or USAFWC test personnel on AI principles and AI T&E processes, practices, and procedures.[26]

The committee recommends that the DAF assess the utility of using the law school analogy for building a cadre of AI T&E personnel across the test enterprise. Just as all lawyers receive common core education on the law followed by extensive additional, specialized training for their planned area of practice (tort law, criminal law, contract law, and so on), DAF AI T&E personnel can participate in a common core test curriculum at the AFTC, with AI T&E-specific training (and training on other emerging technologies) provided either within the AFTC or at other designated DAF or joint organizations, such as the DAF AIA, AFIT, AFRL, or the Defense Acquisition University (DAU). Moreover, the importance of continuing education and training (CET) has its own analogy in the legal profession: as mandated by law, lawyers require so many continuing education units (CEU) annually. The committee suggests that the importance of CET is even greater for AI, considering the exponential rate of technological change.

As noted earlier in the summary, there must be sufficient flexibility at the operational and tactical levels to accommodate agile and CI/CD principles and continuous T&E. This may require deliberate placement of AI T&E experts within operational and training units outside the traditional DAF test community. Some of these people may already be test-certified (similar to how test pilots continually rotate through operational and training squadrons throughout their careers) and only require AI T&E "top-off" training. Others may possess useful skills (a computer science background or previous AI experience, for example) but have not

[25] Similar for example to OSD DOT&E's use of IDA for providing analytic support to DOD's T&E community.

[26] This expectation accounts for the substantive differences between traditional T&E of hardware systems, and AI T&E. The committee acknowledges the potential utility of a hybrid approach that takes advantage of the expertise of both highly experienced "traditional" test personnel, and people with extensive experience in the development, testing, fielding, and sustainment of AI-enabled systems.

received training at the AFTC and thus should receive tailored AI T&E training aligned to their unit responsibilities.[27]

Whenever feasible, the DAF should take advantage of existing AI-related education and training initiatives. For instance, in response to congressional direction, in 2020, the JAIC, now CDAO, developed the *2020 Department of Defense Artificial Intelligence Education Strategy*.[28] In crafting the AI education strategy and implementation plan, the JAIC segmented the entire DoD workforce into six AI archetypes: specifically, personnel grouped by similar AI education and training needs.[29] The committee recommends that the DAF continue to use these same archetypes in developing AI and AI T&E-specific training and education. The JAIC initiated an AI education pilot program in October 2020. The CDAO, in coordination with the DAF AIA, now offers a variety of AI training programs and courses for personnel across DoD. The AIA has compiled a list of AI educational resources for DoD personnel, which can be accessed with a common access card (CAC).[30] Similarly, the 96th Operations Group Commander briefed the committee that the 96th is developing AI T&E educational programs for the test community to address the implications of lethal autonomous weapon systems (LAWS), human factors, and human-systems integration.

The committee also recommends that the DAF AI T&E champion consider using the DAU's approach to modernizing the DoD T&E acquisition workforce as a guidepost for developing DAF-wide AI T&E education, training, and certification. DAU is pivoting from a "one-size-fits-all" certification framework to a component and workforce-centric, tailorable, continuous learning construct.[31] The DoD

[27] The committee suggests the AI T&E champion, in coordination with the AFTC, AFOTEC, USAFWC, DAF CDAO, AFMC Digital Transformation Office (DTO), and DAF Chief Experience Officer (CXO) assess the value of placing "digital natives" at the unit level. Analogous to the practice, for example, of placing unit intelligence officers within DAF squadrons.

[28] Section 256 of the National Defense Authorization Act (NDAA) for Fiscal Year 2020 directed the Secretary of Defense to "develop a strategy for educating service members in relevant occupational fields on matters related to artificial intelligence." It also directed the secretary to develop an implementation plan. (DoD Joint AI Center, 2020, *Department of Defense Artificial Intelligence Education Strategy*, Washington, DC, https://www.ai.mil/docs/2020_DoD_AI_Training_and_Education_Strategy_and_Infographic_10_27_20.pdf.) See Chief Digital and Artificial Intelligence Office, 2023, "Education & Training," https://www.ai.mil/education_training.html for descriptions of the CDAO's AI training programs.

[29] As detailed on p. 7 in *DoD AI Education Strategy*, the six archetypes are Lead AI, Drive AI, Create AI, Employ AI, Facilitate AI, and Embed AI. For detailed descriptions of each archetype, see Appendixes B–G of the *DoD AI Education Strategy*.

[30] C. Del Aguila, 2022, "AI Accelerator Focuses on Education," Air Force Material Command, https://www.afmc.af.mil/News/Article-Display/Article/3013236/ai-accelerator-focuses-on-education.

[31] S. Possehl, 2022, "Test and Evaluation: The Change Is Here Today," Defense Acquisition University, February 1, https://www.dau.edu/library/defense-atl/blog/Test-and-Evaluation-change-today.

acquisition force T&E functional area includes members working in developmental test and evaluation (DT&E), the TRMC, test ranges, and operational test and evaluation (OT&E) throughout all phases of the acquisition life cycle. This DAU initiative focuses on personnel development, streamlining functional areas, reforming the certification framework, modernizing talent management, and equipping acquisition professionals with the tools needed in the digital age. It includes both foundational (within 3 years of position assignment) and practitioner (within 5 years of position assignment) categories.

Moreover, it includes both T&E certification training requirements (basic requirements for working in a designated T&E acquisition position) and T&E credential development (additional training that will provide job-specific, specialty, and point-of-need training for mid- and advanced career jobs and opportunities). The initial set of training credentials includes—among others—T&E of AI, T&E of autonomous systems, evaluating data, T&E of software, and digital engineering (an existing DAU credential). Credentials are intended to be flexible for point-of-need applications. They may be required by senior leaders, functional leaders, supervisors, managers, and others.

Finally, for more advanced AI T&E education and training, the committee suggests that the DAF AI T&E champion review programs offered by the DAU, Air Force Institute of Technology (AFIT), and Air Force Materiel Command (AFMC). For example, AFMC's Air Force Acquisition Instructor Course (AQIC), which is viewed as a "Weapons School for the acquisition career field," includes an entire section on traditional T&E and another on emerging technologies. The committee expects that AFMC and AQIC will be receptive to providing more advanced education and training on AI T&E based on DAF, AFTC, AFOTEC, and USAFWC needs.

Recommendation 4-6: The Department of the Air Force (DAF) should inculcate an artificial intelligence (AI) testing and evaluation (T&E) culture espoused by DAF leaders and led by the AI T&E champion. In particular, the DAF and the DAF AI champion should:

- **Provide AI education, training, and, where applicable, certifications to all personnel, from general officers and senior civilian executives to entry-level personnel**
- **Institute career-long tracking and management of personnel with specific AI and AI T&E skills**
- **Place core AI T&E training under the Air Force Test Center**
- **Take advantage of existing AI-related education and training initiatives**

One of the most important first steps is to survey the entire DAF workforce to determine as accurately as possible the current baseline of AI and AI T&E skills that exist in the DAF today. The committee heard the resounding message from several speakers that such a baseline does not exist—not for general AI skills or

even more important for this report, for AI T&E experience. The DAF AI T&E champion should coordinate with the Air Force Personnel Center (AFPC) and other organizations, such as Air University and the DAF AIA, to develop and administer this DAF-wide survey. The DAF AI T&E champion, in coordination with the DAF CDAO, 96th Operations Group, AFPC, Air University, and the AIA, should consider taking the same approach used when developing the Air Force Computer Language Self-Assessment (CLSA) program in 2019. The CLSA, administered by Air University, allows DAF active duty, reserve, and civilian personnel to assess their knowledge and skills in various computer programming languages.[32] Modifying the CLSA to allow personnel to identify their AI and AI test-specific skills and any formal AI training courses and certifications, while not perfect, is the best way to accelerate developing a DAF-wide baseline of personnel with AI and AI test-specific credentials.

Once a baseline of personnel with AI and AI T&E experience is established, the DAF AI T&E champion should coordinate with the applicable organizations and agencies to develop a tiered approach to AI implementations and AI T&E-specific education, training, and training certification. This includes modifying existing programs to reflect the needs of the test enterprise. Currently, Air University and the AIA use a useful approach for DAF-wide AI education and training: a three-tiered system that begins with basic training on digital skills through online courses offered by Digital University;[33] a second tier focuses on digital skills for basic practitioners and mid-level managers (similar to what exists today for personnel in the cyber field); and a third tier comprises in-depth training, up to and including the designation of expert status. The DAF should consider using this approach for AI T&E training.

The DAF will be unable to build an AI T&E workforce as rapidly as needed to meet expected demands over the next 5 years. However, in the near term, the DAF AI T&E champion, supported by DAF senior leaders, should use the survey results described above to coordinate across the entire DAF to help rebalance the test force by shifting people with needed expertise into the test enterprise. At the same time, DAF test leaders should solicit volunteers from within the test community to be trained specifically on AI T&E. Part of this process includes, with the support of AFPC, formally designating with Air Force Specialty Codes (AFSC) and special experience identifiers (SEI), people who have certain AI and AI T&E skills—similar to how various career fields, including the air force test community—identify special

[32] The CLSA is a self-paced, online program comprising a series of tests and exercises designed to evaluate an individual's knowledge of programming concepts and techniques. Such a survey could also be used to gauge interest in entering the test community as an AI and AI T&E specialist.

[33] Digital University is a joint venture started between the Air Force and Space Force, and is available to members of DoD. It provides access to Silicon Valley-accredited technology training and fosters a community of learners. It includes coding, data science, and product management training.

skill sets today. Once the DAF embarks on this path, it will be equally important to continue to track these skills throughout a person's career. Given the dearth of AI T&E expertise in the DAF today, the Air Force and Space Force can ill-afford to place personnel with these skills in positions unrelated to AI and AI T&E (with normal exceptions granted for career development at more senior levels in the officer, enlisted, and civilian ranks).

Recruiting and retaining AI expertise remains one of DoD's biggest challenges. While this is a multifaceted problem with no single solution, the DAF should take advantage of numerous extant DoD-wide initiatives to find, recruit, and retain the nation's best AI talent. Likewise, the DAF can take advantage of lessons from the standup of U.S. Cyber Command to ensure that military personnel, once trained, are tracked throughout their careers (as noted above) and to the maximum extent feasible retained in AI and AI T&E-related positions. Other creative ideas could include hiring contractors to work within DAF T&E facilities as AI T&E subject matter experts (SME); offering scholarship funds to undergraduate or graduate AI (or related) majors, with the caveat that the individual would serve for a designated period after graduation;[34] reviewing the Science, Mathematics, and Research for Transformation (SMART) Scholarship-for-Service Program to ensure the appropriate emphasis on soliciting undergraduates for AI T&E; and reviewing the DAF's programs for sponsoring graduate-level AI T&E work for military and civilian personnel serving in AI-related positions.

The DAF should also take advantage of Section 605 of the 2019 NDAA to help jump-start building an experienced AI T&E workforce.[35] This section allows accelerated temporary promotion opportunities for officers with skills in areas designated to have a critical shortage of personnel. Section 605 applies as long as the Secretary of the Air Force designates AI and AI T&E as areas that are critically short of personnel.

Recommendation 4-7: The Department of the Air Force (DAF) should determine the current baseline of artificial intelligence (AI) and AI testing and evaluation (T&E) skills across the DAF, develop and maintain a tiered approach

[34] The "National Security Commission on AI Final Report" includes several recommendations along these lines, to include the establish a new digital service academy and civilian national reserve to grow tech talent. See NSCAI, 2021, *National Security Commission on Artificial Intelligence Final Report*, Arlington, VA, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. The FY2023 National Defense Authorization Act contains provisions authorizing DoD to establish a cyber and digital service academy. As proposed, the academy will provide scholarships for up to 5 years, in exchange for equivalent years of service in a civilian DoD position focused on digital technology and cybersecurity. Over time, the committee expects that more computer science and AI degree-granting programs will increase the emphasis on AI TEVV, perhaps even including TEVV as a specific subfield of study.

[35] U.S. Congress, 2018, "John S. McCain National Defense Authorization Act for Fiscal Year 2019," H.R. 5515, 115th Congress (2017–2018), https://www.congress.gov/bill/115th-congress/house-bill/5515.

to AI and AI T&E-specific education and training, rebalance the test force by shifting people with needed expertise into the test enterprise, and consider placing personnel with AI T&E expertise into operational units.

4.5 SUMMARY OF IMPLICATIONS OF FUTURE AI FOR DAF T&E

Even as the DAF addresses its current needs and opportunities, it must evaluate emerging AI trends and their likely implications for T&E. Based on trends that the committee sees today, it has identified areas that will likely have significant implications for DAF AI-based systems and the T&E of these future systems. However, given the pace of AI progress, it is difficult to predict with precision which AI advances will be most impactful for Air Force applications. Therefore, the committee recommends that the DAF pursue a strategy that puts procedures and mechanisms in place to continually track emerging AI trends and investigate T&E implications.

4.6 RECOMMENDATION TIMELINES

This chapter makes numerous recommendations about actions the DAF should take concerning AI T&E. While each recommendation is important, the time horizon associated with the recommendations varies greatly. Therefore, for ease of prioritization, this section groups the recommendations into three groups: recommendations that could be addressed immediately, in the mid-term (3–5 years), and over the long term (over 5 years). These time frames are not hard delineations nor meant to be definitive. They may prove to be overly conservative or overly aggressive.

Action on several recommendations can be taken immediately. Appointing a DAF AI T&E champion (Recommendation 4-1), placing core AI T&E training under the AFTC (bullet 3 of Recommendation 4-6), and committing to establishing independent red teams (Recommendation 4-4) can all be implemented quickly.

In the 3- to 5-year range, many more recommendations can be implemented. This includes adopting an AIOps and MLOps approach for AI-enabled capabilities (Recommendation 4-2) and integrating AI requirements into the program of record (Recommendation 4-3). This would also be the timeframe where the DAF-widevision and strategy for data would be updated and promulgated (Recommendation 4-5), and the AI education parts of Recommendation 4-6 would be implemented. Having established the current baseline of AI and AI T&E skills across the DAF, this would also be the time frame where the DAF should develop a tiered approach to AI and AI T&E education and rebalance the test force by shifting people with needed expertise into the test enterprise (Recommendation 4-7).

Beyond a 5-year window, coordinating investments to explicitly address the findings of red-team activities (Recommendation 4-4) and inculcating an AI T&E culture (Recommendation 4-6) will be key.

5

AI Technical Risks Under Operational Conditions

This chapter will consider the risks of incorporating artificial intelligence (AI) within Department of Defense (DoD) operational systems. These AI-enabled systems have several realistic threats, some based on adversarial AI and others based on the risk in deploying the AI-enabled system in an operational environment. The employment of AI-enabled systems can have significant benefits in augmenting the capabilities of the warfighter, but there are other risks inherent in the use of AI-enabled systems that must be considered. In particular, this chapter answers the second of the committee's three primary tasks, "consider examples of AI corruption under operational conditions and against malicious cyberattacks."

5.1 INTRODUCTION

An AI-enabled system includes technical risks that must be considered during the test and evaluation (T&E) of any system that incorporates AI elements. The most likely risks to AI-enabled systems are enabled through cyber access to the AI components, be they the training or operational data, the model, the software implementing the component, or the output of the AI-enabled system. Thus, the first and most important risk is exploiting vulnerabilities in the system to access, manipulate, or deny the elements of the AI component. Therefore, during T&E, the traditional cybersecurity testing should be augmented by attacking the AI component's availability, integrity, and confidentiality and privacy. These particular attacks are of high consequence to any AI-enabled system as an adversary may be

able to exploit cyber vulnerabilities of the system to affect the operation of the AI component. Strong cyber defense is the first line of defense against many (probably most) adversarial attacks. However, the Department of the Air Force (DAF) T&E process cannot test for all potential cyber breaches of AI models. Therefore, in-depth defense is called for, which employs a zero-trust architecture and T&E of AI-based software defects and operational performance degradation.

Beyond the traditional cyber risks that would enable the adversary to access the AI component, other risks are unique to the operation of the AI component. These involve the data supply chain (training and test data), the model, and the manipulation of the domain to control the operation of the AI component. These risks are described in more detail in this chapter. Mitigating these risks may involve new T&E approaches that mimic an adversary's actions during all phases of the AI-component life cycle to assure that the AI component is resilient during the development, training, deployment, retraining, and adaptation of the AI implementation to different operational environments.

Even with strong cyber defenses in place, the adversary can still attack an air force AI-based system. In particular, physically-based attacks using camouflage, concealment, or deception are possible avenues to biasing, denying, and poisoning training data, training labels, and operational data. Particularly insidious are backdoor attacks, which poison the training data (without changing the training labels) and then use triggers during actual operations to misdirect or degrade the AI models during missions. This motivates the need for DAF red-teaming to uncover these potential attack vectors and their likely effects. DAF-specific adversarial training and detection algorithms that target the vulnerabilities discovered by red-teaming can be incorporated into AI-based systems and tested by DAF T&E processes.

AI-based systems are subject to environmental and adversarial effects that can degrade performance. The current DAF requirements-driven T&E processes can uncover the effect of these attacks if the AI model performance is degraded below the required, acceptable ranges. Thus DAF DT&E that tests performance with hold-out datasets can be employed as a defense against some attacks. In addition, OT&E performance tests can ferret out operational environmental degrading effects, which may motivate model retraining using new datasets.

Some attacks, referred to as backdoor attacks, only manifest once the adversary triggers them. Unfortunately, it is intractable to test against all possible triggers, and techniques that hide triggers to make them difficult to detect by humans and machines have been developed. Thus, the mitigations, in this case, are (a) to make such attacks hard to accomplish (which will be discussed further below) and (b) in the case where the attack intends to degrade overall performance, to build monitor systems that check the AI components during run-time for performance degradation and take appropriate action. The T&E role, in this case, is to cause the

AI model to fail through a simulated backdoor attack and then to test the monitor to ensure it detects the deviation.

The DAF T&E process is responsible for uncovering attacks and environmental situations that generally degrade performance. The DAF will have to depend on cyber defenses to broadly restrict adversarial attacks by denying access to data (training and testing) and queries (to protect against inversion and privacy attacks). T&E can also test monitors that look for out-of-spec performance deviations. Finally, T&E can test against threats that red teaming has uncovered to the extent that these threats are detectable. This implies a very close relationship between red-teaming and T&E.

5.2 GENERAL RISKS OF AI-ENABLED SYSTEMS

From a T&E perspective, the areas of an AI-enabled system that must be protected are the availability, integrity, and confidentiality and privacy of the AI system:

- Availability and integrity of the AI output
- The integrity of the AI model, data, and software
- Confidentiality and privacy of the model and training data

Several risks are inherent to using AI in operational environments. Awareness is the first step in mitigating such risks, so the committee discusses here some main challenges that have been identified.

AI is dependent on its training data, and its predictions are only as good as the data it has been trained on. Limited data in some scenarios can lead to inherent biases and risks. For instance, having no training data from below-freezing weather conditions means the AI model will not be accurate in such operational conditions.

Additionally, AI implementations are very sensitive to distribution shifts, which often happen gradually and slowly degrade the performance of an AI element, independent system, or joint cognitive system. This could be caused by slow changes (e.g., in the landscapes in satellite images, degradation of sensors due to dirt, and other factors). Even updates to sensor software can result in distribution shifts in the data. Detecting these distribution shifts is a major challenge in AI.

At the extreme, out-of-distribution (OOD) AI predictions occur when a model is presented with data outside of the distribution it was trained on. Unfortunately, AI models cannot robustly detect if data are OOD; in other words, they do not know what they do not know. Instead, AI models learn to make predictions based on patterns and relationships specific to the training data. As a result, they do not generalize well to novel data, and even a bit of extra signal noise in a sensor collection, which might be invisible to the human eye, can potentially confuse an AI system. For example, in object classification in image data, small perturbations of the image data can create a movement toward a centroid within the ML classifier that will mis-classify

the object. This was famously demonstrated by the noise introduced in the classification of stop signs, causing the system to classify the object as a speed limit sign.[1]

5.3 AI CORRUPTION UNDER OPERATIONAL CONDITIONS

In January 2023, the Office of the Undersecretary of Defense for Policy issued DoD Directive 3000.09, which states that the Director of Operational Test and Evaluation (DOT&E) "[e]valuates whether autonomous and semi-autonomous weapon systems under DOT&E oversight have met standards for rigorous V&V and T&E in realistic operational conditions, including potential adversary action, to provide sufficient confidence that the probability and consequences of failures have been minimized."[2]

This latest policy on autonomous weapon systems illustrates the nature of placing adversarial attacks against AI-enabled systems in the context of all threats in realistic operating conditions. Thus, while this report focuses specifically on the question of adversarial attacks against the AI components, it is important to continue to test and evaluate AI-enabled systems against a broad spectrum of adversaries and attacks in a realistic operational context. In addition to development and test datasets, the collection and use of operational data, especially in training AI components, is key to mitigating threats to the AI component and the system as a whole.

The consideration of adversarial AI should be in addition to all traditional cyber threats associated with more traditional integrated systems. Software vulnerabilities, supply chain vulnerabilities, insider threats, network vulnerabilities, denial of service attacks, privilege escalation, and root of trust attacks are just a few of the traditional threats that remain in AI-enabled systems and must be addressed. These threats frequently dominate the new categories of AI corruption based on the incorporation of AI technology. While these novel attacks are important to detect and mitigate today, the ability of an adversary to attack the software and networks of the integrated system may be an even larger risk to the operation.

Also of note is the rapid evolution of attacks against AI-enabled systems. This results from the attention on AI systems in academia, the private sector, and the government. This is similar to the rapid evolution of cyberattacks in the early 2000s when attention was centered on network-enabled systems. The committee would expect adversarial attacks against AI-enabled systems to follow the same pattern; rapid evolution of individual attacks followed by a more comprehensive set of mitigations on attack strategies and consistent policy and taxonomies on AI attacks. Some examples of adversarial attacks on AI are shown in Box 5-1.

[1] K. Eykhold, I. Evtimonv, E. Fernandes, et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," arXiv:1707.08945, https://arxiv.org/pdf/1707.08945.pdf.

[2] N. VanHoudnos, B. Draper, J. Richards, J. Schneider, and N. Carlini, 2022, "DoD Zero Trust Strategy," Washington, DC: Department of Defense, https://dodcio.defense.gov/Portals/0/Documents/Library/DoD-ZTStrategy.pdf.

To properly scope the discussion on AI corruption, it is important to clearly define the concept of AI corruption. Although there is significant literature on various attacks against AI systems, there is no standard definition to date for AI corruption. In this context, the committee defines AI corruption as:

> AI corruption is the deliberate or unintentional manipulation of the data, hardware, or software of an AI-enabled system that causes the system to produce missing, inaccurate, or misleading results, to deny or degrade the use of the system, or to force the system to expose hidden information used in the training or configuration of the AI component.

The result of AI corruption is a decrease in the quality attributes of an AI component. This may be in the form of statistical measures such as precision and recall, reduction of a performance envelope required for a mission objective, or in a violation of the system's security requirements, such as maintaining the secrecy

BOX 5-1
Some Examples of Adversarial Attacks Against AI Systems[a]

Adversarial examples: These are inputs to a machine learning model that are specifically designed to mislead the model and cause it to make incorrect predictions. Adversarial examples can be created by adding small, imperceptible perturbations to the input data designed to fool the model.

Poisoning attacks: These attacks involve introducing malicious or misleading data into a machine learning model's training data in an attempt to corrupt the model's output.

Evasion attacks: These attacks involve manipulating the input data specifically to evade detection by a machine learning model that is being used for security or fraud detection purposes.

Model inversion attacks: These attacks involve attempting to recover sensitive information about the data used to train a machine learning model by manipulating the model's output and using it to infer information about the input data. Also known as a data inference attack.

Model stealing attacks: These attacks involve attempting to reverse engineer a machine learning model by studying its output and attempting to recreate the model's underlying structure and parameters.

Overfitting attacks: These attacks involve training a machine learning model on a dataset that is not representative of the data it will encounter in the real world, leading to poor generalization performance.

Explainability attacks: These attacks involve manipulating the output of an explainable AI system in an attempt to mislead or deceive the user.

[a] V. Shepardson, G. McGraw, H. Figueroa, and R. Bonett, "A Taxonomy of ML Attacks," *MLSEC Musings* (blog), Berryville Institute of Machine Learning, https://berryvilleiml.com/taxonomy.

of the training set for the AI component. The source of the AI corruption may be a deliberate cyber or physical attack (e.g., destroying a critical sensor or breaking into the server running the AI component) or a result of accidental or environmental conditions (heavy rain or fog distorting sensor input or fault of the hardware or software supporting the component).

5.4 ATTACK SURFACES FOR AI-ENABLED SYSTEMS

The attack surface of any system is defined by NIST (NIST SP 800-172 from GAO-19-128) as the set of points on the boundary of a system, a system element, or an environment where an attacker can try to enter, cause an effect on, or extract data from, that system, system element or environment. For all systems, AI-enabled or not, this defines a surface that must be secured against threats and tested during T&E. Thus, all traditional testing for adversarial attacks against the attack surface that has been previously defined is still required. Defensive mechanisms that obviate or limit the capability of an adversary to take advantage of an attack surface are also the first line of defense for AI-enabled components.

The traditional cyberattack surface can be considered a starting point for defining an attack surface for an AI-enabled system. The rationale for starting with the cyberattack surface is that the AI-enabled component is a data-driven software system, so it shares much of the same surface for an adversary to disrupt, deny, or degrade the system containing the AI-enabled component, which could also provide access by the adversary to the data or software in the AI component.

In addition to the cyberattack surface, the AI component may enable an attack surface beyond the traditional cyberattack surface. This access is due to the dependencies on data within the deployed environment and in the backend infrastructure, the supply chain of any AI model in the component, and the potential for retraining and adaptation with an adversary-controlled environment. These attacks, as described below, expand the traditional attack surface and should be considered for the T&E of any AI-enabled system.

Vulnerabilities in AI-enabled components may be addressed by limiting adversarial access to this component through traditional separation and protection mechanisms. Thus, removing traditional vulnerabilities and adding robust protections to systems containing AI components can limit an adversary's ability to influence, deny, degrade, or corrupt the functions of the AI component. In all cases, the limitation of access by the adversary to the component is considered the first line of defense in preventing AI corruption in operational conditions. These include but are not limited to network protections, authentication, and authorization to system functions, data at rest and data at motion protections, distributed system protections, rate limiting to prevent denial of service attacks, and robust

sensor and actuator protection. The committee can address the white box-black box considerations in adversarial attacks and defenses.

An AI-enabled system consists of many components—those that are specific to AI and components that are part of a more traditional system supporting other functional components. For example, a system that uses a model for object detection may be part of a larger fix and target system supporting a weapons platform. So, the attack surface of a system that includes AI-enabled components also includes the traditional attack surface of the classic system supporting the AI component. This is especially important in the T&E of the system as a whole, as the details of how the AI component is integrated with the traditional system may expand or limit the entire attack surface, which might lead to AI corruption.

While this definition is relevant for AI-enabled systems, some additional vectors should be considered in the life cycle of the AI component. For example, an AI component relies on sensor data free from adversary manipulation, which may be difficult in operational environments. An adversary may alter the environment to cause the AI component to miss-classify an object, thus attacking the system without going through the traditional attack surface.

Another way an AI-enabled system may be attacked beyond the traditional vulnerabilities is by forcing the AI component to respond to many sensor inputs by exposing some of the data used to train the AI model (known as an inversion attack against the ML system). These attacks are effective even with limited access to the ML model and can expose the limitations of the AI component to the adversary.

Yet another attack that is usually not part of a traditional attack surface is the manipulation of training data to an ML component to cause the model to learn a response beneficial to the adversary in an operational deployment. This can be accomplished not only in the initial training of the model but during retraining when the model is updated to respond to updated environmental conditional post-deployment.

AI model inversion attacks refer to techniques that aim to reverse engineer the internal workings of an AI model. These attacks are a type of adversarial attack in which an attacker seeks to reconstruct the input data or features used to train a model or to generate synthetic inputs that will produce a desired output from the model.

Model inversion attacks are a potential concern because they could allow an attacker to learn sensitive information about the training data or the training process itself, which could exploit the model's vulnerabilities or craft adversarial examples that can fool the model. Model inversion attacks can be particularly dangerous when applied to models used in high-stakes situations, such as DoD

weapons systems or decision support systems, as they could result in incorrect or biased decisions.

Several methods can be used to defend against model inversion attacks, including techniques such as differential privacy to obscure the training data and designing models to be robust against adversarial examples.

In summary, the attack surface of AI-enabled systems encompasses all traditional attack surfaces inherent in software-intensive systems—especially those of ML-enabled systems—but has the additional considerations of AI corruption through the life cycle of the AI system, including the data used in training, test, and operations, and the access to the details in any configuration or model that is part of the AI component.

5.5 RISK OF ADVERSARIAL ATTACKS

It can be useful to divide the risk of adversarial attacks into the different levels of integration of the AI-enabled system. The AI component is tested as a standalone system at the most basic level. At this level, the primary risk of adversarial attack is in the supply chain of the software and data used to construct the AI component. The software may rely on open-source components, usually modified for the specific needs of the DoD mission. These compounds may contain elements contributed by the adversary that have not previously been identified by the open-source community. The adversary may also have identified underlying vulnerabilities in the open-source components that have not yet been publicly released, and these may be incorporated into the delivered DoD component. It is important at the DT&E stage to utilize a well-resourced red team with the most up-to-date attack knowledge to expose any potential vulnerabilities that have become part of the software. This case is no different from any other software component that must be tested, but the complexity of open-source AI solutions may be difficult to thoroughly test. Other types of analysis, such as static and dynamic software testing, would be appropriate to augment the red team approach. Modern DevSecOps software pipelines include such tools and should be considered for complex AI incorporation.

It should also be noted that some legacy languages and software stacks may contain numerous vulnerabilities that enable adversarial access to the key data and software elements of an AI component. The use of modern type-safe languages can help to mitigate some of these potential vulnerabilities. Requirements and subsequent T&E for languages that are resilient to attack are an important mitigation technique for AI-enabled systems. In addition, code generation using large language

models such as ChatGPT may produce vulnerable code with unsafe languages such as C. Generation of code in a type safe language can help to mitigate some of these risks.[3]

During the development of an AI component, especially those related to machine learning (ML), the data used for training and testing the component is an important and critical element of exposure to potential adversarial manipulation. When open-source data are used, even if augmented with specific mission data, the potential for adversarial manipulation of the open datasets to force specific ML behaviors can be difficult to detect and can effectively compromise the ML-enabled system in deployment. Protection of the training and test datasets from adversarial manipulation is an important and specific need for ML components. Note that even exposing the training and test datasets to the adversary without their manipulation can provide the adversary with an important tool to discover operational manipulation techniques that can force the ML-enabled system to fail during deployment, even if these components function properly in DT&E and OT&E. Typically this system is tuned and tested by a data scientist, part of the development team for the ML software component.

One recent advance in defense of ML components is using adversarial training. This augmented dataset specifically attacks the function of the ML component, which can then be used as a training set in the component to drive down that specific behavior. An iterative process of continuing to use adversarial examples and train the component to behave appropriately even in these conditions can add to the robustness of the model. Note, however, that over-training in the adversarial space could create vulnerabilities in the inversion attack of the resulting model, thus exposing which attacks and training sets are used to an adversary with limited access to the resulting deployed system.

Recognizing the planned mission objectives is also important when testing these components. Often the theoretical maximum performance of an ML component is tuned to some imagined optimal point, but operational requirements may dictate a different point to tune the trade-offs inherent in machine learning. For example, it may be necessary to have a high-precision result at the expense of recall for a fix and target application, but for a surveillance mission, a high-recall result may better fit operational needs. This should be recognized during DT&E, shown in Figure 5-1, so that by the time this component is integrated into the overall system, the OT&E will have the optimal mission benefit to the ML component.

[3] J. He and M. Vechev, 2023, "Controlling Large Language Models to Generate Secure and Vulnerable Code," arXiv:2302.05319, https://doi.org/10.48550/arXiv.2302.05319.

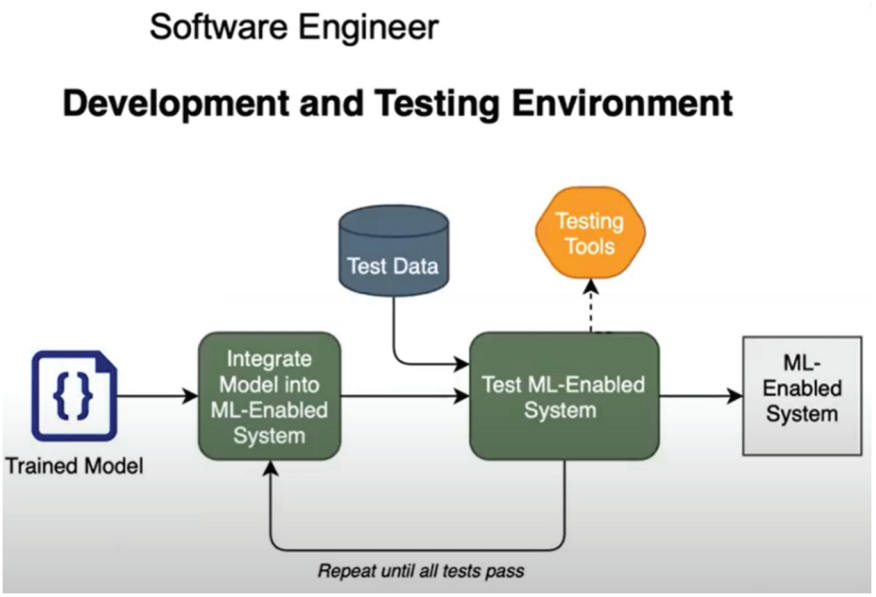

FIGURE 5-1 A model development and testing environment. SOURCES: I. Ozkaya, 2021, "What Is Really Different in Engineering AI-Enabled Systems?" Pittsburgh, PA: Carnegie Mellon University Software Engineering Institute. Images re-used with permission from Carnegie Mellon University. First publication: Grace A. Lewis, Stephany Bellomo, Ipek Ozkaya: "Characterizing and Detecting Mismatch in Machine-Learning-Enabled Systems." WAIN@ICSE. 2021:133–140.

General advice published on testing ML-enabled components consists of the following:[4]

- What are you intending to test (and learn)?
- What logistical challenges might you encounter during testing?
- What are your biggest sources of risk?
- What is the meaning behind your metrics?
- How are you dealing with the scale and level of complexity of your system?
- How are you evaluating for bias and other unintended behaviors?

Once the systems are integrated into an operational component, there will be additional context for the attack surface. These concern the application interfaces to this component and any user interface mechanism and integrated sensor and

[4] V. Turri, R. Dzombak, E. Heim, N. VanHoudnos, J. Palat, and A. Sinha, 2022, "Measuring AI Systems Beyond Accuracy," paper presented at AAAI Spring Symposium Series Workshop on AI Engineering: Creating Scalable, Human-Centered and Robust AI Systems, https://doi.org/10.48550/arXiv.2204.04211.

actuator components that will interface with the component. At this phase, the software engineer must work with the data scientist to assure that the ML-enabled component is sufficiently protected from these interfaces to assure the adversary will not have a path to manipulate the operation of the ML component. During the life cycle, whether DevOps or traditional waterfall, isolation of the model, test data, and testing tools is necessary to prevent exposure of vulnerabilities specific to the trained model that might be exercised during deployment. In particular, if the adversary manipulates the test data, some important operational scenarios may not be tested properly, and the subsequently deployed system may not function properly during these scenarios. This step is demonstrated in Figure 5-2.

At the level of integration of the ML component into the operational system, the addition of software components and data paths expands the attack surface to include many elements that are likely exposed to the adversary, including the sensor input, data streams, and specific APIs of the system to other systems. In this case, many of the more traditional test procedures for an attack surface are relevant and appropriate with the addition of some specific control over these attack surface elements (e.g., sensor input) by the adversary for the express purpose of causing the ML component to fail. In addition, network security, data protection, encryption, and other classic defenses must be integrated at this point into the operational system and tested as part of the OT&E process.

Finally, the risk of adversarial attack during deployment is increased over the traditional software-intensive systems by creating an additional attack surface

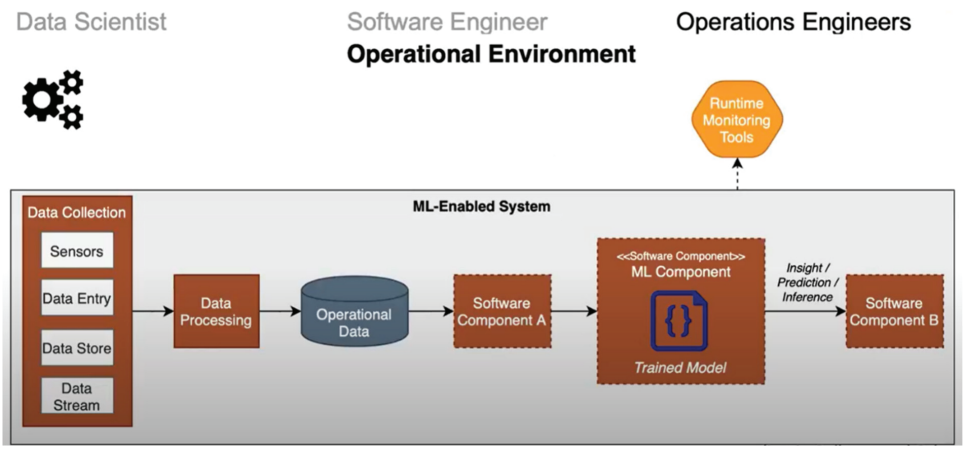

FIGURE 5-2 The operational environment for ML-enabled systems. SOURCE: I. Ozkaya, 2021, "What Is Really Different in Engineering AI-Enabled Systems?" Pittsburgh, PA: Carnegie Mellon University Software Engineering Institute. Images re-used with permission from Carnegie Mellon University. First publication: Grace A. Lewis, Stephany Bellomo, Ipek Ozkaya: "Characterizing and Detecting Mismatch in Machine-Learning-Enabled Systems." WAIN@ICSE 2021:133–140.

in the operational data (sensor and other inputs to the AI component), the life cycle of the AI-enabled component, and the additional vulnerability of model inversion.

5.6 NETWORK SECURITY AND ZERO TRUST IMPLICATIONS

Modern network security within DoD relies on the principles of zero trust (ZT). The framework and approach taken by DoD must be directly applied to all AI-enabled systems during the entire life cycle of these components to address the risks identified above. While not all aspects of AI-enabled systems may be amenable to a ZT approach, the use of ZT where appropriate can decrease the risk of adversarial manipulation of the data or the software of the AI-enabled system. Areas where this may not be applicable might be cases where the sensor network and data collection infrastructure is outside of the ZT boundary for the system (e.g., open data for training purposes). However, in this section, ZT will be discussed where the approach is appropriate and within the system boundary for deployed AI-enabled systems.

Zero trust is the term for an "evolving set of cybersecurity paradigms that move defenses from static, network-based perimeters to focus on users, assets, and resources." Zero trust uses continuous multi-factor authentication, micro-segmentation, advanced encryption, endpoint security, analytics, and robust auditing, among other capabilities, to fortify data, applications, assets, and services to deliver cyber resiliency. DoD is evolving to become a more agile, more mobile, cloud-supported workforce, collaborating with the entirety of the DoD enterprise, including federal and non-federal organizations and mission partners working on various missions. The zero trust framework will reduce the attack surface, reduce risk, offer opportunities to manage the full range of risks (e.g., policy, programming, budgeting, execution, cybersecurity-specific, and others), and enable more effective data sharing in partnership environments. It will also ensure that any adversary damage is quickly contained and remediated if a device, network, user, or credential is compromised.

As the Zero Trust Strategy states:

> Trusted Interoperability Data for Warfighters: Military targeteers need secure access to data at the speed of relevance they can use and trust. Warfighters need to target the right adversaries accurately while minimizing civilian and other casualties. Today, DoD data is often siloed, in impractical formats, and not fully vetted or secured from the point of origin to use. The execution of Zero Trust provides targeteers trusted, tagged, and labeled data so they can confidently employ and share it with trusted partners, assured that the data is protected, secure, and accessed by only the people who need it when they need it, using least privilege principles.[5]

[5] DoD, 2022, "Zero Trust Strategy."

The benefits of securing AI based on the ZT strategy include user, device, application, data, network, automation, and analytic approaches to securing operational systems. Appropriate implementation of zero-trust capabilities across the life cycle of an AI-enabled system covers much, if not all, of the AI attack surface. This is primarily because the nature of supply chain attacks on the training, test, and validation data for AI models require the same level of authentication and authorization as access to the algorithms and models themselves. This is a fundamental principle of zero trust that all elements of the system, data, network, software, and interfaces be authenticated and authorized by the role of the user or software with access to these capabilities. Attacks against the deployed system are also largely addressed by authentication and authorization of all elements, including sensors and other inputs to the AI-enabled system, to prevent adversarial manipulation of any element in the pipeline of the AI-enabled system.

Stages in securing the network to support operational AI-enabled systems include:

- Data flow mapping. Define granular control access rules and policies. Support least privilege access through a full survey of IT assets, including all AI components, the data they rely on for training, test, and operation, and the trained models designed and deployed.
- Macro segmentation. Define software-defined networking (SDN) APIs. Use software-defined networks to isolate network assets and traffic. This will ensure that any network or component corruption is contained in a single segment and does not spread to other networks and functions. Separating ML training and test datasets from operational networks will reduce the attack surface for operational AI-enabled systems and mitigate several specific AI attacks. In addition, using specific APIs to control data flow and access to the AI life cycle and deployed environments will limit the ability of the adversary to propagate attacks to multiple operational systems. Note this is also a trade-off to a dynamic approach to DevSecOps that would enable near real-time updates to operational models based on retraining and updates on changing deployed environments as well as the use of operational data in the development and training of new models. Note that this approach may increase the risk of adversary manipulation of the retraining data to drift the model to be more beneficial to the adversary once the retrained model is deployed. This is an inherent trade-off in retraining with operational data that includes adversarial input.
- Software-defined networking. Assure that all network transport is tightly controlled and adaptable to conditions that may be under adversarial control. Thus, the use of dynamic network controls, bandwidth, routing, and assurance can be monitored and dynamically adjusted based on any changing conditions.

- Datacenter segmentation. Assuring that access to datacenter resources (including access to data repositories and microservices) is not based on network address or reusable tokens but instead on the continuous authentication and authorization of principles of the access to datacenter services.

Finding 5-1: Existing research on attacks on AI-enabled systems and strategies for mitigating them consider attacks that require unimpeded access to an underlying AI model. These attacks are unlikely to be practical with traditional protections and mitigations inherent in deployed DAF systems.

Finding 5-2: Ongoing research on adversarial attacks on AI-enabled systems focus on performance on benchmark datasets which are inadequate for simulating operational attacks. It appears that as robustness to adversarial attacks is improved, the performance often goes down. Even on benchmark datasets, the trade-off between potential performance reduction and improved robustness is not understood. More importantly, the defenses are designed to thwart known attacks. Such pre-trained defenses are not effective for novel attacks.

Finding 5-3: The impact of adversarial attacks on human-AI enabled systems has not been well understood.

At present, the DoD Zero Trust Strategy is only being implemented on enterprise systems. However, it should also be implemented on all DAF AI-enabled systems. This overarching goal may be done in a series of steps.

Recommendation 5-1: The Department of the Air Force (DAF) should fund research activities that investigate the trade-offs between model resilience to adversarial attack and model performance under operational conditions. This research should account for a range of known and novel attacks whose specific effects may be unknown, but can be postulated based on state-of-the-art research. The research should explore mitigation options, up to and including direct human intervention that ensures fielded systems can continue to function even while under attack. The DAF should also simulate, evaluate and generate defenses to known and novel adversarial attacks as well as quantitatively determine the trade-off between potential loss of performance and increased robustness of artificial intelligence–enabled systems.

Recommendation 5-2: The Department of the Air Force (DAF) should apply the DoD Zero Trust Strategy to all DAF artificial intelligence–enabled systems.

The DoD Zero Trust Strategy concludes:

> To achieve the DoD Zero Trust Strategic Vision, the Department must pursue the strategic goals outlined above as an enterprise. While this is an enormous task, DoD has already made significant progress. Dating over a decade, DoD has advanced cybersecurity through initiatives such as continuous monitoring, multifactor authentication, and others. The technologies and solutions that create ZT, and the benefits it provides, must become a part of the Department's lexicon and be accounted for in every plan and operation.
>
> Cybersecurity in the world today is, by definition, a moving target, and while it may move, the concept and the culture will remain the same, even as the Department adapts and refines the strategy. Ongoing and open communication and coordination, underpinned by proper funding and resourcing, are key to the strategy's success.
>
> The Department's ability to protect, and by extension, DoD personnel against the array of increasingly sophisticated cybersecurity threats depends on it.

5.7 ROBUST AND SECURE AI MODELS

A common approach for increasing the robustness and security of AI models is the incorporation of monitoring or watchdog systems that compare the output of an AI-enabled system to pre-defined operational limits. Should the AI system stray from these operations limits, the external monitoring will take control and prevent the system from drifting beyond these predefined limits. This is similar to guardrails that OpenAI has placed on its ChatGPT system to prevent this system from abuse or offensive results during its use.

Recent studies have shown that the incorporation of guardrails on large language models and similar neural-network AI systems may lead to inaccurate results. For the most recent GPT–4 release, a comparison of the calibration curve of the model prior to the guardrails versus after leads to a significant reduction in the correctness of the results, as shown in Figure 5-3.[6] The figure demonstrates the trade-off of the accuracy of the base model (first chart) where provided answers are correct as they become more available. The second chart demonstrates that with the application of guardrails within the models, the availability of answers [P(answer)] have a lower probability of correctness within the critical range of 0.4–0.8 P(answer). This indicates that with the guardrails, the likelihood of incorrect answers (associated with hallucination) is much higher with the same availability of the answer than the base model without the guardrails. The feasibility of guardrails and monitoring of AI-enabled systems is an area of ongoing research, including monitoring to detect attacks and ensure recoverability of the system.

[6] OpenAI, 2023, "GPT-4 Technical Report," arXiv:2303.08774, https://arxiv.org/abs/2303.08774.

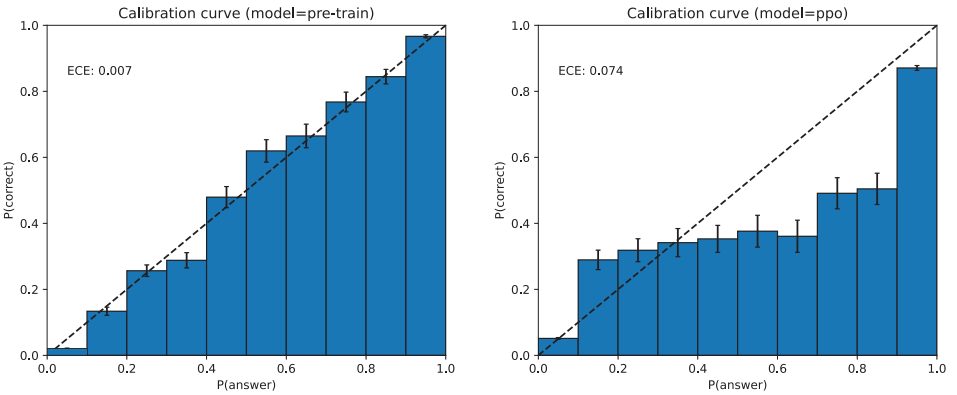

FIGURE 5-3 A comparison of the calibration curve of GPT-4 prior to and after the incorporation of guardrails. SOURCE: OpenAI, 2023, "GPT–4 Technical Report," arXiv:2303.08774, https://arxiv.org/abs/2303.08774.

Robustness is the property that a software component (that includes AI-enabled components) can meet mission requirements with variations in the operational environment.[7] Robustness is measured and tested during T&E with the introduction of deliberate perturbations of the environment beyond the initial configuration and training of the software component. This may be an introduction of environmental variations (weather, background, noise, etc.) or variations in detected signals (new objects, other sensory inputs, or variations on decision support). If the SUT passes these variations, it is deemed to be robust. Note the expectation is that this is similar to stress testing hardware systems to determine the performance envelope of the actual delivered system. This testing also enhances justified confidence (see below) as it can generate a more specific set of operational environments and constraints that is communicated to the operator during deployment.

5.8 RESEARCH IN T&E TO ADDRESS ADVERSARIAL AI

To enable DAF T&E, it will be important to distinguish between practical attacks by near-peer adversaries and academic attacks that would be impractical in deployed systems. In particular, attacks that require unimpeded access to an underlying AI model are unlikely to be practical with traditional protections and mitigations inherent in deployed DAF systems. Nevertheless, as stated

[7] Variations regarding adversarial attacks were discussed in Section 3; in this section robustness also includes both adversarial and non-adversarial variations in the environment.

above, the use of cybersecurity vulnerabilities to reach AI components will continue to be a primary attack vector for the foreseeable future, and research in the mitigation of these vulnerabilities will be increasingly important to AI-enabled systems.

However, there are other AI-specific attacks that do not require unimpeded access to the AI model, its data, or software. These may include model inversion attacks and environmental manipulation attacks that exploit the mechanisms of the AI component without direct access. Even attacks that require unimpeded access to the AI model may be successful due to the transferability of adversarial attacks. Attacks that can be implemented on more accessible models may then be used to attack the target model. Future research to identify and mitigate these attacks should be a priority for DAF T&E.

For identified attacks against AI systems either through cyber vulnerabilities or through manipulation of the data input and model behavior, research to identify these attacks as they are happening in near real time may allow mitigation through response actions. In addition, this traditional approach to intrusion detection and response can include specific attack characterizations of adversarial AI. Research along these lines continues and will be important to include in T&E activities.

Other mitigations, such as external observation and fencing of AI behavior, can also be used to identify adversarial AI and must also be a part of DAF T&E.

Robustness (i.e., graceful degradation) and resilience (countering the effects when detected) against both natural and adversarial corruptions and performance losses are vibrant areas of academic and industry research and constitute an integral part of the OUSD(R&E) trusted AI thrust. Hence, we should see accelerated future progress that the Air Force can exploit. This is discussed briefly in Section 6.1.

The DAF should not just be a fast follower of private sector research and development (R&D) in this area but should prototype advanced applications for DAF-specific situations and systems and should pursue a few key research vectors that are perhaps not as important in academic and industrial settings. Based on the information gathered by the committee, the following R&D thrusts are particularly important:

First, general red teaming R&D of potential adversary attacks for a few distinct scenarios:

1. The scenario where the adversary does not compromise our cyber security but can use camouflage, concealment, and denial to influence model training and achieve model evasion or performance degradation.
2. The scenario where the adversary attempts to compromise our cyber security and then uses a variety of adversarial attacks.

Second, blue teaming of detection and mitigation of the above, and counter AI R&D. Since the adversary may be vulnerable to similar attacks, we need to keep this work at appropriate levels of classification. We should also consider "battle reserve" models as an approach.

There are several promising areas of research that will improve the mitigation of adversarial AI including:

- *Techniques for data sanitization.* In many cases, sensitive data would need to be sanitized to prevent training data from exposure during model inversion or when testing new types of AI-enabled systems. However, current approaches to sanitization do not effectively support the trade-off between the effective training of the system and the possibility of leaking sensitive data. research into new techniques for data sanitization is necessary to resolve this trade-off. The same research may be used in areas such as data privacy, where specific personally identifiable information (PII) or other sensitive data may be used in the training of an AI-enabled system.
- *Quantifiable uncertainty (QU).* DAF systems and operations models should be inherently capable of reporting QU. Through thorough testing, QU metrics should be sufficiently documented to be used confidently in operational contexts or have external monitors or guardrails of performance integrated into their deployed systems. Research into approaches to model-inherent QU is a rich area of enquiry.
- *Certifiable robustness (CR).* The main issue with CR in the recent past is that techniques that work only apply to rather restrictive cases, tend to degrade performance, and often require an inordinate amount of computation. However, recent innovations are showing progress. As Salman et al. (2021), write:

> Certified patch defenses can guarantee robustness of an image classifier to arbitrary changes within a bounded contiguous region. But, currently, this robustness comes at a cost of degraded standard accuracies and slower inference times. The committee demonstrates how using vision transformers enables significantly better-certified patch robustness that is also more computationally efficient and does not incur a substantial drop in standard accuracy. These improvements stem from the inherent ability of the vision transformer to gracefully handle largely masked images.[8]

[8] H. Salman, J. Saachi, E. Wong, and A. Madry, 2022, "Certified Patch Robustness via Smoothed Vision Transformers," Pp. 15137–15147 in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA: IEEE Computer Society and CVF Computer Vision.

Conclusion 5-1: Promising areas of research that will improve the mitigation of adversarial AI include techniques for data sanitization, quantifiable uncertainty, and certifiable robustness.

Additionally, Chapter 6 discusses additional emerging AI technologies and promising areas of research. Thus, the DAF should invest in further R&D at both the foundational level and at the applied level, in particular the DAF in the use of these techniques to DAF AI models/AI-enabled systems.

6

Emerging AI Technologies and Future T&E Implications

New and promising artificial intelligence (AI) techniques and capabilities are on the horizon. Even as the Department of the Air Force (DAF) addresses its current needs and opportunities, it must evaluate these emerging AI trends and their likely implications for test and evaluation (T&E). The committee was tasked with recommending "promising areas of science and technology that may lead to improved detection and mitigation of AI corruption" (see Appendix A). Although it is difficult to predict which AI advances will be most impactful for Air Force applications with precision, five areas seem particularly salient:

- Trustworthy AI
- Foundation Models
- Informed Machine Learning Models
- AI-Based Data Generators
- AI Gaming for Complex Decision-Making

Each of these areas has implications for future Air Force T&E practices and infrastructure needs, as discussed below.

Recommendation 6-1: The Department of the Air Force should focus on the following promising areas of science and technology that may lead to improved detection and mitigation of artificial intelligence (AI) corruption: trustworthy AI, foundation models, informed machine learning, AI-based data generators, AI gaming for complex decision-making, and a foundational understanding of AI.

6.1 TRUSTWORTHY AI

The pressure to reap the benefits of AI technology has encouraged private industry to market AI-based products even though there is no tightly bound theoretical understanding of their performance and robustness. The risk of failure is tolerated because the consequences are acceptable.[1] Commercial AI is generally hardened through continual testing and rapid incremental refinement, usually through extensive user feedback. The DAF should employ this approach whenever possible; nevertheless, it is difficult to confidently engineer robustness and performance into a system when the performance foundations are poorly understood. As military services seek to apply and deploy AI under dynamic and high-risk operational conditions, the need for AI robustness, survivability, resilience, safety, fairness, explainability, ethics, and theoretical performance bounds becomes crucial.

There are several barriers to trustworthy AI. First, current machine learning performance theory lags behind the practical application of AI. For instance, existing theory cannot reliably predict how a neural network architecture will affect performance or how well a learned model will perform in new environments or under new operating conditions. This situation presents a fundamental risk to the trustworthiness of AI and challenges the use of AI in military weapon systems and other safety-critical applications.

A second barrier is a dearth of rigorous testing mechanisms. Testing systems in controlled environments yields over-optimistic evaluations of an AI system's performance, while testing "in the wild" may present significant risks to bystanders; this issue has been observed in catastrophic failures of autonomous vehicles.

A third barrier is limitations in training data. For instance, large language models have made significant strides in English, but their extension to languages with far less online content from which to scrape training data will be challenging and face inherent limitations. Furthermore, biases in training data can harm some stakeholders, as evidenced by Google's and Amazon's AI recruiting tools being biased against women[2] and facial recognition systems not accurately recognizing Black people, partly because those systems had limited training samples from certain subpopulations. Similar challenges will have high-stakes consequences in DAF deployments in various communities, cultures, and environments.

Trustworthy AI depends on reliable human-AI interactions. Humans must be able to see an AI's prediction and assess its confidence in that prediction and

[1] Of course, this observation does not apply to the use of AI in safety-critical commercial systems such as industrial robotics or self-driving cars. Indeed, the T&E approaches and requirements for trustworthy components are similar to those faced by the DAF.

[2] J. Dastin, 2018, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women," *Reuters*, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

characterize the AI's basis for that prediction. Without interpretability and uncertainty quantification, trust in AI will remain limited.

Mechanisms for adapting to distribution drifts are essential to trustworthy AI. Furthermore, such mechanisms are necessary to account for shifting environmental conditions and imbalances in training data (e.g., having different fractions of samples for different subpopulations at test time than at training time).

AI components are often integrated into a larger system, so typical metrics used to assess an AI component's performance in isolation may inaccurately reflect its effect on the overall system performance.

Finally, trustworthy AI systems must be robust in the face of training- and inference-time attacks. Training-time attacks include data poisoning attacks and back doors, while inference-time attacks include making small changes to test samples to induce significant changes to the AI output—both white-box attacks that depend on knowledge of the AI model's inner workings and black-box attacks which pose risks even when details of the AI model are hidden.

DoD has recognized the need to improve the trustworthiness of AI. How AI will interact with the warfighter to improve trustworthiness is becoming a central concern as DoD seeks to adopt AI technologies. Human-AI interaction models, as they exist now, do not account for the dynamic and stressful situations in which warfighters find themselves. Thus, in its February 2022 memo, "Technology Vision for an Era of Competition," the OUSD(R&E) identified "Trusted AI and Autonomy" as one of 14 critical technologies areas and noted that, "[t]rusted AI with trusted autonomous systems are imperative to dominate future conflicts."

Furthermore, the June 2022 DoD report *U.S. DoD Responsible AI Strategy and Implementation Pathway* stated, "[t]o ensure that our citizens, warfighters, and leaders can trust the outputs of DoD AI capabilities, DoD must demonstrate that our military's steadfast commitment to lawful and ethical behavior applies when designing, developing, testing, procuring, deploying, and using Al."

The basic issue is whether a warfighter will trust their life to an AI-based system. These DoD concerns, combined with heightened public concern, have encouraged intensified research and development in trustworthy AI technologies. As a result, we can expect both near- and longer-term progress that will benefit AI-based DAF systems in general and DAF AI T&E specifically.

Finding 6-1: Existing approaches for designing trustworthy AI-enabled systems do not take into account the role of humans who interact with the AI-enabled systems.

Implications of Advances in Trustworthy AI to DAF T&E

While many challenges will undoubtedly persist into the foreseeable future, the continued focus from both private industry and the U.S. government will improve understanding of the theoretical foundations and will lead to the creation of more

trustworthy AI components. Furthermore, these advances will lend greater clarity to the range of appropriate AI applications and will extend that range by developing new ML approaches and improved architectures.

In short, trustworthy AI will enable higher quality and more dependable AI components. Additionally, advances expected in the next few years will permit AI adopters to insist that ML models have their uncertainty comprehensively measured or analytically bounded. Again, this will greatly benefit T&E activities. Ultimately, an ML component should be able to incorporate quantifiable uncertainty as part of its output. Thus, systems based on these components will already have a good test base from which to proceed to system-level testing, which will not only aid T&E but also help guarantee robust and resilient operation and promote user trust.

To reap the benefits of trustworthy AI components, the DAF must adopt new system engineering and T&E practices that explicitly incorporate requirements for trustworthy AI. The DAF will need acquisition approaches that recognize the state of the art in AI trustworthiness, placing realistic but aggressive requirements on AI components. These new practices must be codified in a set of standards and supported by appropriate tools and infrastructure. Developmental testbeds will be needed to explicitly measure AI robustness, resilience, safety, and other trustworthiness attributes. The Air Force will need T&E processes, canonical test datasets, and infrastructure to perform T&E of these higher-quality AI components, including the means to efficiently test performance against out-of-distribution, dynamic, and unexpected operational conditions. AI data generators will likely play a key role. In addition, adversarial T&E processes similar to those emerging in the cyber domain will be important to probe and redress vulnerabilities and deficiencies.

Recommendation 6-2: The Department of the Air Force should invest in developing and testing trustworthy artificial intelligence (AI)-enabled systems. Warfighters are trained to work with reliable hardware and software-based advanced weapon systems. Such trust and justified confidence must be developed with AI-enabled systems.

Trustworthy AI components will be enablers for safety critical systems such as weapon systems and semi or fully autonomous vehicles. However, not all trustworthy AI components must necessarily be fail-safe. As with other complex systems, some failure modes will be acceptable given the operational context of the model. The goal in engineering a trustworthy AI component will be to make its performance significantly more interpretable and predictable than the AI models currently available while tailoring it to its intended application. For example, recommender systems can be more tolerant of errors than autonomous control systems and thus have different AI trustworthiness requirements. There will be a

natural trade between the time and expense to build certain levels of trustworthiness and the intended application of the model.

Uncertainty quantification is an essential ingredient of AI-enabled DAF systems. Two examples of currently common paradigms include Bayesian estimation and conformal inference.[3] Bayesian statistics have a long and rich history, and methods in common use by the DAF, such as Kalman filtering, are grounded in Bayesian methodology. However, some priors, such as priors on weights of neural networks, can be difficult to interpret or validate. Conformal inference uses carefully selected quantiles of training data to quantify the uncertainty in predictions without any distributional assumptions on the data and minimal assumptions on ML algorithms. This framework has a high potential for facilitating T&E. A related challenge is communicating uncertainty measurements to human decision-makers. If the prediction is a scalar value (e.g., predicted amount of precipitation next week), then the DAF will have various excellent tools at its disposal. But when ML systems yield high-dimensional outputs, such as images, visualizing or communicating uncertainty is a persistent challenge.[4] However, the rate of change is such that by the time this report is publicly released, several new relevant examples will have been developed.

6.2 FOUNDATION MODELS

Foundation models (FMs)[5] are deep learning models that have emerged in the past 5 years, initially for language processing applications, where they are called large language models (LLMs) (exemplified in Figure 6-1). However, FMs have recently been applied to visual, multimodal, and multitask applications. These models are extremely large deep neural networks that use immense training sets, with some models exceeding 100 billion learning parameters. FMs employ self-supervised learning (SSL) where the model is presented with (x', x) pairs, where x' is an edited version of x with some of the constituents of x having been excised. The model is taught to predict the excised constituents and uses as a training signal its understanding of the full contents of each x to generate a loss function. Typical examples of edits for image and video-based SSL are coloring, rearranging the sections of an image or frames of a video, and other geometric transformations. One of the main advantages of SSL is that the costly process of labeling training data is avoided. This can greatly simplify data curation for both training and testing.

[3] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018, "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association* 113(523):1094–1111.

[4] A. Angelopoulos, S. Bates, J. Malik, and M.I. Jordan, 2020, "Uncertainty Sets for Image Classifiers Using Conformal Prediction," arXiv:2009.14193.

[5] M. Casey, 2023, "Foundation Models 101: A Guide with Essential FAQs," *Snorkel AI*, March 1, https://snorkel.ai/foundation-models.

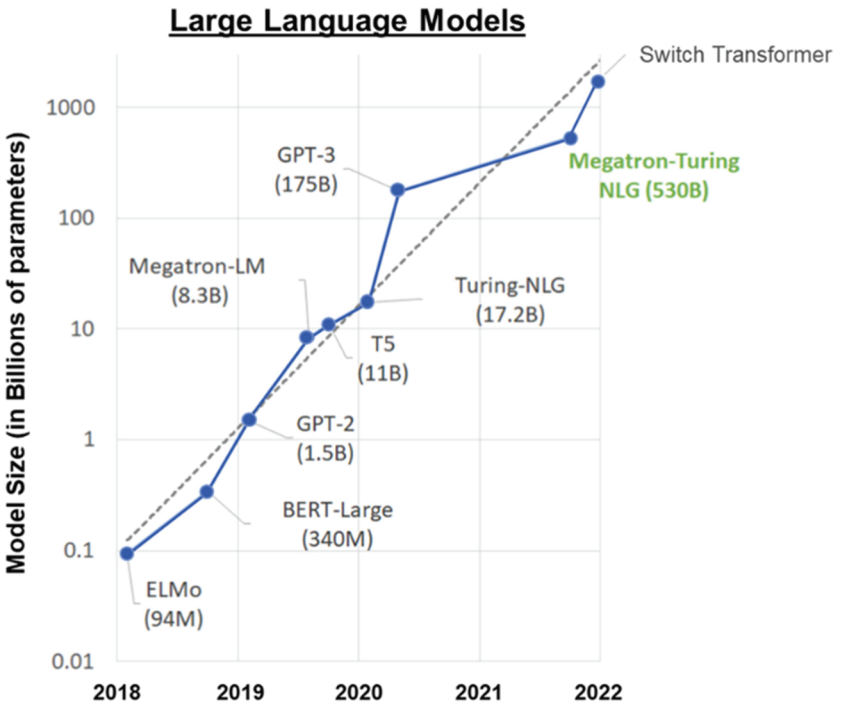

FIGURE 6-1 The growth in the size of Large Language Models (LLMs). Due to computational requirements, it is unlikely that the exponential rate shown can continue indefinitely, but if the trend plateaus near its current size, only a small set of organizations will be able to develop future LLMs. SOURCE: Courtesy of NVIDIA.

Today, FMs represent the state-of-art for natural language processing (NLP) tasks and consistently outperform the previous leaders, recurrent neural networks (RNNs), and long-short-time memory (LSTM) models.

Candidate applications for early Air Force adoption include language translation, communications denoising, language and speaker identification, human-machine interfaces that use DAF-specific nomenclature and idioms, recommender systems for training and intelligence analyses, and data summarization for intelligence reporting. As CV-based FMs become mainstream, the Air Force can leverage them in missions that involve large amounts of ISR data, where FMs will perform state-of-the-art detection, classification, ID, and tracking tasks. Ultimately, multi-modal and multi-tasking FMs will help fuse data from multiple sources and will help analysts, pilots, and commanders perform complex and time-critical tasks.

The Implication of Foundation Models for DAF T&E

The Air Force may elect to build its own FMs or procure pre-trained FMs and adapt them to Air Force applications. In either case, it must address the testing challenges that accompany these huge and complex models. For example, methods are still being developed to test FMs, and what methods exist are highly human-intensive. The allure of FMs is that while initial training and T&E require huge datasets and computing resources, once the base FM has been trained and tested, it can be readily adapted to a broad suite of downstream applications. As a result, the amount of adaptation development and T&E required for each application is less than would be required if the application were created from scratch without the FM as a base model. Moreover, as improvements are made to the base FM, these newer versions can be re-integrated readily into the applications, thereby efficiently propagating improvements across the entire suite.

Unfortunately, today's FMs are huge "black-box" components—literally 100s of billions of learning parameters—that lack transparency, explainability, and interpretability. T&E failure isolation can be a major challenge. For example, if the adapted FM has a failure mode, it may be unclear if the failure is due to the base FM, the adaptation, or the interaction between these two parts. Assigning accountability and correcting failures may be difficult, especially when the failures are due to the complex and subtle interplay between the components.

There are other issues, as well. For example, FMs will likely propagate their failure modes and biases to their adaptations. Thus, if the DAF has used a base FM for many applications, they may all exhibit the same base FM vulnerabilities. Furthermore, while FMs adaptations can perform extremely well, performance under transfer to new environments or continual learning in evolving environments can be suboptimal compared to dedicated models.

The DAF may consider using commercial FMs and adapting them to Air Force applications. For instance, an FM trained on images may lead to "off-the-shelf" image feature representations that could be used to train an Air Force EO image classifier. This framework is tantalizing in terms of the relatively fast development time and small computational resources required for training. However, the pre-trained FMs may also present significant security risks. In particular, commercial FMs are generally trained using massive collections of uncurated data scraped from the internet. This means that an adversary may post images or other data online to be scraped by the FM, which are explicitly designed to poison the FM for a particular task. For instance, an adversary might upload a series of images of jets designed to shift how images of jets are represented by FMs and affect downstream classifiers. Such attacks are almost impossible to detect, and accounting for this possibility is essential for accurate T&E.

Large FMs and data generators (discussed in the next section) will require massive computational resources for training, T&E. Therefore, the DAF must consider strategies for access to supercomputing class computers. One possibility is to partner with the national DOE laboratories, such as Sandia National Laboratory or Oak Ridge National Laboratory. Another possibility is to upgrade the DoD high-performance computing capability to handle these demanding AI workloads. Leasing capability from a major cloud provider should also be investigated. In any case, the solution must be readily accessible to both AI developers and T&E professionals and be able to protect data and AI software at multiple security levels.

Finding 6-2: Large language FMs exhibit a behavior termed "hallucination," where the model output is either non-sensical or is not consistent with the provided input or context. The effects of hallucination are task-dependent. There are no metrics to assess the impact of large FMs on the various downstream applications they have been applied to.

Finding 6-3: Several Large FMs are available for single modality, language being the most dominant one. DAF tasks may involve multi-modal sensing and inference. SSL-based Large Language Models are just recently becoming available for multi-modal paired or unpaired data.

6.3 INFORMED MACHINE LEARNING MODELS

Although foundation and other data-driven deep neural network (DNN) models have become the mainstay of machine learning applications, newer approaches to deep learning are emerging that seek to explicitly incorporate more application-domain knowledge into the learning process.[6] The committee refers to these approaches collectively as informed machine learning (IML).[7] IML models seek variously to incorporate knowledge in the form of algebraic equations, differential equations, simulation results, spatial invariances, logic rules, knowledge graphs, probabilistic relations, and human feedback into the learning process or the model architecture.

IML approaches increase model performance, generalizability across targeted domains, robustness, interpretability, and explainability. Fundamentally, IML models

[6] Conventional deep learning, of course, also integrates knowledge into its learning processes, through labeled data, feature engineering, and by exploiting invariances or equivariances (in convolutional neural networks, for example); but the IML techniques seek to integrate more knowledge and do so in a principled manner that does not depend on the data itself but, rather, on the domain whence the data derives.

[7] L.V. Rueden, S. Mayer, K. Beckh, et al., 2021, "Informed Machine Learning—A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," *IEEE Transactions on Knowledge and Data Engineering* 35(1):614–633.

aim to improve the utility and trustworthiness of deep learning. Compared to FMs and other deep learning models, IML models can be relatively small and trained using fewer data samples. Furthermore, when an IML model incorporates general laws or constraints (of physics or geometry, for example) it can offer an improved ability to handle non-stationary environments and to generalize better beyond the scope of its training set.

The DAF will find numerous uses for IML models, especially in physics-based applications such as radar, sonar, EO/IR processing, and where the models need to be embedded in size, weight, and power-constrained applications. IMLs will also apply to applied science research, such as the discovery of new materials for hypersonic systems or the assessment of aircraft design under various operational conditions.[8]

Implications of Informed Machine Learning Models for DAF T&E

IML models represent an emerging area in machine learning, with many applications and research directions. The DAF needs to assess the T&E needs of these models in the context of relevant applications. Notwithstanding the nascent nature of these models, it is likely that the principled incorporation of knowledge into machine learning will reduce and better characterize test space coverage for DAF applications; this will allow for more efficient testing at both the component and system levels. Also, these models may be more amenable to analytical verification processes based on, for example, the physical constraints programmed into the models. The reduced size of these models and their ability to leverage and focus on causally related environmental features will contribute to model explainability and interpretability, thereby facilitating failure analysis and improving trustworthiness.

There will also be challenges to overcome for effective T&E. IML models get their power from the integration of prior knowledge about the application domain. But this human-directed incorporation of knowledge may lead to unconscious biases or unintended limitations embedded in the models. Also, the development of IML models requires close collaboration between domain experts and machine learning experts. Thus, T&E teams and processes must be multidisciplinary to properly implement efficient test approaches and interpret test results. Furthermore, physics-based information incorporated into ML systems may be approximations of the true physics or may change over time or across instances. For example, an ML model may be trained for one radar sensor and work well in that context but yield poor results when used for a different sensor. Accounting for shifts in the physical knowledge between the training and testing phases is critical;

[8] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, et al., 2021, "Physics-Informed Machine Learning," *Nature Review Physics* 3:422–440, https://doi.org/10.1038/s42254-021-00314-5.

while methods such as model adaptation can help overcome this challenge, such considerations are a vital component of T&E for IML systems. Finally, adversarial robustness may manifest differently in IML systems than in their more generic counterparts. While the side information embedded into IML systems may help reduce opportunities for data poisoning, for instance, it may also mean that new methods are necessary for identifying, counteracting, or safeguarding against poisoning attacks.

6.4 AI-BASED DATA GENERATORS

AI-based data generation is an active and rapidly advancing area in machine learning research and development, with many novel AI techniques appearing in the past 10 years. In the visual domain, for example, generators range from generative adversarial networks (GANS)[9] to variational autoencoders,[10] autoregressive models,[11] normalizing flow techniques,[12] and denoising diffusion models.[13] Neural radiance field (NeRF) models have recently emerged that can generate multi-view 3D volumetric images from multiple 2D images. NeRFs are a type of informed machine learning (covered in the next section) that combine neural networks and traditional geometry-based rendering techniques. In the text domain, generators include transformer-based architectures such as GPT-3 and ChatGPT.[14]

Data generators can create realistic augmented reality and virtual reality simulations; they can fill in missing data, extrapolate or predict data based on existing datasets, and realistically (or otherwise) perturb existing datasets. In short, they can simulate an existing reality or can create fake but realistic variants of reality. This is demonstrated in Figure 6-2, which shows examples of photorealistic faces generated using a denoising diffusion model. They can create realistic images of all sorts and sizes that are often hard for humans to detect as fabrications. They can create photorealistic faces (and other objects or gestures) and can morph one face

[9] Google Machine Learning Education, 2022, "Generative Adversarial Network," updated July 18, https://developers.google.com/machine-learning/gan.

[10] J. Rocca and B. Rocca, 2019, "Understanding Variational Autoencoders (VAEs)," *Towards Data Science*, September 23, https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.

[11] Author, "Guide to Autoregressive Models," *Turing*, https://www.turing.com/kb/guide-to-auto regressive-models, accessed April 25, 2023.

[12] A. Omray, 2021, "Introduction to Normalizing Flows," *Towards Data Science*. https://towards datascience.com/introduction-to-normalizing-flows-d002af262a4b.

[13] J. Ho, A. Jain, and P. Abbeel, 2020, "Denoising Diffusion Probabilistic Models," University of California, Berkley.

[14] T. Brown, B. Mann, N. Ryder, et al., 2020, "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901.

FIGURE 6-2 Photorealistic faces generated using a denoising diffusion model. SOURCE: Courtesy of University of California, Berkeley.

(or object) into another face (or object) or into a different aspect view of the same face (or object). They can extend beyond images to create realistic videos and audio. FMs, discussed earlier, can be used as text data generators, creating realistic sentences, full paragraphs, and even essays that could plausibly come from intelligent (or otherwise) humans. By combining video generators and text generators, text-to-image and image-to-text transcriptions are possible. The DALL-E system is a modern example of the power of text-to-image generation, and ChatGPT is a modern example of text generation in response to prompts.

Potential DAF uses of AI-based data generators are extensive. These include creating training scenarios for combat games or pilot training, generating data for influence operations, and training machine learning algorithms and autonomous systems for operation in simulations of denied or contested environments. There are also numerous applications to ISR in data extrapolation, smoothing, or interpolation. Today, for example, an AI-based Global Synthetic Weather Radar (GSWR) system has been prototyped for the DAF. The GSWR uses AI data generation techniques that integrate multiple data sources to predict how weather radar returns would appear in regions where they are absent.

Finding 6-4: Physics-based and other knowledge-informed models have the potential to increase the robustness and computational efficiency of data-driven methods. These models can also help enforce physics or knowledge-based performance boundaries, which can increase the efficiency of the T&E process. However, to successfully deploy such models the DAF must ensure that the parameters and assumptions upon which they are based are actually present during operations, which requires additional attention to operational T&E.

Implications of AI-Based Data Generators to DAF T&E

Data generators will likely play a significant role in future DAF T&E activities. For example, they offer the capability to automate and accelerate the exploration of large test spaces using simulation; they can extrapolate from real data to generate unusual or special test datasets; they can be combined with live data and hardware-in-the-loop to support integration testing; and they can help evaluate concepts of operation and human-machine interactions.

However, the effective use of data generators for T&E will require rigorous T&E of the generators themselves. This need, in turn, calls for standardized evaluation and test metrics for these generators that probe their vulnerabilities and limitations. Indeed, generated data may appear valid when in fact, it is erroneous. For example, GANs are quite capable of generating "fake" images, but the fidelity of the fakes may be crucial in certain T&E activities, such as evaluating the robustness of an AI-based system in new environments. The quintessential question that T&E needs to answer is: does the generated data properly represent the important aspects of its domain and intended use? More specifically, generative models can be considered a tool for drawing samples from an estimate of the probability density underlying the training data, where that density is represented using a neural network. Generated images may look realistic, but testing procedures must ensure that all modes of distribution are accurately captured and that rare but mission-critical events or samples are not ignored by the generative model

The cost and effort to produce sufficiently realistic and useful data must be weighed against the cost and effort of other approaches, such as operational testing and analytical methods. Simulation testbeds that leverage data generators may be expensive to build initially, but their ability to test many situations rapidly could readily amortize the initial investment and lead to more cost-effective T&E overall.

Recommendation 6-3: The Department of the Air Force should assess the capabilities of data generators to enhance testing and evaluation in the context of relevant applications.

Data generators can exhibit significant biases. This phenomenon has been well-documented in the context of racial and gender biases, but in DAF settings, the

biases may be unpredictable, and the DAF lacks tools for detecting unanticipated biases. Furthermore, data generators typically are focused on generating "typical" samples from a distribution corresponding to the training distribution, whereas in some settings, the DAF may have a stronger interest in extreme or anomalous events. Therefore, DAF T&E must consider how data generators may affect our ability to understand systems in atypical operating conditions.

In addition, training state-of-the-art generative models can incur significant costs. For instance, it has been reported that training GPT-3 on 0.5 trillion words required $4.6M and generated 500 metric tons of carbon dioxide. Yet, it is unclear how this trend toward larger models will evolve, and it is also unknown what the scale of generative models for various DAF applications needs to be. Generative models for computer-vision applications such as image generation are generally smaller by orders of magnitude compared to GPT-3-scale generative language models, but training can nevertheless require substantial computing resources. Future applications that combine language and computer vision models (multi-modal generative models) will require substantial training and will undoubtedly also pose computational challenges. In summary, generative AI model development and training costs may affect the use of such models in varied USAF contexts and limit the DAF's ability to address problems with generative models uncovered in T&E.

6.5 AI GAMING FOR COMPLEX DECISION-MAKING

Recent AI gaming technology, such as Alpha Zero (Go, Chess, and Shogi) and Pluribus (poker), has demonstrated superhuman capabilities in extremely complex albeit constrained adversarial decision-making contexts. Reinforcement learning combined with deep learning is at the heart of these technologies. Typically, very large computational resources are required. These systems are often boot-strapped with labeled training sets and then further trained through self-play. Recent models (e.g., Alpha Zero) use self-play exclusively and require no labeled training data.

AI board game systems have developed strategies of play that surpass those that humans have developed across centuries of over-the-board play. AlphaStar is even more sophisticated, with the ability to play Starcraft II at the grandmaster level. StarCraft is more challenging than typical board games, as shown in Table 6-1. AI researchers continue to develop more sophisticated AI gaming and decision-making capabilities, aiming to achieve superhuman-level decision-making in demanding and realistic situations.

As AI gaming technology continues to increase in sophistication, it will be an important technology choice to augment complex decision-making in air force autonomous systems, robotics, command and control, logistics, planning, and scheduling applications. In most circumstances, human-AI teaming will be a crucial element of success (to include reinforcement learning with human feedback, or RLHF). In other circumstances, the gaming technology may need to operate

TABLE 6-1 A Comparison of the Challenges Presented by Games Such as Chess and Go Versus the Challenges Presented by Wargames Such as StarCraft

"Simple" Board Games (e.g., Chess and Go)	StarCraft-Like Environments
Huge state space of possible moves	Huge state space of possible moves
Fully observable	Partially observable
Single-player	Multiple agents and types of agents
Turn-taking	Simultaneous movement
Deterministic	Stochastic observations and effects
Few rules, some context coupling	Many rules, often context-dependent
Non–real time	Real time

completely autonomously for periods of time, such as for EW or cyber applications where superhuman response times are required. Coordination of multiple assets at a large scale is another example where AI gaming technology may excel. For example, research today on using deep reinforcement learning to coordinate multiple drones in real time[15] may translate to new swarm warfighting capabilities in the future. Based on recent successes as well as their future promise, the DAF should stay abreast of the latest advances in AI-enabled gaming technologies and explore how these capabilities might help enhance DAF missions. At the same time, the DAF AI T&E champion should ensure that such systems undergo the same type and level of T&E as any other AI-enabled weapon system.

Implications of AI Gaming for Complex Decision-Making to DAF T&E

Future advances in AI gaming and its foundational deep reinforcement learning (DRL) techniques will enable the Air Force to build systems that are more capable than ever before and that involve AI in more sophisticated and complex ways. This increased system complexity will mean more challenges facing T&E. Also, the teaming relationship between the human and AI elements will likely be much more interrelated and complex. Thus, many tests will need to assess human-AI interactions and overall teaming effectiveness and will require more intricate user participation. This is typical of operational testing today, but the key point is that the T&E process will need to engage the user continuously, from the early stages of development to the operation of the system. Indeed, one important way to address this challenge is for the Air Force to adopt the agile and continuous

[15] A.T. Azar, A. Koubaa, N. Mohamed, et al., 2021, "Drone Deep Reinforcement Learning: A Review," Electronics 10(9):999, https://doi.org/10.3390/electronics10090999.

testing approaches that are currently being used by commercial industry for its complex AI-based systems.

In the cases where the AI agent is acting autonomously or is generating a set of complex decisions that exceed human capability, the systems can fail in non-intuitive and potentially catastrophic ways. Thus, the Air Force should require that explainability and interpretability be key engineering goals not just for individual AI components but for the entire system. Additional fail-safes and data-logging capabilities will need to be built into such systems. Safeguarding systems that provide appropriate performance guarantees can help narrow the test space of the overall system. Lessons can be learned from the private industry efforts to build autonomous automobiles, where continual testing, ghost AI hosting, early user involvement, human-AI teaming, and other techniques are being pioneered.

Notably, researchers are making important advances in safety-critical reinforcement learning. For example, control barrier functions have been shown to provide control-theoretic guarantees for obstacle avoidance. Hamilton-Jacobi Reachability[16] provides an exact formulation of the states that may lead to failure and can be used to formulate optimal safety control policies. While these techniques have difficulties generalizing and scaling, recent machine learning approaches are emerging that use such techniques offline to learn approximate but effective safety control policies. Computationally efficient, approximate but guaranteed safety-critical control is then applied online.[17] This is a very active area of research, and the committee expects continued progress that will aid in DAF autonomous systems T&E.

> Finding 6-5: Recent and anticipated advances in AI gaming technologies will enable the Air Force to build systems that are more capable than ever before and that involve AI in more sophisticated ways, but this increased system complexity will make the teaming relationship between the human and AI elements much more interrelated and complex, thereby placing additional challenges on effective T&E.

AI Foundations

In addition to the core area discussed above, important research is progressing in foundational and theoretical AI research. A foundational understanding of AI is akin to investments in a foundational understanding of medicine, biology, chemistry, and materials science. The DAF must have strong pillars on which to build, test, and evaluate AI systems. Testing and evaluation of AI-enabled systems

[16] S. Herbert, University of California, San Diego, "The Safe Autonomous Systems Lab," http://sylviaherbert.com/hamilton-jacobi-reachability-analysis, accessed April 27, 2023.

require understanding the implicit biases and generalization properties of learned models, and when all potential operational scenarios cannot be tested explicitly, theory can provide invaluable insights. For instance, many modern neural networks are "overparameterized," so the number of parameters learned during training far outstrips the amount of available training data. In these settings, we can often interpolate the training data, and the architecture of the neural network determines the nature of the learned interpolator; theory may provide insights into the nature of the interpolator as a function of architecture. Interpretable machine learning, which is essential to our ability to debug faulty systems, is a further foundational research challenge. Several empirical studies have shown interpretability may come at the expense of accuracy, but there is no evidence that this is a fundamental or insurmountable challenge. Foundations are essential to understanding how a learned model will perform under new operating conditions or how a model trained in one setting will perform in a shifted environment. Theory can also inform trustworthiness assessment through the development of new metrics. Privacy and stability guarantees, important safeguards in trustworthy AI, depend on a cadre of theoretical tools. Model compression is also ripe for theoretical advances and important to air force deployment or continual learning settings with limited power. Finally, theory is essential to the development of new tools for uncertainty quantification without assumptions on the distribution underlying data or properties of the learning algorithms or models.

7

Concluding Thoughts

The committee was tasked with the following questions:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force (DAF) and in commercial industry.
2. Consider examples of artificial intelligence (AI) corruption under operational conditions and against malicious cyberattacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

What was uncovered through investigating these questions was a significant overlap between the answers. For example, examples of AI corruption (question 2) were a common topic of promising areas of science and technology (question 3) and reflected as an issue in the current testing and assessment methods employed by the DAF (question 1).

In structuring this report, the committee organized chapters such that each question was primarily addressed in a specific chapter in the report, and thus the recommendations were primarily reported in those sections. However, each chapter of the report contains relevant findings and recommendations for each of the questions addressed by the committee, so in practice, it was not possible to isolate the questions to individual chapters.

Task 1, "Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry," is primarily answered in Chapters 3 and 4. In these chapters, the current testing and

assessment methods found by the committee are described and referenced, and a comparison to best practices in commercial industry are directly covered. However, to fully understand this contrast, findings and recommendations from Chapters 2 and 5 must be also considered.

Task 2, "Consider examples of AI corruption under operational conditions and against malicious cyberattacks," is primarily addressed in Chapter 5. However, the topic of AI corruption is a primary challenge throughout all of the test and evaluation (T&E) of AI-enabled systems and thus is mentioned throughout the study, especially in Chapter 6.

Task 3, "Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption," is primarily addressed in Chapter 6. However, AI corruption is under active S&T research and in the spirit of DevSecOps and AIOps, solutions are being deployed as rapidly as they are discovered. Thus, many of the approaches in Chapters 3, 4, and 5 also support findings and recommendations to mitigate AI corruption.

In short, the complexity, interconnection, and coupling of issues throughout T&E with AI-enabled systems will require a reassessment of all T&E policies, processes, and procedures to assure that validation and verification of all systems, not just the AI components, will support the necessities of a dynamic and risky deployment and operational environment. In this case, the committee concluded that while the questions appeared quite straightforward on an initial reading, in fact the DAF has now caught the AI tiger by its tail. Taming that tiger will be challenging, especially as AI-enabled components become commonplace in all platforms and MDAPs, but it is not at all an insurmountable problem. It requires vision, hands-on leadership, prioritization, and a shared commitment to an AI-enabled future DAF.

Appendixes

A

Statement of Task

The National Academies of Sciences, Engineering, and Medicine will establish an ad hoc committee to (1) plan and convene a multi-day workshop and (2) conduct a consensus study to examine the Air Force Test Center's technical capabilities and capacity to conduct rigorous and objective test, evaluation, and assessments of artificial intelligence (AI)-enabled systems under operational conditions and against realistic threats. Specifically, the committee will:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyberattacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

The committee will provide workshop proceedings—in brief and in a report summarizing the results from the consensus study.

B

Public Meeting Agendas

APRIL 22, 2022
KICK-OFF MEETING DAY 1

Executive Session

3:00–3:05 PM	Welcome and Introductions
3:05–4:00 PM	Bias and Conflict of Interest Discussion Scott Weidman, Deputy Executive Director, Division on Engineering and Physical Sciences

Open Session

4:00–5:00 PM	Sponsor Remarks Major General Evan Dertien, Commander, Air Force Test Center Colonel Keith Roessig, Vice Commander, Air Force Test Center Air Force Test Center Technical Experts
5:00 PM	Adjourn

APRIL 25, 2022
KICK-OFF MEETING DAY 2

Executive Session

11:00 AM–1:00 PM	Bias and Conflict of Interest Discussion

JUNE 27, 2022
DATA-GATHERING WORKSHOP, DAY 1

Room 208
The Keck Center, 500 Fifth Street, NW
Washington, DC 20001

11:00–11:15 AM	Workshop Welcome and Introductions
11:15 AM–12:00 PM	Speaker: Mr. Jacob Martinez, Technical Director, 47th Cyberspace Test Squadron, United States Air Force
12:00–12:45 PM	Speaker: Mr. David Coppler, Technical Director, 46 Test Squadron, United States Air Force
12:45–1:15 PM	Lunch Break
1:15–2:00 PM	Speaker: Mr. Marshall Kendrick, 45th Test Squadron, United States Air Force
2:00–2:45 PM	Speaker: Dr. Jane Pinelis, Chief, AI Assurance, Office of the Department of Defense Chief Digital and Artificial Intelligence Officer (CDAO)
2:45–3:15 PM	Discussion and Day 1 Wrap-Up
3:15 PM	Adjourn

JUNE 28, 2022
DATA-GATHERING WORKSHOP, DAY 2

Room 208
The Keck Center, 500 Fifth Street, NW
Washington, DC 20001

11:00–11:15 AM	Welcome and Recap of Day 1
11:15 AM–12:00 PM	Speaker: Dr. Chad Bieber, Director, T&E Operations, Joint AI Center, Johns Hopkins University Applied Physics Laboratory
12:00–12:30 PM	Discussion: Workshop Day 1 and Future Committee Meetings
12:30–1:15 PM	Lunch Break
1:15–2:00 PM	Speaker: Olivia Brown, Technical Staff, AI Technology Group, MIT Lincoln Laboratory
2:00–2:45 PM	Speaker: Dr. Michael Wellman, Richard H. Orenstein Division Chair of Computer Science and Engineering and Lynn A. Conway Collegiate Professor of Computer Science and Engineering, University of Michigan
2:45–3:00 PM	Break
3:00–3:45 PM	Speaker: Dr. Thomas Strat, President and CEO, DZYNE Technologies
3:45–4:30 PM	Speaker: Jim Bellingham, Executive Director, Johns Hopkins Institute for Assured Autonomy
4:30–4:45 PM	Break
4:45–5:30 PM	Speaker: Dr. Matt Turek, Deputy Director Information Innovation Office

5:30–6:15 PM	Speaker: Dr. Nancy Cooke, Professor and Graduate Program Chair in Human Systems Engineering, Polytechnic School, Arizona State University
6:15–6:30 PM	Discussion and Day 2 Wrap-Up
6:30 PM	Adjourn

JUNE 28, 2022
DATA-GATHERING WORKSHOP, DAY 3

Room 208
The Keck Center, 500 Fifth Street, NW
Washington, DC 20001

11:00–11:05 AM	Welcome and Recap of Day 2
11:05–11:50 AM	Speaker: Dr. Bin Yu (NAS), Chancellor's Professor in the Departments of Statistics and Electrical Engineering & Computer Sciences, University of California, Berkeley
11:50 AM–12:35 PM	Speaker: Dr. Nathan VanHoudnos, Senior Machine Learning Research Scientist, Software Engineering Institute
12:35–1:15 PM	Lunch Break
1:15–2:00 PM	Speaker: Dr. Bruce Draper, Program Manager, Defense Advanced Research Projects Agency
2:00–2:45 PM	Speaker: Mr. Ed Zelnio, Principal Research Physicist, Air Force Research Laboratory
2:45–3:00 PM	Break
3:00–3:45 PM	Speaker: Dr. Eileen Bjorkman, Executive Director, Air Force Test Center
3:45–4:30 PM	Workshop Wrap-Up and Discussion
4:30 PM	Adjourn

AUGUST 23, 2022
DATA-GATHERING MEETING #2

11:00–11:10 AM	Welcome, Introductions, and Quick Updates
11:10 AM–12:10 PM	Speaker: LCDR Joseph W. Geeseman, Program Manager, Smart Sensor, CDAO Algorithmic Warfare Division
12:10–12:45 PM	Lunch Break
12:45–1:45 PM	Speaker: Dr. Lori Westerkamp, Air Force Research Laboratory Sensors Directorate
1:45–2:00 PM	Break
2:00–3:00 PM	Speaker: Dr. John Richards, Sandia National Laboratories
3:00–4:30 PM	Committee Discussion: Draft Report Outline, September Meeting, Site Visits, Writing Assignments, Fall Meeting Schedule
4:30 PM	Adjourn

SEPTEMBER 28, 2022
DATA-GATHERING MEETING #3, DAY 1

Open Session

11:00 AM–12:00 PM	Speaker: Prof. Yolanda Gil, Senior Director for Major Strategic AI and Data Science Initiatives, Information Sciences Institute, University of Southern California
12:00–12:30 PM	Lunch Break
12:30–1:30 PM	Speaker: Professor Jeff Schneider, Research Professor, Carnegie Mellon University, Founding Member—Uber Advanced Technologies Group

1:30–2:30 PM	Speaker: Dr. Mitch Crosswait, Director of Operational Test and Evaluation Deputy Director of Net-Centric, Space, and Missile Defense Systems
2:30–2:45 PM	Break

Closed Session

2:45–4:30 PM	Committee Writing Session/Discussion
4:30 PM	Adjourn

SEPTEMBER 29, 2022
DATA-GATHERING MEETING #3, DAY 2

Open Session

11:00 AM–12:00 PM	Speaker: Mr. Eric Nelson, CTO of Software, Morse Corp.
12:00–12:30 PM	Lunch Break
12:30–1:30 PM	Speaker: Mr. Neil Serebryany, Founder and CEO, CalypsoAI
1:30–1:45 PM	Break

Closed Session

1:45–4:00 PM	Committee Writing Session/Discussion
4:00 PM	Adjourn

NOVEMBER 30, 2022
DATA-GATHERING MEETING #4

1:00–2:00 PM	Speaker: Dr. Riccardo Mariani, VP, Industry Safety, NVIDIA Italy
2:00–3:00 PM	Committee Discussion, Planning, and Writing Session

3:00–3:45 PM	Speaker: Col. Tucker "Cinco" Hamilton, Chief of AI Test and Operations, Department of the Air Force Chief Data and AI Office & Commander, 96th Operations Group, Eglin Air Force Base
3:45–5:00 PM	Committee Discussion, Planning, and Writing Session (continued)
5:00 PM	Adjourn

DECEMBER 6, 2022
DATA-GATHERING MEETING #5, DAY 1

1:00–1:45 PM	Committee Discussion: Finalizing Writing Assignments, January Writing Meeting Planning, Discussion of the Outline and Draft Sections
1:45–2:45 PM	Speaker: Dr. Nicholas Carlini, Research Scientist, Google Brain
2:45–3:00 PM	Break
3:00–4:00 PM	Speakers: Alex Kotran, Co-Founder and CEO, AI Education Project & Michael Kanaan, Chief of Staff of the U.S. Air Force Fellow at Harvard Kennedy School, AI Education Project Board Member
4:00 PM	Adjourn

DECEMBER 7, 2022
DATA-GATHERING MEETING #5, DAY 2

12:30–1:30 PM	Speaker: Michael Cox, Vice President and Chief Architect, NVIDIA
1:30–1:45 PM	Break
1:45–2:30 PM	Committee Discussion, Planning, and Writing Session

C

Committee Member Biographical Information

May Casterline, *Co-Chair*, is an image scientist and software developer with a background in satellite and airborne imaging systems. Her research interests include deep learning, hyperspectral and multispectral imaging, innovative applications of machine learning (ML) approaches to remote sensing data, multimodal data fusion, data workflow design, high-performance computing applications, and creative software solutions to challenging geospatial problems. She holds a PhD and a BS in imaging science from the Rochester Institute of Technology, with a focus on remote sensing. In industry, Dr. Casterline has acted as a product owner, technical lead, lead developer, and image scientist on both research initiatives and development projects. She joined the NVIDIA federal solution architecture team in 2017 focusing on deep learning and artificial intelligence (AI) applications.

Thomas A. Longstaff, *Co-Chair*, is the chief technology officer (CTO) of the Software Engineering Institute (SEI) at Carnegie Mellon University. As CTO, Dr. Longstaff is responsible for formulating a technical strategy and leading the funded research program of the institute based on current and predicted future trends in technology, government, and industry. Before joining the SEI as CTO in 2018, he was a program manager and principal cybersecurity strategist for the Asymmetric Operations Sector of the Johns Hopkins University Applied Physics Laboratory (JHU APL), where he led projects on behalf of the U.S. government, including nuclear command and control, automated incident response, technology transition of cyber research and development (R&D), information assurance, intelligence, and global information networks. Dr. Longstaff also was the chair of

the Computer Science, Cybersecurity, and Information Systems Engineering Programs and the co-chair of Data Science in the Whiting School at JHU. His academic publications span topics such as malware analysis, information survivability, insider threat, intruder modeling, and intrusion detection. He maintains an active role in the information assurance community and regularly advises organizations on the future of network threat and information assurance. Dr. Longstaff is an editor for *Computers and Security*. He has previously served as the associate editor for *IEEE Security and Privacy* and general chair for the New Security Paradigms Workshop and Homeland Security Technology Conference and numerous other program and advisory committees. Prior to joining the staff at JHU APL, Dr. Longstaff was the deputy director for technology for the CERT Division at the SEI. In his 15-year tenure at the SEI CERT Division, he helped create many of the projects and centers that made the program an internationally recognized network security organization. His work included assisting the Department of Homeland Security and other agencies to use response and vulnerability data to define and direct a research and operations program in analysis and prediction of network security and cyber terrorism events. Dr. Longstaff received a bachelor's degree in physics and mathematics from Boston University and a master's degree in applied science and a PhD in computer science from the University of California, Davis.

Craig R. Baker is the president of Baker Development Group, LLC, a consulting, leadership, and teaching company. He is a trusted executive leader widely known as a strategic planner and executor of large, important, highly visible projects and products to mitigate risks. Mr. Baker retired from the U.S. Air Force as a Brigadier General in July 2021. He graduated from the U.S. Military Academy at West Point in 1992, was a command combat test pilot, and instructed at the U.S. Air Force Fighter Weapons School ("Top Gun"). Mr. Baker commanded at both the squadron and wing levels. Additionally, he was the technical director GM/test program manager of the 59th Test and Evaluation Squadron achieving $500 million in savings cultivated and lives saved by creating and establishing an innovative process and software program that required 60 percent fewer assets and personnel; and met 60 percent of worldwide objectives in developing and integrating a congressionally directed fleet capability while delivering a historical first milestone highlighted to the enterprise president weekly. Mr. Baker led multi-service weapons assessment teams into Iraq and Afghanistan after OIF and OEF, which resulted in revolutionary software program and new weapon developments. He earned two MS degrees in strategic intelligence and strategic studies.

Robert A. Bond, prior to his appointment as principal staff, served for 5 years as the CTO at the Massachusetts Institute of Technology Lincoln Laboratory (MIT LL). He was formerly the associate head of the Intelligence, Surveillance,

Reconnaissance and Tactical Systems Division. In his 42-year career, Mr. Bond has led research initiatives in very-large-scale integrated (VLSI) circuits, software technology, parallel processors, adaptive and nonlinear signal processing, AI, C2ISR systems, and big-data analytics. He joined the MIT LL in 1987 and led the software and integration activities for the Radar Surveillance Technology Experimental Radar. In the 1990s, Mr. Bond conducted seminal studies on the use of massively parallel processors (MPP) for real-time signal processing. He then pioneered the development of a custom VLSI processor and a 1,000-node MPP for radar space-time adaptive processing. Mr. Bond led the development of a middleware technology for portable and scalable parallel signal processors that evolved into the Parallel Vector Tile Optimized Library (PVTOL), which won an R&D 100 award. In 2003, he received the MIT LL's prestigious Technical Excellence Award, for his "technical vision and leadership in the application of high-performance embedded processing architectures to real-time digital signal processing systems." Since 2015, Mr. Bond has led the MIT LL's strategic initiatives in AI and autonomous systems. As the CTO, he oversaw and funded the applied research portfolios in these areas. In 2018, Mr. Bond founded the Recent Advances in AI for National Security workshop. He is currently the MIT LL's program manager for the Air Force-MIT AI Accelerator program. Mr. Bond has a BS (honors) in physics, is a member of the Association for the Advancement of Artificial Intelligence (AAAI), and a senior member of the Institute of Electrical and Electronics Engineers (IEEE).

Rama Chellappa is a Bloomberg Distinguished Professor in the Departments of Electrical and Computer Engineering (Whiting School of Engineering) and Biomedical Engineering (School of Medicine) at JHU. At JHU, he is also affiliated with the Center for Imaging Sciences, the Center for Language and Speech Processing, the Institute for Assured Autonomy, and the Mathematical Institute for Data Science. Dr. Chellappa holds a non-tenure position as a College Park Professor in the Electrical and Computer Engineering (ECE) department at the University of Maryland (UMD). From 1981 to 1991, he was an assistant and associate professor in the Department of EE-Systems at University of Southern California. He received an MSEE (1978) and a PhD (1981) in electrical engineering from Purdue University. His current research interests span many areas in computer vision, pattern recognition, AI, and ML. Dr. Chellappa is an elected member of the National Academy of Engineering (NAE). He received the 2023 Distinguished Career Award from the Washington Academy of Sciences, the 2020 Jack S. Kilby Medal for Signal Processing from the IEEE, and the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). Additionally, Dr. Chellappa is a recipient of the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society, the Technical Achievement and Meritorious Service

Awards from the IEEE Computer Society, and the inaugural Leadership Award from the IEEE Biometrics Council. At UMD, he received numerous college- and university-level recognitions for research, teaching, innovation, and mentoring of undergraduate students. He was recognized as an Outstanding ECE by Purdue University and as a Distinguished Alumni by the Indian Institute of Science. Dr. Chellappa is a Golden Core Member of the IEEE Computer Society and has served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the president of IEEE Biometrics Council. He is a fellow of AAAI, the American Association for the Advancement of Science (AAAS), the Association for Computing Machinery (ACM), the American Institute for Medical and Biological Engineering, IAPR, IEEE, the Optical Society of America, and the National Academy of Inventors and holds nine patents.

Melvin Greer is an Intel Fellow and the chief data scientist at the Americas, Intel Corporation. He is responsible for building Intel's data science platform through graph analytics, ML, and cognitive computing to accelerate transformation of data into a strategic asset for public sector and commercial enterprises. His systems and software engineering experience has resulted in patented inventions in cloud computing, synthetic biology, and Internet of Things bio-sensors for edge analytics. Dr. Greer functions as a principal investigator in advanced research studies, including nanotechnology, additive manufacturing, and gamification. He significantly advances the body of knowledge in basic research and critical, highly advanced engineering, and scientific disciplines. Dr. Greer is a member of the AAAS and serves on the board of directors for the National Academies of Sciences, Engineering, and Medicine. He has been appointed to senior advisor and fellow at the Federal Bureau of Investigation (FBI) IT and Data Division and is charged with acceleration of the FBI mission by supporting appropriate data collection, data analytics, discovery, and visualization via advanced data science and AI techniques. Dr. Greer is one of the 2018 LinkedIn Top 10 Voices in data science and analytics. He also received the Washington Exec Inaugural Pinnacle Award as the 2018 Artificial Intelligence Executive of the Year, and received the 2017 Black Data Processing Associates Lifetime Achievement Award and the 2012 Black Engineer of the Year Awards Technologist of the Year Award, which recognized his outstanding technical contributions that have had a material impact and high value to society as a whole. Dr. Greer has been appointed a fellow of the National Cybersecurity Institute where he assists government, industry, military, and academic sectors on meeting the challenges in cyber security policy, technology and education. He is professor for the MS of science in data science program at Southern Methodist University and adjunct faculty, advanced academic program at JHU, where he teaches the MS course on practical applications of AI. In addition to his professional and investment roles, Dr. Greer is the founder and managing director of the

Greer Institute for Leadership and Innovation, focused on research and deployment of a 21st-century leadership model. He is a frequent speaker at conferences and universities and is an accomplished author; his fifth book, *Practical Cloud Security A Cross Industry View*, is his most recently published book. Dr. Greer is a board of director member at the National GEM Consortium where he oversees and aligns its strategic direction, educational policy, finances, and operations with the mission of the fellowship program. As a popular educator and board member at a number of Historical Black Colleges and Universities, Dr. Greer is leading science, technology, mathematics, and engineering research initiatives, directly trying to shape a more diverse generation of up-and-coming technical talent.

Tamara G. Kolda is an independent mathematical consultant under the auspices of her company MathSci.ai based in California. She is also a distinguished visiting professor in the Department of Industrial Engineering and Management Science at Northwestern University. From 1999 to 2021, Dr. Kolda was a researcher at Sandia National Laboratories in Livermore, California. She specializes in mathematical algorithms and computation methods for tensor decompositions, tensor eigenvalues, graph algorithms, randomized algorithms, ML, network science, numerical optimization, and distributed and parallel computing. Dr. Kolda is a member of the NAE, a fellow of the Society for Industrial and Applied Mathematics (SIAM), and a fellow of the ACM. She holds a PhD in applied mathematics from UMD.

Robin R. Murphy is the Raytheon Professor of Computer Science and Engineering at Texas A&M University and a director of the Center for Robot-Assisted Search and Rescue. Her research focuses on AI, robotics, and human–robot interaction for emergency management. Dr. Murphy has deployed ground, aerial, and marine robots to over 30 disasters in five countries including the 9/11 World Trade Center, Fukushima, Hurricane Harvey, and the Surfside collapse. She is an AAAS, ACM, and IEEE fellow, a TED speaker, and her contributions to robotics have been recognized with the ACM Eugene L. Lawler Award for Humanitarian Contributions and a United States Air Force Exemplary Civilian Service Award medal. She holds a PhD (1992) and an MS (1989) in computer science and a BME (1980) in mechanical engineering from the Georgia Institute of Technology.

David S. Rosenblum is the Planning Research Corporation Professor and the chair of the Department of Computer Science at George Mason University. Since receiving his PhD from Stanford University, Dr. Rosenblum has held positions as a member of the technical staff at AT&T Bell Laboratories (Murray Hill); associate professor and associate chair at the University of California, Irvine; CTO and principal architect at PreCache, Inc.; professor of software systems at University College London; and Provost's Chair Professor, dean of the School of Computing,

and founding director of the NUS-Singtel Cyber Security R&D Lab at National University of Singapore. He has made significant contributions to a broad array of research problem areas in computer science, including software engineering, distributed systems, ubiquitous computing, and ML. Among his most highly cited research are works on Internet-scale publish/subscribe computing; assertion processing techniques and regression testing methods for software engineering; and ML and deep learning techniques for recommendation systems, activity recognition, and social media analytics. Dr. Rosenblum is a fellow of the ACM and IEEE and has received two 10-year, test-of-time awards, including the International Conference on Software Engineering (ICSE) 2002 Most Influential Paper Award for his ICSE 1992 paper on assertion processing, and the inaugural 2008 ACM SIGSOFT Impact Paper Award for his Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering 1997 paper on Internet-scale event observation and notification (co-authored with Alexander L. Wolf). He also received the 2018 ACM SIGSOFT Distinguished Service Award.

John (Jack) N.T. Shanahan retired from the United States Air Force in 2020 after a 36-year military career. In his final assignment, he served as the inaugural director of the Department of Defense (DoD) Joint Artificial Center. Mr. Shanahan served in a variety of operational and staff positions in various fields, including flying, intelligence, policy, and command and control. He established and led DoD's Pathfinder AI fielding program (Project Maven) and is an adjunct senior fellow with the Technology and National Security Program at the Center for a New American Security. Mr. Shanahan is a member of the IEEE Standards Association Autonomous Weapons Systems Assurance and Safety Subcommittee.

Rebecca Willett is a professor of statistics and computer science at The University of Chicago. Her research is focused on ML, signal processing, and large-scale data science. Dr. Willett received the National Science Foundation (NSF) CAREER Award in 2007, was a member of the Defense Advanced Research Projects Agency Computer Science Study Group, received an Air Force Office of Scientific Research Young Investigator Program award in 2010, was named a fellow of the SIAM in 2021, and was named a fellow of the IEEE in 2022. She is a co-principal investigator and member of the executive committee for the Institute for the Foundations of Data Science, helps direct the Air Force Research Laboratory University Center of Excellence on Machine Learning, and currently leads The University of Chicago's AI+Science Initiative. In addition, Dr. Willett serves on advisory committees for the NSF's Institute for Mathematical and Statistical Innovation, the AI for Science Committee for the Department of Energy's Advanced Scientific Computing Research program, the Sandia National Laboratories Computing and Information Sciences Program, and the University of Tokyo Institute for AI and Beyond.

She completed her PhD in electrical and computer engineering at Rice University (2005) and was an assistant and then tenured associate professor of electrical and computer engineering at Duke University (2005–2013). Additionally, Dr. Willett was an associate professor of electrical and computer engineering, the Harvey D. Spangler Faculty Scholar, and a fellow of the Wisconsin Institutes for Discovery at the University of Wisconsin–Madison (2013–2018).

D

Acronyms and Abbreviations

AAIT	Autonomy and Artificial Intelligence Test
ABMS	Advanced Battle Management System
ACC	Air Combat Command
ACE	Air Combat Evolution
ADAX	Autonomy, Data, and AI Experimentation
AEDC	Arnold Engineering Development Complex
AEIS	Artificial Intelligence Enabled Systems
AF DCGS	Air Force Distributed Common Ground Station
AFB	Air Force Base
AFIT	Air Force Institute of Technology
AFLCMC	Air Force Life Cycle Management Center
AFMC	Air Force Materiel Command
AFOTEC	Air Force Operational Test and Evaluation Center
AFPC	Air Force Personnel Center
AFRL	Air Force Research Laboratory
AFSC	Air Force Specialty Code
AFTC	Air Force Test Center
AI	artificial intelligence
AI4NSL	Artificial Intelligence for National Security Leaders
AIA	AI Accelerator
AICI	Artificial Intelligence Criticality Indicator
AIOps	Artificial Intelligence for IT Operations

API	Application Programming Interface
APIGEE	Automated Pipeline for Imagery Geospatial Enhancement and Enrichment
ASIL	Automotive Safety Integrity Level
ATR	automatic target recognition
AWCFT	Algorithmic Warfare Cross-Functional Team
AWS	Amazon Web Services
BPA	bulk purchase agreement
C2	Command and Control
C2BMC	Command, Control, Battle Management, Communications
C4I	Command, Control, Communications, Computers, and Intelligence
C4ISR	Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance
CAC	common access card
CBA	capabilities-based analysis
CCA	Collaborative Combat Aircraft
CCF	common cause failure
CDAO	Chief Digital and AI Office (OSD); Chief Data and AI Office (DAF)
CDC	Capability Development Council
CET	continuing education and training
CEU	continuing education unit
CI/CD	continuous integration/continuous delivery
CIP	critical intelligence parameter
CLSA	Computer Language Self-Assessment
CMCC	Common Mission Control Center
CNN	Convolutional Neural Network
CoI	community of interest
CONOPS	concept of operations
CR	certifiable robustness
CSAF	Chief of Staff of the Air Force
CTP	critical technical parameter
CV	computer vision
CXO	chief experience officer
DAF	Department of Air Force
DARPA	Defense Advanced Research Projects Agency

DASD(DT&E)	Deputy Assistant Secretary of Defense for Developmental Test and Evaluation
DAU	Defense Acquisition University
DDB	Dynamic Database
DevOps	development operations
DevSecOps	development, security, and operations
DGS	Defense Geospatial Services
DIE	Defense Intelligence Enterprise
DL	deep learning
DMP	data management pipeline
DNN	deep neural networks
DoD	Department of Defense
DOT&E	Director of Operational Test and Evaluation
DOTmLPF-P	Doctrine, Organization, Training, materiel, Leadership, Personnel, Facilities, Policy
DRL	deep reinforcement learning
DT	developmental test
DT&E	developmental test and evaluation
DTO	Digital Transformation Office
ELINT	electronic intelligence
EO	electro-optical
EOB	Electronic Order of Battle
EW	electronic warfare
EWIR	electronic warfare integrated reprogramming
FAA	Federal Aviation Administration
FERET	Face Recognition Technology
FFRDC	federally funded research and development center
FM	foundation model
FMV	full-motion video
FOC	fully operational capability
FOT&E	follow-on operational T&E
FPGA	field programmable gate array
GAO	Government Accountability Office
GeoINT	geospatial intelligence
GOTS	government off-the-shelf
GPU	graphics processing unit
GSWR	Global Synthetic Weather Radar

HAI	human-AI
HARA	Hazard and Risk Assessment
HMC	human-machine cognitive collaboration
HMT	human-machine team
HRL	human readiness level
HSI	human-systems integration
HW	hardware
IDA	Institute for Defense Analyses
IEC	International Electrotechnical Commission
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IML	informed machine learning
IOC	initial operating capability
IOT&E	initial operational test and evaluation
IP	intellectual property
IR	infrared
ISO	International Organization for Standardization
ISR	intelligence, surveillance, and reconnaissance
IT	information technology
JAIC	Joint Artificial Intelligence Center
JATIC	Joint Artificial Intelligence Test Infrastructure Capability
JCIDS	Joint Capabilities Integration and Development System
JEON	joint emergent operational need
JHU APL	Johns Hopkins University Applied Physics Laboratory
JROC	Joint Requirements Oversight Council
JSE	joint simulation environment
JUON	joint urgent operational need
KPI	key performance indicator
KPP	key performance parameter
KSA	key system attribute
LAWS	lethal autonomous weapon system
LFT&E	live-fire test and evaluation
LL	Lincoln Laboratory
LLM	large language model
LOR	level of rigor
LSTM	long-short-time memory
LVC	live-virtual-constructive

M&S	modeling and simulation
MAGE	Machine Assisted GeoINT Exploitation
MAIS	major automated information system
MDA	Missile Defense Agency
MDAP	major defense acquisition program
MIT	Massachusetts Institute of Technology
ML	machine learning
MLOps	machine learning operations
MRTFB	Major Range and Test Facility Base
MSTAR	Moving and Stationary Target Acquisition and Recognition
MTA	middle tier of acquisition
MVP	minimal viable product
NAIIO	National AI Initiative Office
NAITIC	National Artificial Intelligence Test and Evaluation Infrastructure Capability
NDAA	National Defense Authorization Act
NeRF	neural radiance field
NFS	NextGen Federal Systems
NGAD	next generation air dominance
NIST	National Institute of Standards and Technology
NLP	natural language processing
NPS	Naval Postgraduate School
NR-KPP	net ready key performance parameter
NSCAI	National Security Commission on AI
O&M	operations and maintenance
OCR	operational change request
ODD	operational design domain
ODNI	Office of the Director of National Intelligence
OFP-CTF	Operational Flight Program-Combined Test Force
OOD	out-of-distribution
OSD	Office of the Secretary of Defense
OT	operational test
OT&E	operational test and evaluation
OTA	Operational Test Agency
OUSD(R&E)	Office of the Under Secretary of Defense (Research and Engineering)
PaaS	platform-as-a-service
PED	processing, exploitation, and dissemination
PEO	program executive officer

PII	personally identifiable information
PMO	Program Management Office
QU	quantifiable uncertainty
R&D	research and development
R&E	research and engineering
RADIUS	Research and Development for Image Understanding Systems
RAI	responsible AI
RAIDEN	Robust AI Development Environment
RDT&E	research, development, test, and evaluation
RFI	request for information
RFP	request for proposal
RL	reinforcement learning
RLHF	reinforcement learning with human feedback
RMF	Risk Management Framework
RNN	recurrent neural network
SAR	synthetic aperture radar
SCSP	Special Competitive Studies Project
SDN	software-defined networking
SDPE	strategic development planning and experimentation
SEI	special experience identifier
SME	subject-matter expert
SPO	system program office
SWAP	size, weight, and power
T&E	test and evaluation
TEMP	T&E Master Plan
TEVV	Test, Evaluation, Verification, and Validation
TL	transfer learning
TPS	Test Pilot School
TQD	training-quality data
TRL	technology readiness level
TRMC	Test Resource Management Center
TW	Test Wing
UARC	University-Affiliated Research Center
UAS	unmanned aerial system
UGV	uncrewed ground vehicle
UI/UX	user interface/user experience

UON	urgent operational need
USAFWC	United States Air Force Warfare Center
USC	University of Southern California
USDI	Under Secretary for Defense for Intelligence
V&V	verification and validation
VENOM	Viper Experimentation and Next-Gen Operations Model
VISTA	Variable In-Flight Simulator Aircraft
VOC	Visual Object Classes
VTTC	Virtual Test and Training Center
ZT	zero trust

E

Testing, Evaluating, and Assessing Artificial Intelligence–Enabled Systems Under Operational Conditions for the Department of the Air Force: Proceedings of a Workshop—in Brief

NATIONAL ACADEMIES Sciences Engineering Medicine

Proceedings of a Workshop—in Brief

Testing, Evaluating, and Assessing Artificial Intelligence–Enabled Systems Under Operational Conditions for the Department of the Air Force

Proceedings of a Workshop—in Brief

On June 28–30, 2022, the National Academies of Sciences, Engineering, and Medicine's Air Force Studies Board (AFSB) convened a hybrid workshop in support of its consensus study on testing, evaluating, and assessing artificial intelligence (AI)–enabled systems under operational conditions. The goals of the study are as follows:

1. Evaluate and contrast current testing and assessment methods employed by the Department of the Air Force and in commercial industry.
2. Consider examples of AI corruption under operational conditions and against malicious cyber-attacks.
3. Recommend promising areas of science and technology that may lead to improved detection and mitigation of AI corruption.

The information summarized in this Proceedings of a Workshop—in Brief reflects the opinions of individual workshop participants. It should not be viewed as a consensus of the workshop's participants, the AFSB, or the National Academies. The workshop planning committee heard from a wide range of experts from government, industry, and academia to help inform them about the Air Force Test Center's (AFTC's) ability to test, evaluate, and assess AI-enabled systems. The purpose of this workshop was to hear about how the U.S. Air Force (USAF) currently approaches AI testing and evaluation (T&E), industry approaches to testing AI, and challenges to AI testing. Exploration into other topic areas from the statement of task will be done in future data-gathering meetings by the workshop planning committee.

47TH CYBER TEST SQUADRON OVERVIEW

The first speaker was Jacob Martinez (47th Cyberspace Test Squadron [CTS]). Martinez began by giving a brief overview of the 47th CTS, which is part of the AFTC, and its two primary mission areas: providing test environments for hardware and software type cloud environments and conducting cybersecurity and resiliency activities for the Air Force's kinetic and non-kinetic weapons. In essence, the 47th CTS looks at not only the physical capabilities but also software capabilities. Martinez also noted that the 47th CTS is primarily a "fee for service" organization. He explained that the squadron relies on normal, agile, and continuous methods of T&E, with the intent to focus on continuous T&E in the future. He also stated that the 47th CTS is primarily a developmental testing (DT) organization.

The discussion shifted toward Unified Platform (UP), a project that aims to integrate cyber capabilities, systems, infrastructure, and data analytics while allowing cyber operators to conduct numerous tasks across the full spectrum of cyber operations. It is also one of the five elements of the Joint Cyber Warfare Architecture. The 47th CTS worked on this project and looked at several vendors to help support the application of AI/machine learning (ML) in UP. It determined that investments to begin integrating AI/ML into UP are estimated to be anywhere from $75,000 to $255,000 per year in licensing costs alone.

Thomas A. Longstaff (Software Engineering Institute; workshop planning committee co-chair) was curious if, within UP especially, Martinez's group is focusing more on the tools and techniques within UP or on what is within the development, security, and operations (DevSecOps) chain on the testing side from the software factory. Martinez responded that the 47th CTS is tied into the DevSecOps pipeline process. Discussion then ensued about ownership and responsibility. Martinez stated that, ultimately, the end user is the one who assumes the risk and takes responsibility. Rama Chellappa (Johns Hopkins University; workshop planning committee member) asked for an explanation of how they currently recruit people who can be a step ahead and fully understand the implications of the system design, AI, and so on. Martinez responded that industry is paying individuals, with that level of expertise, more than what he can provide. Instead of using high salaries to entice talent, he suggested using the PALACE Acquire (PAQ) Internship Program.[1] Martinez stated, "by embracing and offering training positions and PAQ internship positions, we not only get the latest training from academia, but we also can hold those individuals for 2 or 3 years and invest in them, in their education, and they invest in us by providing us new techniques and capability." This policy is not official, but an idea proposed by Martinez, he clarified. Longstaff asked a final question regarding applying resilience testing to things that may have adaptive behavior. Martinez responded that the 47th CTS does have a mission in which they do cyber resilience testing. The point was also made that resiliency testing will, in Martinez's opinion, probably become integrated with future AI/ML requirements as they may develop. The only issue is that acquiring and funding technological concepts takes a long time. Martinez has usually seen, within the Department of Defense (DoD), a 2- to 3-year gap between the time it takes for a concept to be accepted, funded, and explored.

[1] The PAQ Internship Program is a paid, full-time, 2- to 3-year USAF program for graduates interested in a number of disciplines. More information can be found at the AFCS website, at https://afciviliancareers.com/recentgraduates.

46TH TEST SQUADRON "KILL CHAIN DEVELOPMENTAL TEST"

Dave Coppler (46th Test Squadron [TS]) talked to the workshop planning committee about the 46th TS, a subordinate to the AFTC's 96th Test Wing, and the importance of DT. He noted the squadron's importance in considering all stages of the "kill cycle," also using the term "kill chain DT." He gave an overview of the organization's chain of command and mission statements. One of the squadron's primary focuses is on the testing of kill chain–relevant systems.

Coppler transitioned to talking about DT and why it is essential. He stated that DT is necessary government work that helps to accelerate acquisition by leveraging unique expertise, facilities, equipment, and capabilities. The 46th TS supports the entire system life cycle to ensure that upgraded systems do not break any of the system's initial capabilities. They can also provide upgrades to the software with new capabilities and ensure that they work properly. They also provide highly qualified experts with proper clearances to engage customers on any level and provide the necessary support. Coppler also discussed the importance of the test environment that the 46th TS provides for DT.

Coppler fielded questions from the workshop planning committee. Longstaff asked if, within the simulated emulated systems that the 46th TS is already using, it is considering incorporating more AI systems behavior into its simulated systems (e.g., the F-15E, etc.). Coppler responded that until the F-15Es, F-22s, and F-35s start incorporating AI into their platforms, the TS has no desire to do that. Longstaff followed up by asking if the TS is thinking about doing any automated AI-based behavior within the hardware for testing. Coppler stated

that he thinks that is way off in the future. The TS is still in the very early stages of developing the art of the possible. Chellappa asked about annotation and who does it. Coppler responded that the 46th TS does provide the truth data for physical things in real time, but it is not involved when the AI, for example, takes a deeper look at how data are being generated and used.

AI DT FOR COMMAND AND CONTROL

The next speaker, Marshall Kendrick (Air Operations Center Combined Test Force), opened by saying that the 45th TS is just getting started in the AI business. He then talked about the different efforts that the 45th TS is undertaking, many of which are in the big data/algorithm stage. In the future, he noted that most of the efforts have the potential to move into full AI/ML capabilities. Last, Kendrick posed two questions that his organization has been tracking for the past few years: how to test AI and how to use AI to test and test better.

Kendrick talked about some of the squadron-level flight programs his organization is involved in, such as Air Ops Command and Control (C2), a space flight that uses DoD's Kobayashi Maru C2 program, and other programs. He also discussed the need for real-time data processing, as everything is constantly changing (potential threats, environments, etc.). AI can assist in this effort, particularly with the Advanced Battle Management System (ABMS) and the Joint All-Domain Command and Control (JADC2) vision. Kendrick then talked about ongoing efforts within the 45th TS. Lt. Gen. (Ret.) John N. Shanahan (USAF; workshop planning committee member) asked if the 45th TS would play a role in helping to develop some of the C2 capabilities that the Air Force is working on. Kendrick responded that they could absolutely play a role, particularly on the software side. Kendrick and Shanahan also discussed the operator's role throughout the test process, the need to identify risk, and who accepts the risk.

Kendrick also discussed other ongoing efforts, such as cloud-based C2. He explained that this effort comes from the Air Force's Rapid Capabilities Office as part of the ABMS work. They have already built the test data sets and are working directly with developers. Kendrick mentioned that he has people meeting with the developers to ensure that the 45th TS understands the developers' test methodology and their data. The goal is to see how they can extend the test data and ensure that it covers all of the operational boundaries.

AI ASSURANCE

Jane Pinelis (Office of the DoD Chief Digital and Artificial Intelligence Officer) led the final presentation of the first day. She opened by defining AI assurance. She described AI assurance as the combination of T&E and responsible AI. She explained that the AI assurance process provides arguments and evidence to establish trustworthiness and justified confidence. She defined the goal as providing stakeholders with "justified confidence" that DoD AI-enabled systems meet requirements and support missions through ethical action. Stakeholders include the warfighter, commander, program manager, regulators, taxpayers, and others. She also talked extensively about the existing partnerships that the Chief Digital and Artificial Intelligence Office (CDAO) has and the different support it provides to these stakeholders.

Pinelis moved on to talk about the AI T&E process. The first step, algorithmic testing, is when reserved test data are used against a vendor's model in a laboratory environment. Next, the model tests four areas: integrity testing, confidence assessment, robustness, and resilience. Integrity testing shows the model's effectiveness using metrics such as the number of false positives, F1 score, precision recall, and other data points. Pinelis also talked about a new method called calibrating "model competency," where someone uses trained data on a specific data set deployed in an operational environment. She noted the importance of the model competency step in assessing "domain adaptation," or the model's ability to perform in different operational environments at the same level as observed in the bench testing environment. Confidence assessment calculates the distance between a data point and everything on which the model has previously been trained. Pinelis mentioned that this type of test helps with things such as label prioritization. She then talked about robustness—specifically, natural perturbations—and how they are transitioning a tool from the Test

Research Management Center that will help identify edge cases in a test set. Resilience was the final test, where they specifically focused on adversarial action and whether it comes through adversarial AI or cyber. It also measures the data set's ability to diagnose and recover from those attacks.

The second step is system integration, Pinelis said. This measures how well a model performs when plugged into a legacy system not intended to interact with AI. The key things that the CDAO looks for are functionality, reliability, interoperability, compatibility, and security.

The third step Pinelis described is human–system integration (HSI). This step involves inserting a human in the loop—that is, when a model is mounted to a platform and works. They tied the observe, orient, decide, act loop to DoD AI ethical principles to describe the HSI framework. She emphasized that human interactions with machines need to be maximally informative.

The final step is an operational test. Pinelis described this as both the easiest and the most challenging step. It is the toughest because, in her opinion, operationally testing AI-enabled systems, particularly autonomous ones, is very difficult. It is also the easiest because the CDAO always gets to collaborate with somebody when doing it. She then stated that the theory and methods behind operational testing are extraordinarily well developed and established. With AI, things have changed slightly. Tactical testing is an important part of the culture shift that avoids doing one big test at the end of the process and instead focuses on doing smaller but more frequent tests in multiple contexts and environments. There is also a push to evaluate the quality of decision making as performance. This attempts to evaluate the quality of decision making. The final point focused on the idea that one cannot test for everything and that test culture needs to shift to becoming more risk accepting rather than risk averse.

Pinelis noted that the CDAO was working with the Office of the Secretary of Defense Director, Operational Test and Evaluation, on various AI T&E products that would be available throughout DoD (to include T&E best practices, cloud-native test harnesses, a T&E bulk purchasing agreement, T&E tools, test products, and an AI Red Teaming handbook, among others). She ended her presentation by briefly discussing the different challenges of T&E and responsible AI. Longstaff asked how industry best practices would interact with a newly established T&E factory.[2] Pinelis responded that they would absolutely continue to get industry's tools and host them in the factory. She also stated that they try to keep the CDAO's tools available to industry for items they build for the CDAO, but they do not share the test data. However, some tools are ones that the CDAO does not want widely advertised, for national security purposes. Discussion took place about how there are lessons to be learned from the private sector's safety community for using AI in safety systems. Chellappa asked about domain adaptation and how Pinelis's group will tackle it. Pinelis responded that they will do their best to train the system with the data that they have but that a lot of emphasis should be placed on learning after the system is fielded. She also talked about privacy and how data transformation and governance can be significant in keeping data useful while ensuring that identity is not recoverable. Last, Shanahan asked about the cultural shift between the traditional developmental testing/operational testing and how that is coming along. Pinelis responded that integrative testing had been discussed for a long time but had not yet been implemented. Shanahan also touched on an AI mishap database and whether any thought had been put into that. Pinelis affirmed that they had thought about that and are establishing a database for responsible AI that will be a repository not just for incidents but also for tools and data.

DAY 1: WORKSHOP PLANNING COMMITTEE DISCUSSION

May Casterline (NVIDIA; workshop planning committee co-chair) raised a go-back question to Kendrick on whether or not the testing rigor that Pinelis described in her presentation was captured in their requirements. Kendrick responded that he has assessed whether rigor

[2] A T&E factory is a broad set of tools to empower non-experts in DoD to test a model when it arrives as a black box (i.e., when the model's inner workings are difficult to understand). K. Foy, 2022, "Graph Exploitation Symposium Emphasizes Responsible Artificial Intelligence," Massachusetts Institute of Technology, https://www.ll.mit.edu/news/graph-exploitation-symposium-emphasizes-responsible-artificial-intelligence.

has been properly addressed, but that his assessment would probably not be the same as what Pinelis described in her presentation. Kendrick pointed out that it is difficult for a fee-for-service organization to solve a problem when they need a contract before hiring, tasking, building, and testing are available to address the problem. Shanahan observed that a philosophical question needs answering at the Air Force level, writ large, on establishing "who owns what part" of this difficulty and looking into the requirements process. Another point was that some of the language used, such as F1, F2 scores, ROC curves, and false positives, is new for many people involved with Air Force T&E. He noted that this is not a typical T&E discussion. He followed on by saying that it sounds like the Air Force would like these terms to become part of the T&E discussion, but wondered how the Air Force builds toward that.

Chellappa and Shanahan discussed how someone would know if a new AI system is performing much better than what is already out there. This thought was a central question for some in figuring out "what is good enough?"—something that is still unresolved. Coppler commented that during his time on active duty with the 53rd wing, they would test "good enough" by measuring against what they already had. Chad Bieber (CDAO) agreed with Coppler and added that there are many ways to be good. Coppler jumped back in and posited that if an AI/ML algorithm does not perform as expected when tested, it may be doing something better than one ever thought possible. Longstaff resonated with that point and brought up his concern that sticking with the old regime of "testing to requirements," may result in the discarding of systems that yield surprisingly better results.

A final discussion ensued regarding the testing of large systems. Longstaff used JADC2 as an example—once one starts incorporating more AI capabilities, the nature of the entire system changes. How does one test that and begin to think about what to do to test an integrated system of that size and scale—an integrated system incorporating behavior and change based on how an adversary changes?

DAY 2: MORNING DISCUSSION

The workshop planning committee opened the day with a recap and discussion of the previous day. There was discussion regarding unknowns, such as the lack of ownership regarding liability and requirements. One workshop planning committee member commented on a contrast in approaches between the CDAO's office and the test squadrons. Shanahan commented that at the end of the day, the Air Force has to come in at an Air Force level and decide the best way forward—the test center versus the warfare center—regarding roles and responsibilities. Tamara G. Kolda (MathSci.ai; workshop planning committee member) asked if there was a way to audit decisions and collect them as AI systems deploy. Coppler responded that without hooks in the AI algorithms, the test community has no idea how to look into those algorithms and understand what they are doing. Kolda asked if the inputs and outputs of an AI system are logged. Coppler responded that they were. Bieber added that it is not always a given that one can check the inputs and outputs of an AI in a box. AI might be a larger component of the software, and it has a fundamental problem: it cannot be instrumented after the fact because that might change how the software operates.

PRACTICAL GUIDE TO AI TESTING

Bieber spoke briefly about his background as a tester and his previous work at the Joint Artificial Intelligence Center.

Initially, Bieber spoke about metrics and metric development. When developing metrics, one needs to understand what metrics developers are using, understand how program management has defined requirements, and understand how to measure operational success—that is, the importance of soliciting the end user's assessments of operational performance. He also talked about standards and how everyone makes their own tools and products. Unfortunately, this does not allow much in the way of interchangeability. Bieber also talked about tools and the CDAO's work establishing a T&E software factory, as well as its vision for developing a "suitcase" test kit, which would allow AI T&E in situ. Bieber explained that such a capability would not only be invaluable in assessing competency

(domain adaptation) but would likely also lead to the ability to "tune results" under operational conditions. He then touched on modeling and simulation (M&S). He stated that M&S is vital to AI T&E. He talked specifically about the common worry, or complaint, regarding the exploding state of the AI space. Bieber does not think of that as the biggest problem. AI's unique problem is that it does not understand the performance across that space well enough to predict behavior between two points, much less outside the area it tests.

Bieber then presented a scenario regarding a dog-finding uncrewed aerial vehicle (UAV) used by emergency services. Within this scenario, he talked about different metrics and their uses, such as mean average precision (mAP), average precision, Recall, and f-scores. Longstaff asked if Bieber could contrast mAP to accuracy. Bieber responded that precision, in the computer vision world, has a smaller, less overloaded definition than accuracy. Casterline and Chellappa discussed the applicability of some of the metrics that Bieber mentioned and commented that they are very computer-vision-centric. Chellappa stressed the need to understand the metrics and think more about what would work for AI-based systems. The workshop planning committee also discussed the idea of an algorithm deployed in the field that continuously learns during deployment. Bieber mentioned SmartSensor,[3] which does have the ability to retrain rapidly. Trevor Darrell (University of California, Berkeley; workshop planning committee member) asked for Bieber's thoughts on the idea of merging the culture of testing and development. He also asked for thoughts regarding identifying specific entity labels and not just a broad category, such as identifying a T-72 versus a tank. Bieber responded that he had seen the opposite problem, where they have tried to use computer vision to detect too far down the ontological hierarchy. Bieber also stressed the need for continuous testing. He stated, "We have to have the ability, if we're doing continuous development, to do testing at the same speed as the development process." He also spoke about competency testing and the different ways one can do it.

The discussion then shifted to autonomous vehicles and testing metrics, such as the number of user interrupts, to measure operational performance. Darrell spoke about how coming up with a commonsense tool that could look at and summarize a performance dump[4] could be helpful. He also spoke about how it would be valuable to require some disclosure and the ability to benchmark against open systems. Bieber followed on and spoke about the challenges of a black-box system and being unable to look inside it. Although he did say that while it would be useful to have full access to everything, the financial cost of having full access may not be feasible. Darrell suggested model cards and documentation confidentiality as middle-ground solutions. Bieber stated that the CDAO requires model cards. In closing, Kolda and Bieber engaged in discussion regarding data sequestration. Kolda also asked about model learning and whether the models that Bieber's group receives are already trained. Bieber responded that once the algorithm was delivered and deployed, it did not change.

ROBUST AND RESILIENT AI

Olivia Brown (Massachusetts Institute of Technology [MIT] Lincoln Laboratory [LL]) spoke about how AI systems have great promise for DoD. However, they are demonstrably brittle and often vulnerable to different forms of data corruption, Brown said. She specifically named post-sensor digital perturbations as a form of corruption. Brown explained that there are sources of natural and adversarial forms of vulnerability. A natural source could be when an AI model trains on upright chairs. When tipped over, the model could suffer a significant performance drop. Adversarial forms of vulnerability could involve deliberately manipulating an image's pixels, causing the model to fail in correctly classifying inputs, according to Brown.

[3] Smart Sensor is a CDAO project delivering an on-platform, AI-enabled autonomy package that allows a UAV to conduct automated surveillance and reconnaissance functions in contested environments. Satnews, 2022, "DoD CDAO Partners with USAF to Conduct Developmental Test Flight of AI and Autonomy-Enabled Unmanned Aerial Vehicle," Satnews, https://news.satnews.com/2022/06/23/dod-cdao-partners-with-usaf-to-conduct-developmental-test-flight-of-ai-and-autonomy-enabled-unmanned-aerial-vehicle.

[4] A performance dump of the system is a collection of data from a service processor after a failure of the system, an external reset of the system, or a manual request. IBM, 2021, "Initiating the Performance Dump," https://www.ibm.com/docs/en/power9/0009-ESS?topic=menus-initiating-performance-dump.

Brown then talked about the current way that machine models train. First, they undergo a design phase, where training data are collected and validated. The model then tests on a test data set similar to the one on which it trained. The system then deploys. She noted that they often observed that performance of the deployed system in the operational domain was much worse than predicted during the test phase. This degraded performance reduces operator and user trust and results in the system going offline, reoptimization, and ultimately redesign, Brown noted.

Brown stated that the path to creating a more robust system starts at the opposite side of the development process. Talking to operators at the beginning of the system's design phase is essential. In this way, the developer understands the operational environment into which the system will deploy. This awareness allows the programmers to consider potential sources of variation in the data that the system is likely to encounter. Next, the developer should establish a testing process that avoids experimenting against a test set similar to the system's training. Instead, one should test against that training data's perturbations or that training distribution. Last, Brown advised training the model to perform better against perturbed data. Brown then spoke about the work at MIT LL in robust AI research that addresses new ways to tackle natural and adversarial sources of vulnerabilities. Brown highlighted tools like HydraZen[5] and the Responsible AI Toolbox (rAI-toolbox),[6] which will help Brown's team at MIT continue its research on evaluating AI robustness. The workshop planning committee conversation then shifted toward different use cases that utilized these tools. Brown concluded by describing MIT LL's next steps in supporting the development of robust and responsible AI.

Longstaff asked if there was a way to specify a requirement that would allow them to test against the requirement once the robustness training was complete. He also asked how well the robustness pipeline works with non-vision-oriented AI. Brown responded that she does not necessarily have an answer, but that setting the requirements is very important. She responded that MIT LL was exploring ways to use simulators and to figure out how to integrate those into the training process. Regarding the second question, Brown stated that MIT LL is moving beyond natural images and looking at radar and time series. Longstaff followed up and asked about data augmentation strategies. Brown responded that these strategies exist to train against a single type of perturbation, but you will (normally) have a suite of them.

[5] See the Hydra-Zen site, at https://github.com/mit-ll-responsible-ai/hydra-zen.
[6] See the Responsible AI Toolbox site, at https://github.com/mit-ll-responsible-ai/responsible-ai-toolbox.

AI TRUST AND TRANSPARENCY

Michael Wellman (University of Michigan) opened his presentation with a brief discussion of his past work. He started with how trust and transparency are nothing new for AI. Trust in an AI system is ill-defined because people have different ways of defining it, Wellman said. Moreover, trust goes beyond AI systems—it applies to any software system or system that generates recommendations, information, or decisions. To Wellman, however, trust is not a necessary condition to use a system. Many instances exist where people use technology without understanding its full consequences, Wellman said.

Wellman then discussed an example of autonomous AI–stock trading. In certain instances, companies have employed AI to control large trading accounts that act autonomously in financial markets. Indeed, inserting a human in the loop is not feasible. By the time a human can do anything, the opportunity evaporates. He cited a company, Knight Capital, where a software configuration error led to a loss of around $400 million that took the company down. Nevertheless, even with that kind of outcome, people did not stop trusting or using this technology, Wellman said.

Wellman then discussed transparency in AI systems. Specifically, he spoke about the common approach, called the explanation approach, used to interrogate the underlying model so that one can explain the decision or recommendation it produces. However, this approach has some dangers—mainly that it is easy to come up with an explanation that seems plausible and could be the reason

for an underlying decision, but that might not necessarily have a causal connection. Wellman then presented a different approach—to limit oneself to models that are interpretable in the first place. In other words, the model has a certain simplicity or structure that one can discern directly—the explanation that the model deduces is causally related to an actual decision or recommendation. He maintained, however, that it is not always possible to do this.

Wellman then introduced strategic domains. This approach considers decisions in worlds where the outcome depends on other agents' actions. The finance and trading example discussed earlier is one example. He mentioned cybersecurity as a strategic domain because an attack or defense is always relative to the other party's actions. Negotiation, monitoring, war gaming, or anything in conflict is also considered a strategic domain. Strategic domains present a transparency challenge—the decisions made in a strategic situation often require unpredictability. So, something like debugging is more challenging. Wellman concluded that from a designer's perspective, it requires extra care to preserve transparency.

Wellman ended his presentation and opened up the discussion. Chris Serrano (HRL Laboratories) pointed out that, while we may not have a theorem on whether an attacker or a defender of a system will win, there is undoubtedly an idea of how the cost grows when defending a system versus attacking one. Wellman added that cost also determines who wins in the end. Longstaff and Wellman discussed counterfactuals and how to utilize them in dealing with the issue of inferring intent. Wellman explained that using counterfactual queries could infer intent—in this instance, identifying whether someone is a scammer. Shanahan asked a question regarding Wellman's statement on trust not being a prerequisite for adoption. He asked if we are getting too detailed or "cute" with some of our existing systems, particularly given how basic their capabilities are right now. Wellman said he framed his stock trading example as a cautionary tale to show that it may not be possible to stop a system without full trust or confidence because it will be compelling. Sometimes that is worth embracing. However, there is always going to be a matter of measured risk. Chellappa said that there are four things to look at: domain adaptation, adversity of attacks, bias, and privacy. Wellman responded with adversarial approaches in black-box situations. He said that the risk with domain adaptation is that things get deployed in situations for which they were not designed.

EVALUATION OF AI

Thomas Strat (DZYNE Technologies) started by going through the background of his company and presented some case studies about the types of AI work in which DZYNE Technologies is currently involved.

DZYNE Technologies is a small company that designs, builds, and operates autonomous aircraft, Strat said. These aircraft can range anywhere from small 6-pound aircraft to the largest UAVs that the military currently operates. The company also has an AI group of around 25 individuals who help to deploy AI capabilities on their aircraft, Strat said.

Strat then discussed semantic labeling from satellite imagery. Specifically, he talked about determining which algorithm is better, given two different image classifications. A qualitative approach to answering the question is useful because it allows you to look at the data from a visual perspective and ask if they represent your intuition. A quantitative approach allows one to use several metrics to determine accuracy. Strat pointed out that while there are many commonly used metrics, there is not one single obvious best metric to use in any given situation. He then stated that it is seldom clear what metrics to use from the outset, which depends on many factors. Strat also pointed out that to do an evaluation, you must have some form of ground truth to compare it to, and ground truth is not always complete and correct. The quality of that ground truth makes an important difference, he said. Overall, some key challenges concern trade-offs between the evaluation metrics that someone chooses and the quality of ground truth making a big difference. Strat stated that some solutions to help with these challenges include having multiple metrics

and pretraining a model without annotation. Longstaff asked if any metrics explain the quality of ground truth. Strat responded that he was not sure, but during his time as a Defense Advanced Research Projects Agency (DARPA) program manager, they played around with that. Chellappa stated that there are some models of label noise, but it is hard to figure out how good the ground truth is.

Strat spoke about another case study regarding the area of building extraction. This case study aimed to highlight all of the buildings in an image and use brightness to determine building height. Strat pointed out that as one gets toward the perimeter of the image, off-axis pixels increase. He posed the question, "How do you evaluate the accuracy of these data sets?" You do not have ground truth that covers city-size areas with any accuracy, according to Strat.

Additionally, said Strat, any algorithm's accuracy will not be uniform across something the size of an entire city. Cities are not uniform and have many factors that could affect the algorithm. For example, he stated that the heights of trees in a certain area could affect the ability to extract data properly.

Strat then shifted the discussion toward autonomous vehicles. First, he considered how to evaluate progress, and that speed may not necessarily be the way to measure that. He then talked about the DARPA Grand Challenge. This challenge aimed to put autonomous vehicles to the test in a real-world environment—in particular, an operationally relevant environment such as a desert. Strat said that he favors attempting system-level tests in operational environments whenever possible, as he believes that there is nothing more convincing than doing that. He then covered autonomous aircraft—specifically, a long-endurance air platform (LEAP). LEAP has been in operation since 2016 in the Middle East in numerous combat operations. First, it was evaluated and tested using simulation and takeoff tests at military bases. LEAP then moved on to formal operational assessments in theater in the hands of the military, where it has been continually reassessed since its first use. At this point, Strat said, it has amassed more than 50,000 hours of operational use by the military in the Middle East. Strat then talked about the mishaps. Most of them have been mechanical, some were owing to hostile action, and a number were attributed to operator error or the human in the loop. According to Strat, zero mishaps were attributed to the AI error. Instead, when the operators did not trust the AI, problems occurred, such as intervening in the aircraft's landing approach. Last, Strat presented a final video showing off ROBOPilot, an autonomous system that can fly an airplane. Over the span of a few years, the system was developed and trained by Strat's team to fly an airplane with no human in the cockpit. Strat then spoke about the potential application of this robotic technology for the military.

Chellappa asked for Strat's assessment of the efficacy of simulations. Strat stated that the answer to whether it is useful for AI algorithms is complicated, but why would it not be? The more veracity the simulation has, the less reason there is to doubt its efficacy for training or evaluating AI algorithms. Longstaff asked about ROBOPilot and how it compares to the full auto function that a 787 Max has. Strat responded that there is a market for autonomous flight. He specifically mentioned aircraft that may have been deemed unsafe for human flight. Strat also said that fly by wire is the way to go and that he would not necessarily put ROBOPilot up against a fully integrated autopilot system. He ended his talk by briefly discussing how interfacing with a human being is one of the most difficult challenges for AI because the human brain is so complex. It is much easier to interface with physics than it is with humans. As such, the ROBOPilot program is a much easier challenge to solve than what DARPA set out to do with the ALIAS program.

ASSURANCE: THE ROAD TO AUTONOMY

Jim Bellingham (Johns Hopkins Institute for Assured Autonomy) discussed his background in marine robotics and autonomous marine vehicles. He also talked extensively about the application of AI and autonomy in many vehicles that he had helped develop and utilize. In this talk, he also referenced several tools.

Bellingham explained that autonomy is everywhere: finance, logistics, military systems, the medical environment, and others. The big problem is assurance. He explained that for industry, it is a trade-off between assurance and ensurance. He also said that assurance is a key to accelerating AI and autonomy.

Bellingham wrapped up his presentation by stating that robotics and autonomy will transform society. He reviewed some current societal drivers regarding future conflict, such as the lack of guidelines for managing escalation and the changing geography for future conflict (land, sea, air, space, cyber, etc.). Bellingham also shared that an enormous amount of research needs to be done regarding the connection between humans and AI. He ended by noting that getting ahead of the curve is important to slow down adversaries.

AI: CURRENT AND FUTURE CHALLENGES

Matt Turek (DARPA) began discussing current AI breakthrough applications, such as AlphaGo, DeepBlue, and more. However, even with all of this success, we may not be on the right trajectory with AI. For example, he brought up self-driving cars—specifically Tesla's autopilot feature, and how it relies on computer vision, not multimodal sensing. He also spoke about how even when autopilot mode is engaged, Tesla holds the human drivers responsible. He continued by saying that users are not at a point where they can reliably delegate critical decisions to autonomy. He mentioned that some people have been excited in parts of the AI/ML community—but they are not working in ways comparable to humans. He noted that state-of-the-art large language models lack basic comprehension, fail to answer simple but uncommon questions or match simple captions and pictures, do not understand social reasoning, do not understand psychological reasoning, and do not understand how to track objects.

Turek then commented on his belief that the evaluation of AI/ML systems is broken. He talked about how we are chasing very narrow benchmarks and optimizing performance against those benchmarks. According to Turek, this just reinforces the building of narrow and relatively fragile AI systems. He then spoke about how current evaluation techniques do not encourage AI/ML systems to generalize. He also explained how current evaluation techniques do not reveal AI/ML fragilities.

Turek identified how DoD needs do not align with the focus of the AI/ML industry, as follows:

Industry is profit-driven, has access to massive amounts of data, has a low cost of errors, and faces threats from commercial adversaries. DoD is purpose-driven, has access to limited amounts of data, has a high cost of errors, and faces threats from active nation-state-level adversaries.

Turek then highlighted what may be possible as future national security–relevant capabilities:

- Trustworthy autonomous agents who sense and act with superhuman speed and can adapt to new situations;
- Intuitive AI teammates who can communicate fluently in human–native forms;
- Agents that promote national security; and
- Knowledge navigation for intelligence and accelerating defense technology development.

To realize some of these capabilities, Turek stressed the importance of investment in AI engineering, human context, and theory to help build robust DoD AI systems.

Turek closed by stressing the importance of:

- Developing theories of deep learning and ML;
- Measuring real-world performance by developing a rigorous experimental design that measures fundamental capabilities and produces generalizable systems;
- Focusing not only on performance but also on resource efficiency;
- Developing compositional models using principles approaches to exchange knowledge; and

- Developing appropriately trustable AI systems that have predictable adherence to agreed-upon principles, processes, and alignment of purpose.

Shanahan responded to Turek's comment that the trial-and-error approach to AI testing is no longer acceptable by saying that human beings do an awful lot of trial-and-error learning. He then asked if DARPA has been looking at hybrid approaches to solving the AI T&E problem. Turek responded that DARPA is interested in hybrid AI, particularly across statistical and symbolic approaches. Last, Longstaff asked Turek if we are going in the right direction in regard to making advances in the fundamentals of AI. Turek responded that he does not have a magic solution, but that his team is trying to set a vision for things that they think need to be done.

HUMAN AI: TEAMING IS UBIQUITOUS

Nancy Cooke (Arizona State University) was the day's final speaker. She began by discussing human-AI teaming. She stated that AI could not be effectively developed or implemented without consideration of the human. AI does not operate in a vacuum and will interface with multiple humans and other AI agents.

Cooke then spoke about different aspects of teaming, specifically regarding team composition and role assignment, processes, development, and effectiveness measurement. She then talked about the Synthetic Teammate Project on which she is working. The project's objective is to develop a synthetic AI teammate to take the place of air vehicle operators and work with two humans in the remotely piloted aircraft system task, Cooke said.

Cooke affirms taking human–machine teaming seriously. She defined a team as two or more teammates with heterogeneous roles and responsibilities who work independently toward a common goal. She then commented on what is currently known regarding human–AI teaming. Her first point was that team members have different roles and responsibilities and that this argues against having AI replicate humans. It also upholds that narrower AI allows AI to do what it is best at, such as big data analytics and visualization for humans. Her second point was that effective teams understand that each member has different roles and responsibilities that avoid role confusion but back each other up as necessary. Cooke stated that AI should understand the whole task to provide effective backup. Her third point was that effective teams share knowledge about the team goals and the current situation; over time, this facilitates coordination and implicit communication. Cooke stated that human–AI team training should be considered and that we should not expect a human to be matched with an AI system and immediately know how to work well together. Her fourth point was that effective teams have team members who are independent and thus need to interact or communicate, even when direct communication is not possible. Cooke said this argues not necessarily for natural language but maybe some other communication model. The fifth thing we know is that interpersonal trust is important to human teams. Cooke stated that AI needs to explain, provide a reason for its decision, and be explicable.

Cooke then spoke about the challenge with human–AI teaming. She stated that research on human–AI teaming cannot wait until AI is developed; it is then too late to provide meaningful input. Instead, a research environment, or testbed, is needed to get ahead of the curve and conduct research that can guide AI development. She then discussed different examples of physical and virtual synthetic test environments. Next, she introduced a concept called the "Wizard of Oz" paradigm in which a human plays the role of the AI or even remotely operates a robot to simulate very intelligent AI in a task environment. She also spoke about the importance of measures and models when measuring different aspects of human–AI teaming effectiveness.

Longstaff asked, regarding the Synthetic Teammate Project, how she would write a requirement for someone else to develop that pilot program? He also asked how she would test what she got back from the developer to know if she got the right product. Cooke responded that she would write down the details and results about her team's experiment and say, make it better so that it succeeds. She also stated that they would test it the

same way her team tested it the first time. Robin R. Murphy (Texas A&M University; workshop planning committee member) jumped in and asked if the way to ensure that the synthetic agent is aware of its team responsibilities would be to develop an Adaptive Control of Thought—Rational model of the entire operational space. Cooke responded that she thinks so. Chellappa asked how she sees human–AI as different, better, or more complicated than human–computer interaction (HCI). Cooke responded that much is known about human systems integration that can be brought to bear on human–AI systems that people do not consider on HCI when one person is interacting with a product. She also stated that HCI had not done much in massive system areas such as JADC2. Shanahan asked if Cooke had a separately controlled experiment where it was a three-AI team with no humans involved. He also asked if there were any takeaways regarding their work with HSI. Cooke responded that they had not done any three-agent teaming. She also stated that one of her takeaways was that AI is too often optimized on task work when it is important, in these complex systems, to optimize teamwork. Casterline commented that there are concepts of multiple agents being able to learn how to work better to serve an objective function in robotics, game theory, and more. Robin Murphy asked about metrics. Specifically, how can they estimate whether one team is more likely to produce the right answers than another? Cooke responded that they have metrics and are trying to develop more, such as domain-independent measures. Longstaff and Cooke talked about AI training in the context of human–animal teaming. Robin Murphy asked about the feasibility of predicting the performance of a human–AI team. Cooke responded that she could not think of a way to do it without seeing them perform, potentially in a training scenario.

DAY 2: WRAP-UP DISCUSSION

Casterline started the day's wrap-up discussion by stating that she found it interesting how one presentation spoke about how people will not really be able to trust AI, so they just have to accept the risk, versus when AI trust is really a requirement and more rigor is necessary. David S. Rosenblum (George Mason University; workshop planning committee member) said that he was struck by the fact that many presentations made it seem hard to separate any discussion of T&E from the requirements against which the T&E is being performed. Owing to the narrow statement of task, Rosenblum questioned the extent to which the workshop planning committee would be concerned about saying anything about requirements. The workshop planning committee also broadly discussed the inability to avoid the question or discussion surrounding requirements when it comes to T&E. The workshop planning committee also spoke about other sectors where the consequences of AI would be high. Serrano mentioned that the question of liability resonated with him throughout the day and that we build these systems of systems inside some organization designed by committee. He asked, "Who is going to take ownership for how this thing should perform?" The workshop planning committee ended the day by discussing the break between what happens in the development community and the research community.

VERTICAL DATA SCIENCE

Bin Yu (University of California, Berkeley) defined vertical data science as the process of extracting reliable and reproducible information from data with an enriched, technical language to communicate and evaluate empirical evidence in the context of human decisions and domain knowledge. Yu introduced the predictability, computability, and stability (PCS) framework. She stated that PCS is a way to unify, streamline, and expand on ideas and best practices in both ML and statistics. She also spoke about the importance of documentation.

Yu broke down each part of the PCS framework. Concerning problem formulation, predictability reminds us to keep in mind future situations where AI/ML algorithms will be used while developing AI/ML algorithms. Concerning data collection and data comparability, predictability reminds us to keep in mind future situations where AI/ML algorithms will be used while developing AI/ML algorithms. Concerning data comparability, stability reminds us to keep in mind that there are multiple reasonable ways to clean or curate a given data set from the current situation. Regarding data

partitioning, stability reminds us to keep in mind that there could be multiple reasonable ways to partition a given data set from the current situation to ensure that the test set is as similar to future situations as possible. Last, regarding other forms of data perturbations, stability reminds us to keep in mind that data perturbations should reflect future situations.

Regarding comparing different predictive techniques, Longstaff asked if any quantitative measures were currently incorporated into the framework to help choose the best algorithm or technique. Yu responded by identifying two measures, sensitivity and specificity. Longstaff followed up by asking how to reason about the trade-offs between predictability and stability. Yu responded that her team screens for predictive performance before seeking stability. Casterline and Yu discussed translating operational requirements into the mathematical statistics to which Yu referred. Shanahan asked about justified confidence and how doctors and nurses attain that. Yu responded that it is important to understand their work and profession as much as possible when developing models. Kolda and Yu talked about embedding T&E with operations. Yu ended her talk by speaking about the importance of documentation and metrics.

AN APPLIED RESEARCH PERSPECTIVE ON ADVERSARIAL ROBUSTNESS AND TESTING

Nathan VanHoudnos (Software Engineering Institute) spoke about AI security and making systems do the wrong thing. First, he introduced the Bieler taxonomy, where an adversary can make you learn, do, and reveal the wrong thing. Next, he compared these different things to data poisoning, adversarial patches, model inversion, and membership inference attacks. He then stated that in their laboratory, they focus on training systems to learn correctly, do things correctly, and not reveal secrets. Next, when it comes to verifying a system, VanHoudnos spoke about their "Train and Verify" project, where they try to make robust ML systems that do not reveal secrets, as well as private ML systems that are not fooled as easily. Last, he introduced several other projects focusing on protecting systems from many adversarial techniques mentioned above.

VanHoudnos defined AI corruption as a decrease in a quality attribute of an AI system. He then spoke about the different roles played by different people as teams try to accomplish different missions. The discussion then shifted toward evaluating ML models—specifically, the evaluations should reflect how models will be used in practice and specific scenarios of importance to the application of the model. Thought should also be given to metrics you care about when evaluating. Chellappa commented that one of the reasons he thinks many benchmarks are averages is because there is a desire to avoid somebody optimizing the algorithm for just one point on the plot that may be operationally relevant. VanHoudnos then spoke about different examples of AI corruption. Casterline and VanHoudnos discussed adversarial patches in classification and the idea of using an adversarial patch to test against a retrained model. Casterline compared it to a cat-and-mouse game and wondered if this is truly the right approach, to constantly devise counterattacks for the continual stream of adversarial attacks that continue to evolve no matter what is done. Longstaff asked where in the requirements process they would know that a certain quality attribute of an AI-enabled system will be tested. VanHoudnos responded that he would have to defer to the DevSecOps folks. Longstaff followed and provided his thoughts on the question: creating the quality attributes is a collaboration between a team of operators, testers, and development folks. VanHoudnos wrapped up by discussing the concept of justified confidence[7] with Longstaff.

DEFENDING AI SYSTEMS AGAINST ADVERSARIAL ATTACKS

Bruce Draper (DARPA) began his presentation by talking about different types of adversarial attacks against data models. He then spoke about algorithmic defenses for AI systems—specifically, regarding five best practices.

The first best practice discussed was cyber defense. Draper stated that networks are vulnerable, and most AI systems are attached to a network. He also stated that

[7] Justified confidence is about developing AI systems that are robust, reliable, and accountable, and ensuring that these attributes can be verified and validated. Northrop Grumman, 2021, "AI Development Aligns with US Department of Defense's Ethics Principles," https://news.northropgrumman.com/news/features/northrop-grumman-building-justified-confidence-for-integrated-artificial-intelligence-systems.

it is relatively easier to attack a network than an AI. Therefore, he suggested that to defend the AI, the focus must be on defending the network.

The second best practice discussed was protecting the input data—specifically, sensor-inspired data. Draper described two types of attacks, one revolving around having access to the actual digital signal, which makes spoofing very easy. The other type of attack is physical. These attacks revolve around altering items in the physical world to trick a system. Draper noted that physical attacks are harder for an adversary to launch and easier to defend. He ended by stressing the idea of protecting your data.

The third best practice was about collecting inputs from multiple sources. Draper stated that it is harder to spoof multiple sensors than one sensor. He also noted that different types of sensors make it even harder to disrupt, and having different instances of sensors also offers some benefits.

The fourth best practice discussed was about protecting model development. Draper urged everyone to be wary of externally acquired models. They may have back doors, either unintentionally or from poisoning, and if an adversary has access to the model, it enables white-box attacks. Draper also stated that when training your models, you should avoid using untrusted training data, avoid boot-strapping from untrusted models, and keep information about training data private.

The fifth best practice was quality assurance post-fielding. Draper advised, when possible, to have a person double-check sampled AI outputs.

Draper then spoke about how one can increase system robustness. He introduced a few methods, such as adversarial training and randomized smoothing. Adversarial training is when you attack your sample during the training process. It does not slow you down at runtime but slows training. Randomized smoothing is where you wait to get the input and then make different versions of that input. It has the opposite pros and cons, where it makes training faster but will slow you down at runtime. Draper noted that both of these methods require a known threat model. The danger is that if the adversary does something you do not anticipate, these methods will not work. Draper also spoke about some methods against physical attacks, such as tile-based defenses and patch detection defenses.

Draper ended his talk by speaking about evaluation software and tools. He concluded with DARPA's Guaranteeing AI Robustness Against Deception Armory, an evaluation tool that allows analysts to run adversarial AI experiments at scale quickly and repeatedly.

RECOGNITION SYSTEM EVALUATION

Ed Zelnio (Air Force Research Laboratory) spoke about imaging systems and the different types of data: sensor data, metadata, and labels. He also spoke about labeling—specifically, regarding granularity. Zelnio then spoke about different categories of target data and introduced the categories in library mission targets, library confusers, out-of-library confusers, and clutter. He also went over the difference between developmental and operational data.

Zelnio introduced "some things that would be nice to measure in terms of evaluation." He spoke about measuring the reliability and confidence in a system, measuring understandability and trust of a system, measuring the robustness of a system, measuring the effectiveness of out-of-library confusers, measuring the performance of a performance model, figuring out what to do with limited data, and the need to talk about a sustainable end-to-end training process.

Zelnio ended by speaking about best practices. The first best practice is coming up with an expectation management agreement. These tell you under what operating conditions you can expect a given system to work. The second best practice is the use of a test harness than can help to reproduce training and aid in evaluating the algorithm and the training process. Last, the third best practice is testing to break, to see what does and does not work.

Longstaff asked what has helped get customers over the barrier that allows them to increase their confidence in using an experimental system in an operational environment. Zelnio responded that the expectation management agreements are important in increasing that confidence. He also spoke about keeping demonstrations as relevant as possible to excite operators. Longstaff also asked if they could receive additional feedback from an operational customer over time that would allow retraining or retesting opportunities. Zelnio responded that it would be great to have a laboratory in the loop to help with this, but it happens more informally.

AFTC AI INFRASTRUCTURE NEEDS

Eileen Bjorkman (AFTC) was the workshop's final speaker. Bjorkman spoke about AFTC's current objective of looking at the unique infrastructure needs within the test center and across different organizations to set itself up to test autonomous systems.

Bjorkman spoke about three main things to think about in the testing process. First, test safety, particularly in making sure that an operator can contain a system if it begins to perform in ways that they do not expect. Second, early tester involvement and how testing strategies must be built into system design. The final need revolved around test infrastructure—specifically, instrumentation, data collection and storage, and range support.

Bjorkman also spoke about current T&E needs and how there is no enterprise-level T&E infrastructure to support autonomy testing. She also stated that there is no DoD enterprise-level software T&E infrastructure that supports the testing of AI. She then spoke about different investment areas that she thinks need to happen. These investment areas focused on architectures, frameworks, modular subsystems, data management, virtual ranges, agile workforce, and surrogate platforms, Bjorkman said.

Bjorkman ended by discussing an autonomous system and AI roadmap that showed funding for different programs over a 7-year timeline. Next, Casterline and Bjorkman began discussing the use of simulation work in their virtual environments. Longstaff asked if they have begun incorporating digital twins work into the testing process for autonomous systems. Bjorkman responded that she had seen that happening. Next, Shanahan, Casterline, and Bjorkman discussed data and the use of virtual testing environments. Bjorkman commented on how you cannot get enough replications of things in the real world to test a system fully. She posed an example about autonomous cars and how you cannot just go and drive every road 100 times. She also pointed out how often she has observed so many different tests that they perform where they cannot collect anywhere near sufficient live data. As a result, it forces them into a virtual or even a constructive environment.

WORKSHOP WRAP-UP AND DISCUSSION

The workshop planning committee began its wrap-up by discussing its final thoughts from the workshop. Casterline commented that she does not think that any of the systems are prepared for the iteration that they will have to facilitate. She also said that there is a lot to "grab from" and apply here regarding the DevSecOps model for software. Shanahan commented about the culture shift of iteration and adaptability. If the Air Force does not get that right, everything else is just another discussion about T&E. Chellappa commented about the idea of a centralized facility to test AI. He also commented that he does not think that we know what it means to test AI right now. He specifically pointed out the metrics mAP and Recall and commented how these are ideas from the 1970s. He questioned why they had become a metric for current AI systems. Kolda wanted to stress that not everything is in the data. She also warned against the idea of trusting AI too much.

Additionally, Kolda commented that humans need to evaluate the answers that come out of an AI system and not just blindly accept them. Casterline commented about a gap in the vernacular between algorithmic tests and measurements and operational relevance and tests. Kolda questioned how the process of continuous integration, evaluation, and feedback would work. Last, Robin Murphy offered her thoughts and spoke about how human work processes need to be considered in the discussion regarding AI corruption. She also commented about her fear that security is viewed as an algorithmic

problem and that it is just going to come up with another algorithm that will detect when the AI is not working correctly.

Longstaff discussed the major questions from the statement of task. His first point, regarding the task of evaluating and contrasting current T&E, was that there is very little overlap between the way T&E is done commercially and the way that the workshop planning committee experienced it through the examples in the workshop presentations. Next, he focused on AI corruption and stated that the third question from the statement of task goes beyond being just a scientific technology question. It could also ask how DoD and the Air Force can improve the nature of how technological advances are incorporated. Rosenblum commented that, regarding the third question, he is worried that anything the workshop planning committee says will be outdated in a year or two owing to the rapid pace of technological change. Longstaff responded and stated that the workshop planning committee could point toward general trends instead of specific scientific advances. Last, Longstaff thanked everyone for their contributions, and staff member Ryan Murphy officially closed the workshop.

DISCLAIMER This Proceedings of a Workshop—in Brief was prepared by **EVAN ELWELL** as a factual summary of what occurred at the workshop. The statements made are those of the rapporteur or individual workshop participants and do not necessarily represent the views of all workshop participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine

REVIEWERS To ensure that it meets institutional standards for quality and objectivity, this Proceedings of a Workshop—in Brief was reviewed by **LIDA BENINSON,** National Academies of Sciences, Engineering, and Medicine; **TED BOWLDS,** U.S. Air Force (Retired); and **JOHN N. SHANAHAN,** U.S. Air Force (Retired). **KATIRIA ORTIZ,** National Academies of Sciences, Engineering, and Medicine, served as the review coordinator.

SPONSOR This workshop was supported by the U.S. Air Force.

For additional information regarding the workshop, visit https://www.nationalacademies.org/event/06-27-2022/testing-evaluating-and-assessing-artificial-intelligence-enabled-systems-under-operational-conditions-for-the-department-of-the-air-force-workshop.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2023. *Testing, Evaluating, and Assessing Artificial Intelligence–Enabled Systems Under Operational Conditions for the Department of the Air Force: Proceedings of a Workshop—in Brief*. Washington, DC: The National Academies Press, https://doi.org/10.17226/26885.

Division on Engineering and Physical Sciences

NATIONAL ACADEMIES *Sciences Engineering Medicine*

The National Academies provide independent, trustworthy advice that advances solutions to society's most complex challenges.
www.nationalacademies.org