REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
03-08-2020	Final Report	31-May-2019 - 30-May-2020

4. TITLE AND SUBTITLE
Final Report: Artificial Intelligence Cybersecurity Workshop

5a. CONTRACT NUMBER
W911NF-19-1-0345

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Vanderbilt University
PMB 407749
2301 Vanderbilt Place
Nashville, TN 37240 -7749

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75472-NS-CF.8

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Mary Dey
UU	UU	UU	UU		19b. TELEPHONE NUMBER 615-875-2415

RPPR Final Report
as of 31-Aug-2020

Agency Code:

Proposal Number: 75472NSCF **Agreement Number: W911NF-19-1-0345**
INVESTIGATOR(S):

Name: Mary K. Dey
Email: katie.dey@vanderbilt.edu
Phone Number: 6158752415
Principal: Y

Organization: **Vanderbilt University**
Address: PMB 407749, Nashville, TN 372407749
Country: USA
DUNS Number: 965717143 EIN: 620476822
Report Date: 30-Aug-2020 Date Received: 03-Aug-2020
Final Report for Period Beginning 31-May-2019 and Ending 30-May-2020
Title: Artificial Intelligence Cybersecurity Workshop
Begin Performance Period: 31-May-2019 **End Performance Period:** 30-May-2020
Report Term: 0-Other
Submitted By: Mary Dey Email: katie.dey@vanderbilt.edu
Phone: (615) 875-2415
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: In collaboration with the NSTC's NITRD Subcommittee and NSTC's MLAI Subcommittee a workshop was held to assess the key research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). The goal of the workshop was to discuss current and future research activities in this space, including potential research gaps to help the federal government plan future R&D in this area.

The workshop focused on identifying capability gaps and prospective research directions relating to the use of AI and machine learning (ML), including: (1) Using AI to improve security, (1a) Using AI to understand threat models and to improve the detection of threats, the protection of systems, and the adaptation of systems to increase their resiliency to future cyber attacks (1b) Developing defensive methods that effectively counter cyber-attacks that utilize AI capabilities and (2) Improving security of AI, (2a) Defining and understanding AI vulnerabilities, (2b) Improving resiliency of AI methods and algorithms to various forms of attacks

Accomplishments: An Artificial Intelligence and Cybersecurity Workshop was held to assess the key research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). The workshop brought together top academic, commercial, and government subject matter experts to discuss capability gaps and prospective research directions relating to the use of AI and machine learning. This funded effort focuses on providing travel support for a diverse community of participants in the AI/ML workshop.

The workshop was held June 4-6, 2019 at the University of Maryland College Park, Maryland. The program agenda featured a mix of invited talks, small group writing sessions, and a poster session. The workshop opened with a charge from the co-chairs: John Launchbury (Galois) and Patrick McDaniel (Penn State). Sessions were organized around the following topics: AI for Security, Security of AI, Industry Research, Academic Research, and Government Mission Context.

Invited Speakers included Yan Shoshitaishvili (Arizona State University), Michael Kearns (University of Pennsylvania), Ulfar Erlingsson (Google Brain), David Wagner (UC Berkeley), and Dean Souleles (Office of the Director of National Intelligence).

Each invited presentation was followed by small group breakout sessions. Each small group had a facilitator and approximately 6-8 participants. The groups were each assigned a different topic area. Groups were tasked with brainstorm specific needs and opportunities within that topic area. After an initial brainstorming session each group was tasked with writing an initial draft for that section of the report. A initial draft of the report was produced during the workshop and the participants continued to refine the report in the months after.

The workshop concluded with a student poster session. 11 students presented posters on AI/ML research.

The outcome of the workshop was a detailed technical workshop report that summarizes research challenges and opportunities at the intersection of AI and cybersecurity. The report was presented to science and technology leadership across the U.S. government and beyond.

No travel funds were used for government attendees.

Training Opportunities: Nothing to Report

Results Dissemination: 1. A technical workshop summary report was disseminated by the NSTC on March 2, 2020.
2. A final detail technical workshop report was released to the community on June 2, 2020.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Mary K Dey
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

WEBSITES:

URL: https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019
Date Received:
Title: AI and Cybersecurity Workshop Website
Description: Workshop website

Artificial Intelligence and Cybersecurity Workshop

The Hotel at the University of Maryland

7777 Baltimore Ave. College Park, MD 20740

June 4-6, 2019

Agenda

TUESDAY, JUNE 4

07:30 – 08:00 ARRIVAL / CHECK-IN

08:00 – 08:30 **Welcome and Charge**
John Launchbury (Galois) and Patrick McDaniel (Penn State)

08:30 – 09:15 **Plenary I: AI for Security**

The Dangers of the Subconscious Mind (of Cyber Reasoning Systems)
Yan Shoshitaishvili (Arizona State University)

09:15 – 09:30 BREAK

09:30 – 12:00 **Breakouts I: AI for Security** *(with informal break as needed)*

12:00 – 13:00 LUNCH

13:00 – 13:45 **Plenary II: Security of AI**

Individual Fairness for Machine Learning
Michael Kearns (University of Pennsylvania)

13:45 – 14:00 BREAK

14:00 – 16:30 **Breakouts II: Security of AI** *(with informal break as needed)*

16:30 ADJOURN

WEDNESDAY, JUNE 5

08:00 – 08:30 ARRIVAL / CHECK-IN

08:30 – 09:15 **Plenary III: Industry Research**
Úlfar Erlingsson (Google)

09:15 – 09:30 BREAK

09:30 – 12:00 **BREAKOUTS III: Broader questions** *(with informal break as needed)*

12:00 – 13:00 LUNCH

13:00 – 13:45 **PLENARY IV: Academic Research**

Security Against Adversarial Examples
David Wagner (University California, Berkeley)

13:45 – 14:00 BREAK

14:00 – 16:30 **BREAKOUTS IV: Writing Session** *(with informal break as needed)*

16:30 ADJOURN

THURSDAY, JUNE 6

08:00 – 08:30 ARRIVAL / CHECK-IN

08:30 – 09:15 **PLENARY V: Government Mission Context**

Mona Lisa Talks
Dean Souleles (Office of the Director of National Intelligence)

09:15 – 09:30 BREAK

09:30 – 11:45 **Poster Session**

11:45 – 12:00 **Thanks and Adjourn**

Artificial Intelligence & Cybersecurity Workshop

June 4-6, 2019

Attendees

Last Name	First Name	Affiliation
Alstott	Jeff	Intelligence Advanced Research Projects Activity (IARPA)
Bauer	Lujo	Carnegie Mellon University
Beieler	John	IARPA
Burke	Quinn	Pennsylvania State University
Butler	Nekeia	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Chellappa	Rama	University of Maryland
Clouse	Dan	Department of Defense
Corbett	Matthieu	United States Navy
Dey	Katie	Vanderbilt University
DSouza	Faisal	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Dumitras	Tudor	University of Maryland
Dupree	Lynn	Privacy and Civil Liberties Oversight Board
Dyson	Anne	Cyber Pack Ventures
Edwards	Christine	National Security Agency
Erlingsson	Ulfar	Google Brain
Evans	David	University of Virginia
Everett	John	Defense Advanced Research Projects Agency (DARPA)
Garris	Michael	National Institute of Standards and Technology
Gaston	Matthew	Carnegie Mellon University Software Engineering Institute
Geck	Frank	United States Army
Giorgio	Edward	Bridgery Technologies
Groth	Jacob	Defense Point Security
Jajodia	Sushil	George Mason University
Jayaraman	Bargav	University of Virginia
Jere	Malhar	University of California San Diego
Joseph	Sarah	Department of Defense
Kantarcioglu	Murat	University of Texas at Dallas
Kautz	Henry	National Science Foundation
Kearns	Michael	University of Pennsylvania
Keromytis	Angelos	Georgia Institute of Technology
Lincoln	Patrick	SRI International
Liu	Gabrielle	Student
Madry	Aleksander	Massachusetts Institute of Technology (MIT)

Artificial Intelligence & Cybersecurity Workshop

June 4-6, 2019

Attendees

Magill	Stephen	Galois, Inc.
Martin	William	National Security Agency
McDaniel	Patrick	Pennsylvania State University
Molina-Markham	Andres	MITRE Corporation
Nemr	Christopher	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Nguyen	Tristan	Air Force Office of Scientific Research
Papernot	Nicolas	Google Brain
Patwardhan	Dinesh	Food and Drug Administration
Piotrowski	Victor	National Science Foundation
Ray	Indrajit	National Science Foundation
Richards	Raymond	DARPA
Ridley	Ahmad	Department of Defense
Roberts	Kamie	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Scherlis	William	Carnegie Mellon University
Schwartz Drobnis	Ann	Computing Community Consortium
Sharif	Mahmood	Carnegie Mellon University
Sheatsley	Ryan	Pennsylvania State University
Shoshitaishvili	Yan	Arizona State University
Shrobe	Howard	MIT Computer Science and Artificial Intelligence Laboratory
Souleles	Dean	Office of the Director of National Intelligence
Stanley	Martin	Department of Homeland Security
Streilein	William	MIT Lincoln Laboratory
Sun	Kun	George Mason University
Suya	Fnu	University of Virginia
Thai	Alex	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Thuraisingham	Bhavani	University of Texas at Dallas
Tobin	Noah	Georgia Tech Research Institute
Tong	Liang	Washington University in St. Louis
Vagoun	Tomas	Networking and Information Technology Research and Development (NITRD) Program National Coordination Office (NCO)
Vorobeychik	Yevgeniy	Washington University in St. Louis
Wagner	David	University of California, Berkeley
Wang	Shu	George Mason University
Wellman	Michael	University of Michigan
Zhang	Xiao	University of Virginia

Artificial Intelligence and Cybersecurity: A Detailed Technical Workshop Report

The Networking & Information Technology R&D Program

June 2020

Table of Contents

Executive Summary

On June 4-6, 2019, the National Information Technology and Networking Research and Development (NITRD) Program's Artificial Intelligence Research and Development (R&D) and Cyber Security and Information Assurance Interagency Working Groups (IWG), held a workshop[1] to assess the research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). This document summarizes the workshop discussions.

Technology is at an inflection point in history. AI and machine learning (ML) are advancing faster than society's ability to absorb and understand them; at the same time, computing systems that employ AI and ML are becoming more pervasive and critical. These new capabilities can make the world safer and more affordable, just, and environmentally sound; conversely, they introduce security challenges that could imperil public and private life.

Though often used interchangeably, the terms AI and ML refer to two interrelated concepts. Coined in the 1950s, AI is the field of computer science that refers to programs intended to model "intelligence." In practice, this refers to algorithms that can reason or learn given the necessary inputs and base knowledge and are used for tasks such as planning, recognition, and autonomous decision-making (e.g., weather prediction). ML is a specialized branch of AI that uses algorithms to understand models of phenomena from examples (i.e., statistical machine learning) or experience (i.e., reinforcement learning). Throughout this document the term AI will be used to discuss topics that apply to the broad field, and ML will be used when discussing topics specific to machine learning.

The challenges are manifold. AI systems need to be secure, which includes understanding what it means for them to "be secure." Additionally, AI techniques could change the current asymmetric defender-versus-adversary balance in cybersecurity. The speed and accuracy of these advances will enable systems to act autonomously, to react and defend at wire speed,[2] and to detect overt and covert adversarial reconnaissance and attacks. Therefore, securing the Nation's future requires substantial research investment in both AI and cybersecurity.

AI investments must advance the theory and practice of secure AI-enabled system construction and deployment. Considerable efforts in managing AI are needed to produce secure training; defend models from adversarial inputs and reconnaissance; and verify model robustness, fairness, and privacy. This includes secure AI-based decision-making and methods for the trustworthy use of AI-human systems and environments. This will require a science, practice, and engineering discipline for the integration of AI into computational and cyber-physical systems that includes the collection and distribution of an AI corpus—including systems, models and datasets—for educational, research, and validation.

For cybersecurity, research investments must apply AI-systems within critical infrastructure to help resolve persistent cybersecurity challenges. Current techniques include network monitoring for detecting anomalies, software analysis techniques to identify vulnerabilities in code, and cyber-reasoning systems to synthesize defensive patches at first indication of attack. AI systems can perform these analyses in seconds instead of days or weeks; in principle, cyber-attacks could be observed and defended against as they occur. However, safe deployment will require understanding the multiple dimensions and implications of these AI actions.

[1] https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019

[2] *Wire speed* is the rate of data transfer that a telecommunication technology provides at the physical level (hardware wire, box, or function) and that supports the data transfer rate without slowing it down.

Introduction

The Networking and Information Technology Research and Development (NITRD) Program's Artificial Intelligence R&D, and Cyber Security and Information Assurance, IWGs held a workshop to assess the research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). The workshop, held June 4–6, 2019, brought together senior members of the government, academic, and industrial communities. The participants discussed the current state of the art, future research needs, and key research and capability gaps. This document is a summary of those discussions. For more details, including the agenda, please go to the workshop webpage.[3]

The document is divided into three topic areas: AI for Cybersecurity, Security of AI, and Science and Engineering Community Needs. These areas intentionally overlap and intertwine to reflect the multiple contexts and vantage points discussed. Therefore, the reader should not consider the document's organization to provide rigid structure to any larger initiative, but rather to provide a free form for discussion of the relevant topics. Developing a specific structure or prescriptive task list for this pressing domain is outside the scope of the workshop effort. Such a determination and resulting plan will require substantial effort across many organizations over many years.

Security of AI

Recent advances in AI have vastly improved the capabilities of computational reasoning and exceed human-level performance in tasks like image recognition, natural language processing, and data analytics. The applications of these new technologies are transformative. Autonomous vehicles will soon transform transportation, and virtual assistants have already become part of everyday life. The economic drivers of these technologies will result in their broad adoption and will disrupt almost every aspect of the enterprise.

However, when AI-systems are exposed to adversarial behavior, they can be manipulated, fooled, evaded, and misled in ways that can have profound security implications. As more critical systems employ AI, whether financial systems, self-driving cars, network monitoring tools, or military applications, it is vitally important to develop techniques and best practices to make them more robust.

Specification and Verification of AI Systems

Integrated AI systems involve perception, learning, decisions, and actions in complex environments. These four components employ diverse AI technologies including both statistical and symbolic approaches. There are interactions and interdependencies among these components (e.g., errors made in perception can cause an otherwise intact decision-making component to behave incorrectly). Furthermore, there are unique vulnerabilities in each of the components (e.g., perceptual systems are prone to training attacks whereas decision-making components are susceptible to classic cyber exploits). Finally, the notion of correctness is not a purely logical matter; every component involves noise and uncertainty that require bounds to protect the system from misbehaving.

There is a pressing need for definitions and methods to formally verify AI and ML components, both independently and in concert. Verification as it relates to logical correctness, decision theory, and risk analysis needs to be explored. New techniques are needed for AI system specification, validation, and

[3] https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019

verification that specify what a system is expected to do and how the system responds when subjected to adversarial manipulation.

Techniques for AI System Specification and Validation

Specification of an engineered system involves clear, quantifiable statements of purpose, design, components, and component interactions that the system will be required to meet. In traditional systems, information is available for the components, and qualities that match the specification are tractable. Because AI components and their interactions are so complex, it is difficult to identify attributes that match the specifications. Research is needed into methods and metrics that enable identification, description, and characterization of complex AI components to measure specification compliance. Methods that provide statistical bounds on AI systems could be leveraged, as could current techniques for identifying and controlling component interactions.

Because AI systems operate in open environments, the range of input values or distributions is difficult to predict. Therefore, research is needed to develop techniques that can reason in the opposite, and more difficult, direction as well. Rather than wait for random inputs, it will be necessary to determine, based on the risk profile of the system, the type of inputs needed for the system to behave as desired.

Verification of AI Systems

Deployed AI systems are often extremely complex, and their implementation and configuration are difficult to assess. Research is needed in architectural structures and analysis techniques that allow verification of these components as part of a larger effort to develop manageable standards, best practices, tools, and methods to reason about the behavior of a system.

A new discipline and science of AI architecture could leverage an AI "building code" that provides guidelines for the composition of such a system. These guidelines would need to consider the overall goals, the AI component goals, and the system's threat model. Such a building code could come from theory (statistical or numerical) or experience (hard-won lessons) and capture justified best practices. Leveraging codes or guidelines from other fields, such as software, databases, and computer systems, may be possible. In addition, analysis of the building code would lead to a better understanding of the successful underlying (e.g., latent) AI mechanisms and thus move the field forward.

However, specification and verification must also use pairwise (or more) comparisons of aspects such as performance, security, robustness, and fairness. Research is needed to better understand the tradeoffs and determine when the environment can safely support specific operations. In many domains, defining the correct, incorrect, and desired behavior of a system will require a domain expert on the team. And finally, an engineer must be identified to take these frameworks and implement, deploy, and maintain the AI system.

Trustworthy AI Decision-Making

As AI systems are deployed in high-value environments, the issue of ensuring that the decision process is trustworthy, particularly in adversarial scenarios, is paramount. Given the potential for harm, it is crucial to develop methods and principles for trustworthiness. Research is needed to develop principles for a wide array of AI systems, including ML, planning, reasoning, and knowledge representation. Areas that need to be addressed include:

- Defining performance metrics for trustworthy decision-making
- Making AI systems explainable and accountable
- Developing techniques for trustworthy decision-making

- Improving domain-specific training and reasoning
- Managing training data

Defining Performance Metrics for Trustworthy Decision-Making

In conjunction with research to develop threat models that capture realistic assumptions about adversarial capabilities and goals, research must also identify a set of measurable properties that define trustworthiness. A defender can then incorporate these measurable properties (e.g., robustness, privacy, and fairness) when designing decision-making algorithms. Metrics resulting from analysis, such as decision accuracy on a set of test points, will lead to trustworthiness metrics.

Threat models will have to reason about an adversary's ability to interact with a system as it is making decisions. One possibility is to draw from standard definitions established in cryptography or other areas of computer security research. Adversarial goals and performance metrics will be defined to capture all facets of trustworthy decision-making. One possibility is to unify these properties in a single reasoning framework and treat them as variants of a single notion of (in)stability in ML and AI. Note that metrics defined here will differ from existing notions of average case properties in the scientific literature. Analytical efforts will also need to define requisite conditions for an aspect of trustworthiness to be achievable given a specific threat model.

For trustworthy decision-making, research is needed on frameworks to reason about security properties in the broad sense, as well as analytical and empirical tools to measure how well these properties are satisfied by an AI decision-making system.

Making AI Systems Explainable and Accountable

It is often difficult to explain why an AI system produces a particular output for a given input. This problem is particularly acute in sophisticated ML systems such as deep learning. This issue increases the effort to debug a faulty system and creates many challenges in assessing accountability.

Research is needed in methods for understanding the learned reasoning of AI methods. For example, it is currently difficult to identify training points that define the boundaries separating different ML decision outcomes. How certain data points influence the optimization procedures (and hence the reasoning) involved in ML systems is a necessary research direction. This research could involve either analyzing the optimization procedure itself or the AI system outcome.

To link decisions made by an AI system to the relevant training data, techniques need to capture contributions from both the training data and the learning method. Techniques that can estimate the influence of each training point on individual predictions could become the basis for mechanisms that assess the relevance of a model in a decision environment.

Developing Techniques for Trustworthy Decision-Making

While there are numerous illustrations of ML vulnerabilities, science-based techniques to predict trustworthiness (e.g., understanding what makes path planning or probabilistic graphical models robust in adversarial settings) are elusive. Research is needed on techniques for either identifying trustworthy models or improving the trustworthiness of existing models.

In ML, there is an emerging set of approaches that provides decision guarantees using a variety of techniques (e.g., convex relaxation of the adversarial optimization problem and randomized smoothing). However, the approaches are currently focused almost exclusively on supervised learning and are difficult to achieve without degrading system performance. Important research questions remain:

- Are there innovative techniques that provide robustness guarantees?

- Beyond machine learning, for AI approaches such as planning, knowledge representation, and reasoning, what are realistic notions of vulnerabilities and trust?
- Are there empirical techniques that improve the trustworthiness of AI systems and degrade gracefully when exposed to unanticipated attacks?

Research is needed to identify when AI and ML fail. It is important to know when an output is unreliable (e.g., not supported by evidence) and should not be used for mission-critical decisions. A related area of research, AI systems that request guidance when they are uncertain, can improve trust in the eventual decision and allow the system to obtain information for future decision-making.

Improving Domain-Specific Training and Reasoning

The accuracy of AI is sensitive to the domain where it is used. AI systems can exhibit security vulnerabilities when training data is not representative of the deployment environment. Conversely, vulnerability analysis of AI without consideration of constraints in the application domain may result in overly pessimistic assessments. Research is needed on how input data is acquired and maintained within domain-specific AI environments, and on evaluation methods for the security of AI systems as they become a part of the full-use ecosystem.

AI systems consist of a collection of integral pieces, including data acquisition, model development and implementation, evaluation and validation, and their application. An autonomous vehicle system is trained with images and situations acquired from realistic environments and constantly maintained as its environment changes. Research must evaluate numerous domain-specific vulnerabilities such as perception, planning, reinforcement learning, knowledge representation, and reasoning. This may include reasoning about streaming data (as opposed to static images); consequences (such as causing a car to crash or go in the wrong direction); and adapting to unanticipated developments (such as weather conditions and road construction). It is important to understand how the data used to calibrate the model impacts the performance in the application domain. This research necessitates a rethinking of threat models and will lead to a science of deploying and maintaining AI systems in real-world environments.

Managing Training Data

Datasets are valuable (e.g., large network datasets can reveal everything about network vulnerabilities), but is the collected data more valuable for offense or defense? If the data is of higher value for an adversary, should it be collected? Privacy-preserving collection and storage can hide vulnerabilities while still providing information for defense. Research is needed to evaluate the cost/benefit ratio of collecting, protecting, and storing training data.

Detection and Mitigation of Adversarial Inputs

While AI performs well on many tasks, it is often vulnerable to corrupt inputs that produce inaccurate responses from the learning, reasoning, or planning systems. For example, there are examples where sophisticated deep learning methods can be fooled by small amounts of input noise carefully crafted by an adversary.[4] Such capabilities allow adversaries to control the systems with little fear of detection. As systems based on deep networks and other ML and AI algorithms become integrated into operational systems, it is critical to defend against adversarial inputs by considering:

- Making machine learning methods more robust
- Preventing AI reconnaissance

[4] There are many articles available on this topic, for example: Adversarial Attacks and Defenses: A Survey; https://arxiv.org/abs/1810.00069.

- Exploring the space of adversarial models
- Secure training
- Preventing model poisoning
- Training calibration, confidence, and retraining
- Training data privacy and model fairness

Robust Machine Learning Methods

Substantial efforts are also needed to harden learning methods against adversarial inputs. The robust statistics community has studied this problem—it is well understood in the context of linear regression and time series models—and the technical community has rigorous theoretical foundations and practical measures to address similar issues. Both theoretical and empirical research are needed to make the same advances for deep learning and modern ML methods without sacrificing performance or accuracy.

Preventing AI Reconnaissance

Modern AI systems are vulnerable to reconnaissance (i.e., adversaries can query the systems and learn the internal decision logic or knowledge bases and, in some cases, the training data). Reconnaissance can then be a precursor to an adversarial input attack to extract security-relevant training data and sources or to acquire the intellectual property embedded in the AI.

Research is needed to explore methods and mechanisms for preventing system or model reconnaissance; some possibilities follow:

- Increase the attacker workload and reduce attacker effectiveness through model inversion. This could include the use of noise and formal models such as differential privacy.
- Leverage cybersecurity approaches, including rate limiting, access controls, and deception.
- Study the impacts on accuracy, ability to explain, and other important aspects of learning algorithms and systems.
- Design reconnaissance-resistant algorithms and techniques.
- Integrate resistance into learning and reasoning optimizations.
- Embed security guarantees into the model using new multiple-step techniques.
- Expose the presence and goals of the attacker using the cybersecurity honeypot (attractive decoy) concept.

Exploring the Space of Adversarial Models

The vulnerability of an AI system is defined by the capabilities and knowledge of the adversary. Research is needed to classify the different types of attacks and develop appropriate defenses. Defenses need to address both the “white-box” attack where the attacker has complete access to the model architecture and parameters, including knowledge of all defenses that may be present, and the “black-box” attack where the attacker does not have access to the classification model parameters, and defense is more feasible. The space of these models needs to be carefully mapped, and attack and defense strategies identified.

Research is also needed on defending against physically realizable attacks with specific application in domains (e.g., autonomous vehicles and malware detection) where security is especially critical and ML models are most at risk.

Secure Training

AI and ML rely on learning from training data that allows the model to learn how to characterize expected inputs. However, this also introduces possible security risks. If the training instances do not represent all

possible situations, including future situations, then the model outputs will be inaccurate. An attacker who influences the training can manipulate the model, and in some cases, introduce a backdoor that can be exploited. Therefore, methods are needed that are tolerant to noisy, faulty, or poisoned training data and are able recognize an unreliable ML system.

Preventing Model Poisoning

ML methods are susceptible to poisoning: an attacker can control a fraction of the training set and still influence the behavior of the model. With a need to get as much data as possible for training, it is common to use data from many sources, but this is risky. If even one source of data is malicious, the entire model becomes untrustworthy.

Research is needed on designing robust ML algorithms that adversaries cannot influence. Means are needed to limit the influence on model decisions by a single or even small number of training instances and may require the reformulation of learning optimizations. AI best practices should ensure the end-to-end provenance of data collected and used for training. Data that fall outside the normal input space must be detectable to both mitigate adversarial poisoning and improve the quality of training processes.

Training Calibration, Confidence, and Retraining

ML methods work well when they are used on data that is close to what they were trained on and fail with inputs outside the training parameters. For instance, a self-driving car trained in sunny, cloudy, rainy, and snowy weather might operate poorly in sleet or hail; it cannot recognize situations for which it has not been trained. These problems are common because it is difficult to gather data for all possible situations, including changes over time (e.g., self-driving cars need to detect pedestrians even when clothing styles change). Moreover, systems typically do not recognize suspicious input, even when a human would clearly recognize it as anomalous.

Research is needed to allow models and systems to detect inputs and environments outside their training sets. The goals could be to increase the detection of anomalies, improve algorithms for confidence scores, adopt training methods that amplify rare events, and allow the most effective use of existing training data.

Research is needed on curating training data for retraining. For many ML tasks, modeled phenomena change over time; for example, social media posts used for public sentiment analysis change quickly over time as the vocabulary and topics of interest change. To be effective and accurate, models evaluating social media content must be retrained frequently. Research is needed to develop theory and methods to identify what training data to collect, when such training data is no longer relevant, and how aggressively models should be retrained.

Training Data Privacy and Model Fairness

Many applications require ML training using private data and thus risk leaking sensitive information. For example, recent attacks (called membership attacks) have shown that an adversary can determine whether a data item was used in training a model.

Research is needed to improve the theory and practice of training-data privacy. Recent advances, such as differential privacy, provide new pathways to anonymize data and prevent leaks.

Models will learn whatever biases and discriminatory features are present in training data. For example, if historical data used in the training data reflects discrimination against a given community (e.g., in college admissions or loan approvals), that bias will appear in the target task. Therefore, research is needed into training methods that guarantee all communities will be treated equitably. This will require scientific and technical foundations for fairness be developed in ML. Fairness goals must be defined and then

algorithmic techniques developed to measure, detect, and diagnose unfairness in algorithms and methods used to train ML models.

Engineering Trustworthy AI-Augmented Systems

AI is frequently integrated into a data processing pipeline to address some complex system task. New understanding of the vulnerability of AI models to adversarial action raises questions about the safety of the system in which it is embedded. AI components are often opaque and defy conventional software analysis, but the AI information-decision pipeline can introduce new attack vectors in multiple places, including the following:

- Environments where the AI algorithms operate (e.g., a co-located adversary inserts a Trojan[5] in an open source content management system containing hardware fault attacks.
- Implementations of AI frameworks and applications (e.g., ones that include software bugs and vulnerabilities).
- ML models (e.g., teacher models that include trojans).
- Training data (e.g., untrusted training data lead to poisoning, allowing queries from untrusted parties that lead in turn to evasion or model extraction).

Moreover, some attack vectors may be shared by many applications due to hidden dependencies in the supply chain. Research is needed to develop theory, engineering principles, and best practices on applying AI as a component of a system. This would include:

- AI engineering design principles
- Threat modeling, security tools, and domain vulnerabilities
- Securing human-machine teaming

AI Engineering Design Principles

The problem of evaluating and validating models for cybersecurity is difficult. Good abstractions could provide the ability to know when one defense is better than another, like models that support wargaming and simulation. However, threat models are easy to game, especially if adversaries are deploying AI systems of their own. These models need to enable iterative abstractions of attacks and refinements, be designed in accord with an AI expert, and consider the following:

- Data availability and integrity
- Access control policies and mechanisms
- Network orchestration and operations
- Resolution of competing interests
- Privacy and a dynamic policy environment

Research is needed to develop engineering principles, based on science and community experience, to effectively incorporate AI into technology system development. AI systems are immature, and few security-oriented patterns have been studied, developed, or applied widely. Successful research into engineering components that support AI functionality in a redundancy (e.g., ensemble), supervisory (e.g., doer-checker[6]), or other framework would make AI-enabled systems more trustworthy. Understanding the conditions, threat models, and application domains where such patterns can be applied, and the various parameters that govern their implementation and operation, are necessary but subsidiary goals.

[5] A *Trojan horse* or *Trojan* is a type of malware that is often disguised as legitimate software.

[6] *Doer-checker* means that for each transaction, there must be at least two "individuals": a "doer" and a "checker" necessary for its completion.

Threat Modeling, Security Tools, and Domain Vulnerabilities

Threat models for AI systems are needed that acknowledge the vulnerabilities and sensitivities of AI algorithms and practices. Common and precise adversary definitions—a requirement that has enabled security advances in other fields such as cryptography or network security—are lacking in AI. As discussed elsewhere in this document, the scientific community must acknowledge the capabilities of AI adversaries and then harmonize the AI cybersecurity model with larger system goals, threat models, security apparatuses, and deployment environments.

Once the impacts of AI vulnerabilities on the overall system are understood, traditional cybersecurity and robust system design can reduce attack surfaces created by AI; for example, existing cybersecurity techniques may be used to ensure AI training data is more difficult to poison. In addition, redundant and diverse AI models may reduce overall system vulnerability (e.g., an autonomous vehicle may use AI models based on lidar, radar, and image-processing, along with existing map information, to make self-driving vehicles safer and more reliable). Research into robust system architectures that can withstand AI component failures and attacks also will be important.

Finally, the science and engineering community needs to explore domain-specific techniques to counter attacks against AI models. For example, self-driving cars could include a non-AI-controlled brake system that humans could use as needed to prevent crashes during a cyber-attack. Or in the context of a military AI supply chain system, an attacked system sends the wrong bullets to a remote base, leaving the base vulnerable to attack. In addition, domain-specific bounds and safety defaults need to be developed and enforced, such as upper and lower bounds on an AI-controlled temperature system.

Securing Human-Machine Teaming

As AI technologies become ubiquitous, humans and machines will work together seamlessly in more and more aspects of work and life. Human–machine teaming promises to improve the efficiency and accuracy of critical tasks while maintaining moral and subjective involvement by humans. For example, AI-aided doctors will diagnose illnesses faster and more accurately, security operators will detect and thwart adversaries more effectively, and teachers will better recognize and adapt to students' needs to enhance educational outcomes.

However, such integration presents security challenges. The functionality of either the machine or the human part of these systems can be heightened or degraded by any number of factors. Humans and machines must be able to continually assess the trustworthiness of each other and adjust accordingly. They both need the ability to sense, monitor, and assess each other's performance and to provide indicators of trustworthiness.

Research is needed for managing this trust relationship between human and computational performers. How does either side calibrate trust? Trust calibration is a way to measure situational confidence, and different tasks require different degrees of confidence. Human-in-the-loop applications with high complexity and urgency may require that the machine act autonomously when the task's scale and complexity mean the human cannot respond in real-time.

There will also be instances in human-machine teaming when partners disagree on decisions or approaches to solving a problem. In these cases, there are open research questions on conflict management between the human and AI partner. There are examples where the machine's decisions or assessment is correct, but the human analyst does not agree. Theory and techniques for resolving such conflicts must be available to support complex tasks, in real time, where the information is ambiguous or subjective and when a slow resolution would have grave consequences.

Research is needed to explore trust metrics. For example, a potential metric is neglect tolerance. In a scenario where humans and machines are responsible for multiple tasks, the willingness of the human to ignore the tasks that the machine is focused on indicates the amount of trust the human has in the machine. Metrics of this sort must be developed to understand, measure, maintain, and continually assess trust in human–machine teams.

AI for Cybersecurity

The intent of cybersecurity (and cyber-resiliency) is to enable continued operations of networks and systems in the face of attack and compromise. AI has the potential to substantially advance these goals by increasing awareness and reacting to risks and changes in the environment at near wire speed. Such advances provide an important opportunity to alter the attacker-versus-defender asymmetries in current cybersecurity environments. This section presents a range of avenues to explore these opportunities.

Cyber-resiliency includes self-adaptation and adjustment to attack surfaces in the face of ongoing attacks. In active cyber-attack scenarios, AI systems can enable the identification and creation of strategies that allow defenders to identify an adversary's weaknesses, establish methods to observe, and prepare for future cyber-attack campaigns. Proposed resiliency mechanisms include using AI to categorize various kinds of attacks and inform adaptive responses.

Many attacks target relatively simple errors, such as misconfigurations of systems, that are hidden in a vast amount of correct data. Logic-based AI systems are exceptionally good at noticing these kinds of inconsistencies and knowing how to repair them.

Other attacks may show up as departures from standard usage patterns. These patterns may not be obviously anomalous, can be hidden deep within data streams, and are unlikely to be visible to humans. Though often indescribable by humans, these patterns can be learned by machines and noticed at scale.

It is understood that significant leverage is gained from having a small team of highly skilled cyber defenders protecting networks used by thousands. Using AI could enable similar levels of protection to become ubiquitous while providing the domain experience necessary to address other aspects, such as quality-of-service constraints and degradation-of-system behaviors.

Enhancing the Trustworthiness of Systems

Current AI-based techniques can capture and process the enormous amount of data and complex patterns produced by today's most powerful technology systems. Furthermore, the ability to capture large amounts of data—ranging from operational (such as network traces) to software development (such as code commits and contribution patterns)—provides the data needed to train powerful AI-based systems. With additional research investment that is aligned with cybersecurity priorities, the successes of AI-based reasoning can be used to enhance the trustworthiness of technology systems, where the "system" is understood to include humans as well as machines. Two specific areas where AI shows potential are in the creation and deployment of trustworthy software systems, and in identity management.

AI for Creating and Deploying Trustworthy Software Systems

There are potential defensive uses for AI throughout the software (and hardware) development and operational lifecycles. A promising research direction, sometimes called "big code," involves leveraging AI to detect errors in programs, check best practices, and look for security vulnerabilities. These AI techniques enable program analysis to become more adaptive by automatically recognizing and accommodating various approaches used by software engineering teams to design security into their

systems. In cases where the intended behavior of the program is well defined, AI techniques can even synthesize high-assurance code automatically from formal specifications. For legacy code, AI could infer more formal specifications to automate modernization and security hardening.

AI models are also good for modern development practices in which code evolves quickly. Techniques such as online learning can bring decades of work in system analysis to practical application. A valuable long-term outcome would be the use of AI-based "coding partners" to assist less experienced developers and analysts in understanding large, complex software systems and advise them on the security and robustness of proposed code changes.

AI can also play a role in securely deploying and operating software systems. Once code is developed, AI techniques can automatically explore for low-level attack vectors, or where appropriate, domain and application configuration or logic errors. Similarly, AI can also advise IT professionals on best practices for the secure operation and monitoring of critical systems. Automated configuration advice can secure systems against unsophisticated adversaries, whereas AI-based network monitoring can detect patterns of attack that are associated with more sophisticated nation-state adversaries.

Open-source software development offers a unique setting to apply these AI-based software assurance techniques. With its widespread use by commercial and government organizations, open-source security improvements would be extremely high impact (e.g., an automated system that continually proposes security patches for open source software). At the same time, the public nature of open source development adds new challenges concerning the malicious introduction of functionality and corruption of data by an AI-based agent. This requires further exploration.

AI for Identity Management

Identity management and access control are central to securing modern communication systems and data stores. However, an adversary can compromise many of these systems by stealing relatively small authorization tokens. AI-based identity management can make access-control decisions based on a history of interactions, and it is difficult to circumvent. By characterizing expected behavior, AI techniques can provide protection with more lightweight and transparent mechanisms than current approaches (e.g., two-person authorization requirements for certain actions). AI also can enhance accuracy and reduce threats against biometric authentication systems.

However, there is a downside to using AI for identify management. AI monitoring of behavioral patterns to provide authorization and detect insider threats could enable ongoing privacy violations in the system. Research is needed to push monitoring and decision-making procedures closer to where they are needed, and to use techniques such as differential privacy to limit the scope of privacy violations. These efforts should include both the ethical and technical aspects of identity management and examine the potential for abuse.

Autonomous and Semiautonomous Cyber Action

A new paradigm of cybersecurity is emerging where human expertise is augmented using autonomous systems. One of the first successful applications of AI was spam filtering. Today, basic spam filtering works well, but automated cyber defensive techniques have not seen similar improvements. A framework is needed for combining human insight with effective AI techniques.

AI techniques are likely to be used by attackers as well as defenders. Traditional defensive strategy sought to eliminate vulnerabilities or to increase the costs of an attack. The use of AI could dramatically alter the attack risk and cost equations. Automated systems will need to plan for worst cases and anticipate,

respond, and analyze potential and actual threat occurrences. Research is needed to understand how AI changes the attacker and defender balance of capabilities, and how it alters attack economics.

There are multiple stakeholders involved in cyber defensive scenarios, including data owners, service providers, system operators, and those affected by AI-based decisions. How stakeholders are consulted and informed about autonomous operations and how decision-making is delegated and constrained are important considerations.

Two areas of specific interest are autonomous attacks and mission-specific resilience.

Autonomous Attacks

Cyber defenders will face attacks created and orchestrated by AI systems. At the most basic level, where there is a stable cyber environment, attacks could be constructed using classic deterministic planning. At the next level, where the environment is uncertain, attacks may involve planning under uncertainty.

In the extreme case where minimal information about the environment and defenses is available, the attacker could use autonomous techniques to discover information and learn how to attack and execute plans for cyber reconnaissance. The attacker's challenges include the need to remain stealthy and avoid any deception mechanisms.

The attacker may use AI to develop strategies that include building a model of the victim network or system (i.e., AI-enabled program synthesis). An adversary can systematically generate programs that have a fixed behavior to learn about a cybersecurity product—using it as an oracle. At a high level, the attacker can generate code examples and predict whether the defense technology would detect the attacker's presence as malicious. Using the answers, the attacker can build a model of the cybersecurity product.

Methods and techniques are needed to make deployed systems resistant to automated analysis and attack, by either increasing the cost or continuing to close system loopholes. One promising technique is automated isolation (e.g., behavioral restrictions). Attacks can exploit the universality of program execution because most software components are designed to have limited behavior. Sandboxes have proven effective in protecting software from memory corruption attacks, but more precise methods are needed. There is value in exploring AI systems that learn the scope of valid behaviors and limit components to those behaviors.

Another method is to strategically study defensive agility. How and when should plans and systems be updated? Can results from simulation environments be applied to real systems? What are the principles behind simulating? What is possible, and what is useful?

Mission-Specific Resilience

Many cybersecurity techniques are designed to be broadly applicable. While often beneficial, applying techniques without accounting for the objectives of the enterprise can lead to problems, including failure to meet the mission (whether social, industrial, or military). Domain experts must team with the AI experts to categorize system attacks and model responses in the context of the primary mission of the organization.

Mission-driven AI systems must incorporate the intent of the leader (whether commander, chief information officer, etc.) into any autonomous system that is making security-related decisions, especially when those decisions affect access to and operations of the system. A key research question is how to express the leader's intent. This can start with an operations order expressed in natural language, and when clearly understood, it can inform choices of resiliency adaptations. AI techniques can be used to translate a mission briefing or operations order into something that is addressable by an autonomous

decision system (e.g., dormant attackers may be left alone because rooting them out may be even more disruptive than a possible attack).

Mission-oriented AI can also support planning and execution. For example, an important step in security engineering is to identify the cyber assets (i.e., key cyber terrain[7]) that are vital for mission success, and to realize that these can change as the mission purpose or goals change. AI can help identify and prioritize relevant aspects of the data, computation, information classification, and other security factors. Ongoing adaptation of the AI itself is a part of this evolutionary process.

Conflict between security measures designed for distinct computing resources, whether they are run concurrently or in sequence, is a challenge. For example, one autonomous agent may be working to lay a cyber deception trail to confuse a cyber attacker while another agent may be trying to simplify the network structure to reduce the attack surface.

Autonomous Cyber Defense

As adversaries use AI to identify vulnerable systems, amplify points of attack, coordinate resources, and stage attacks at scale, defenders need to respond accordingly.

Current practice is often focused on the detection of individual exploits, but sophisticated attacks can involve multiple stages—including penetration, lateral motion, privilege escalation, malware staging, and/or persistence establishment—before the ultimate target is compromised. Although modern ML techniques can detect the individual events that constitute this "cyber kill chain," a bottom-up approach that sequentially addresses the various stages of attack is inadequate. Progress requires integration activity at the tactical level into a top-down strategic view that reveals the attacker's goals and current status, and helps coordinate, focus, and manage available defensive resources.

Consider the scenario of an attack on a power distribution system. Initial penetration is accomplished through a phishing email and the initial foothold is on a normal workstation. A larger malware package is downloaded that includes a key logger and a "kill disk" that consumes all the space on the workstation disk. The credentials of a system administrator who logs in to repair the workstation are exfiltrated to the attacker, and the attacker moves to the power grid's operator console, able then to disable the entire distribution network. Any of the individual events described in this scenario would typically be detected. However, the ability to intervene before the network is shut down requires a system to understand the attack plan and use a top-down approach.

A top-down strategic approach would include the following actions:

- Identification of adversarial goals and strategies
- Intelligent adaptive sensor deployment
- Proactive defense and online risk analysis
- AI orchestration
- Trustworthy AI-based defenses

Identification of Adversarial Goals and Strategies

Future AI techniques will have the potential to integrate symbolic and probabilistic reasoning, planning, and ML. At the top level, AI planning techniques can automatically generate a library of attack plans and

[7] *Key cyber terrain*, analogous to key terrain in a military sense, refers to systems, devices, protocols, data, software, processes, personas, or other network entities, control of which provides an advantage to an attacker or defender.

a hierarchical network of goals, subgoals, and actions that collectively could achieve an attacker's desired results. Associated with each attack will be a plan recognizer that receives a stream of events generated by detectors, determines the extent to which those events correspond to a plan, and then predicts events and posits defensive responses related to the attacker staging the attack.

Modern AI planners use techniques where ML is trained on search heuristics tuned to derive a single optimal plan; however, a complete set of attack plans is required. Managing the search combinatorics of plan generation is a major challenge that warrants several possible approaches:

- Use Monte-Carlo techniques to generate a representative subset of attack plans. This requires significant research to measure the coverage of this subset.
- Interleave plan generation and plan recognition. This requires new techniques.
- Assemble and organize the attacker's strategies and tactics and represent this knowledge in an effective way.

Defenders need a complete understanding of their network to reason effectively about an attack. Sophisticated attackers move "low and slow" to evade detection. At any time, cyber detectors will signal multiple events, some spurious, others real but unrelated to a major attack being staged. Each of these events causes the plan recognizer to hypothesize about the attacker's intent and progress. Given bounded resources, systems cannot keep all these hypotheses active forever, even when attacks may be staged over a period of many months. How many of these hypotheses should be maintained, for how long, and what heuristics should be used?

Intelligent Adaptive Sensor Deployment

Detectors will need to be integrated into the top-down context of a plan-recognition process. At any point, based on observations, it is more likely that certain attack plans are in progress and less likely that others are in progress. This top-down context could provide bias and choose events to observe. Research is needed to explore the intelligent and adaptive deployment of sensors that integrates event visibility and resource budgets.

Proactive Defense and Online Risk Analysis

As adversarial plan recognition proceeds in scenarios such as that described above, the defender gains confidence that the attacker is pursuing a plan aimed at compromising the defender's system. At this stage, the attacker still needs to achieve subgoals of the attack plan to accomplish the ultimate goal, and the defender still can take actions to prevent the final attack. These defenses might be costly (e.g., shutting down certain machines that provide useful services) or inconvenient (e.g., raising the level of protection in a firewall) and thus require a cost-benefit assessment. Reasoning needs to be automated (with possible human-in-the-loop supervisors) because events are proceeding "at cyber speed."

AI Orchestration

While the use of AI and ML systems improves the performance of individual cybersecurity tools, coordination and orchestration between multiple tools becomes increasingly important. How can these systems cooperate to achieve mission objectives? Successful mission execution may require that models be built to include interactions that involve the goals and objectives of other systems, their cybersecurity tools, and the intent and state of mind of humans in the environment.

Trustworthy AI-Based Defenses

Broad deployment of AI-based decision systems also creates new attack vectors that must be understood. For example, an AI system that analyzes user behavior to provide access to secured resources could be fooled by adversary-provided data, or attackers could corrupt training data to cause poor AI decision-

making. Research is needed to defend against attacks that target the AI system itself. Specifically, for the security domain, work in formal specification and reasoning to capture desired system properties (e.g., communication patterns, application logic, or authorization frameworks) can be leveraged to check AI-based decisions against these requirements and explicit assumptions. AI systems that produce evidence for their decisions and explanations for their techniques would strengthen AI-based defenses.

Predictive Analytics for Security

Cybersecurity may benefit from predictive analytics that process information signals (both internal and external) to a computer system to predict and assess the likelihood of a successful attack on the system.

Initial work has developed techniques for identifying adversary planning or reconnaissance activity early in the cyber operation's lifecycle from data streams (such as dark web traffic) or distributed logs of cyber-relevant activity. Other work has begun to identify patterns and linkages among disparate datasets that tie together the cyber and human domains, taking advantage of *a priori* knowledge (e.g., from classified sources) to automatically augment, discover, and track new activities and campaigns.

Further research is needed to find indicators of adversary intent, capability, and motivation, especially when correlated with signals that track the defensive posture of the system, including the cyber capabilities of human operators. The goal should be not only to predict the likelihood of an attack and whether it might be successful or not, but also to discover new indicators of adversary cyber activities. These indicators can be used to protect sources and methods via parallel story construction and may provide insights to support defensive resilience as attack methods change over time.

Focus areas include data sources, operational security, and adaptation over time.

Data Sources

AI systems require clean, labelled training data for supervised applications. Obtaining real data to train systems for predictive analysis can be challenging. There are a variety of options, including:

- Extend the techniques for learning with less labeled data in order to leverage smaller datasets.
- Develop methods and tools to capture and curate training data that is resilient to data poisoning by adversaries.
- Identify and evaluate novel signals from unconventional massive, diverse, and noisy data streams to help recognize early phases of cyber-attacks that become leading indicators of the main attack phases.
- Generate genuinely realistic synthetic training data that encompasses a wide range of modes and domains.

Operational Security

When diverse datasets and automated analytics are used to monitor, track, and actively counter adversary cyber activities, false flag or misdirection operations can lead to misattribution or even collateral damage. Due to the malleability of cyber indicators, AI analysis for cyber defense may require a higher standard of independent-source verification and confidence scores than when the same or similar techniques are applied to other intelligence problems. Research is needed to develop methods to perform multimodal analysis of combined datasets; provide cross-validation of the analysis and datasets; and identify risks, potential flaws, or gaps in the reasoning.

Correspondingly, AI offers the opportunity to learn and mimic activities based on the analysis of adversary or normal traffic, including activities to reduce or eliminate operator error. For some contexts, a "human-in-the-loop" structure will be necessary; for others, it is likely that a less rigid structure (e.g.,

"human-on-the-loop") will be enough.[8] In both cases, a better understanding of the reasoning behind AI-recommended actions will provide more confidence in the outcomes.

Adaptation Over Time

Is it possible to predict the cybersecurity of a large system over time? Such analysis might consider the internal state of the system (including how regularly patches are applied), which security controls are in place, and the cyber hygiene of the human operators. Combining this with situational awareness that includes current activities and goals of potential adversaries would be beneficial.

The desired output would hypothesize potential cyber operations, including metrics that characterize and prioritize the goals of the adversaries, the threats based on the probability of an attack, and the likelihood of success. It would leverage explainable ML methods to give the rationale for predictions and identify exploitable weaknesses.

Applications of Game Theory

Game-theory models can be useful for understanding adversarial attack plans and reasoning about potential defenses. There has been significant research in this area, but more is required due to the misplaced assumption that an adversary's actions are observable or easily probed. In fact, attacker and defender visibility is so poor that game-theory models that assume they have near-perfect information are inadequate.

In cybersecurity settings, the "game" can change quickly due to adversarial actions (e.g., a new attack tool or capability), a shifting game environment, players with different incentives, or irrational players. Also, equilibrium concepts may not make sense, and optimality concepts will need to be derived to apply noncooperative game theory to cybersecurity.

Two areas of game theory are explored here, cooperative and evolutionary game theory and multi-agent modeling, and using AI for understanding cybersecurity games.

Cooperative and Evolutionary Game Theory and Multi-Agent Modeling

Noncooperative game-theory models are appropriate for modeling many different cybersecurity scenarios; however, there may be instances where different players (e.g., coalition partners) need to cooperate to achieve their goals against an adversary. In some networks it may make sense to treat collections of assets as coalitions, or to consider cooperative orchestration of multiple AI systems (e.g., among different Internet service providers) and teams of AI experts.

In such cooperative environments, game theory combined with multi-agent system modeling approaches could plan and generate potential attacks and reason about their impact. In cyber environments, attackers and defender's plans are continually evolving as their objectives, capabilities, and constraints change. This coevolution must be accounted for to effectively model cybersecurity applications.

Additional research is needed on uncertainty planning in a mixture of cooperative and noncooperative environments. This should also address, in the context of human-machine teaming, how multimodal information is incorporated for more-effective decision support.

[8] The distinction between "human-in" and "human-on" the loop is based on whether humans make key decisions ("in the loop") or whether humans ("on the loop") guide the overall system direction but leave specific actions to an AI system.

Using AI for Understanding Cybersecurity Games

Conversely, game-theory models must assume certain attacker capabilities, incentives, etc. AI systems can be developed to analyze cybersecurity data and extract important game-theory model parameters (e.g., systematically query cybersecurity products to extract their methods for malware detection). By analyzing data related to attacker tools, AI could provide adversarial modeling including capabilities and incentives. Probabilistic modeling using AI tools may help assess the security of a system (i.e., the extent to which defenses will protect the system against a specific set of threats).

Game-theory models can be dual-use. It is possible that a model can be used for cyber offense and cyber defense. More research is needed to model offense and defense scenarios where there is significant uncertainty, equilibrium is not optimal, attacker action visibility is poor, and the game's action space and assumptions are constantly evolving.

Human-AI Interfaces

Coordination between cybersecurity systems that use AI is increasingly important as they become more complex and threats grow more severe. Human cybersecurity "systems" must also be considered. Humans will continue to interact closely with AI systems, both as operators and as end users, and coordination and trust between humans and AI will be essential to optimize system effectiveness.

Insufficient coordination between systems increases the attack surface and the potential for serious system misbehavior. This can occur in AI systems ranging from enterprise IT to self-driving cars. Problems arise when individual system components maximize their own goals without consideration of system-level objectives. Attackers can induce a module to behave in a manner that is locally optimal but globally pathological.

Moreover, in an era of social-information warfare, hybrid approaches are necessary that account for both technological and human perspectives. Research is needed on the orchestration of mixed human-machine contexts that are vulnerable to information that is misinformed, misattributed, or manipulated, and that could result in bad decisions (whether made by human or by machine). Human-machine teaming, building trust between systems and humans, and providing decision-making assistance are three important areas to consider.

Human-Machine Teaming

Because humans and AI systems have very different failure modes, it can be advantageous to leverage both unique human and unique AI capabilities as part of a robust decision-making process. As AI cybersecurity systems team closely with humans, either in-the-loop or on-the-loop, several research areas emerge, depending on the communication paradigm.

Whether humans are the system drivers or simply guide the system, AI systems need to be designed so humans can understand, trust, and explain the AI decisions to others. This will require that humans provide goals, feedback, and data, which are often difficult to collect and label. Non-AI experts will need to be trained to interact with AI systems and provide well-formatted, well-presented, and relevant data. Consider nonverbal human cues; how can they be made AI-consumable?

When AI cybersecurity systems are deployed at scale, what will the failure modes be? Today, AI is often used to automatically lock down suspicious activity, allowing a human time to determine if an activity is allowed. This is acceptable, even if users are inconvenienced, because AI systems are still limited in scope. However, if AI is interwoven into critical systems such as the electrical utility grid, could these automated

"just-in-case" actions be too widespread, too disruptive, or too dangerous? Research is needed on how to incorporate humans in the decision-making loop based on a consequence–impact analysis.

In semiautomated systems, human latencies can be a challenge. Should there be ways to slow down AI systems to accommodate humans in-the-loop? This could be a defensive disadvantage (i.e., too many decision-makers) that reduces agility. However, this could also be an opportunity; AI systems that support a "slow mode" for humans could swap out failing components with human interventions. It should also be noted that some kinds of decisions do not involve conscious processing and can be faster than current machine processing.

In a diverse environment of multiple humans and AI systems, how can interactions be managed and governed to reduce human error and increase safety? Can human-AI interaction expand the applicability of AI by supplying "cybersecurity training wheels"? What is the societal, moral or legal framework where humans can be held responsible for AI system outcomes?

Building Trust Between Systems and Humans

Both operators and users must establish a level of trust in their systems. Stakeholders who impact the adoption and use of the system must understand its operation (i.e., end users are able to interpret that their network connection was cut off in response to an inferred attack). Trust requires that humans can identify a system's state and predict its behavior under various circumstances. Trust allows users to use the system and to continue using it as the system evolves or exhibits unusual behavior. This requires human-readable, rule-based specifications based on approximating system behavior.

Cognitive and other biases should be considered, but the goal is the best possible human-machine interaction. Explanations will need to be at the right level of abstraction and designed to produce the appropriate human response. Both over-trust and under-trust could lead to ineffective interaction between the human and the system. Over-trust, for example, could lead operators to be reluctant to overrule misbehaving systems; under-trust could lead to the abandonment of otherwise effective systems.

Decision-making Assistance Attacking Humans

AI systems will often be deployed in workflows that involve human actors. This extends the "attack surface" to not only the technical components but the humans as well. Bad information can corrupt both AI and human decision-making. Research literature cites AI systems that can generate extremely convincing fake video and audio that humans will trust. Therefore, research must be extended to include support for decision-making assistance. This could include training to inoculate human operators against data falsification attacks, and models that can both defensively predict failure modes and enable AI systems to adapt when humans make erroneous decisions.

Science and Engineering Community Needs

The advancement of AI and its realization within the public and private sectors is dependent on initiatives supporting adoption of science and engineering related to AI. There are several key needs that require support, as summarized below.

Research Testbeds, Datasets, and Tools

To facilitate AI community standards and metrics for deploying and securing future AI systems, investment is required in research testbeds and datasets. Protection mechanisms against comprehensive threats need evaluation, particularly when applying AI to critical application domains (e.g., autonomous vehicles

or medical diagnosis) or to cybersecurity applications (e.g., intrusion detection and network defense). The lack of testbeds and datasets for these critical domains prevents researchers from making progress on effective defenses. Funding opportunities should include the creation and maintenance of realistic simulation environments and datasets in diverse application domains beyond image classification.

The complexity of the AI threat landscape matches that of the AI systems deployed. Testbeds and datasets that evaluate capabilities and defenses in a comprehensive, principled, and sustainable manner need to be understood, developed, and validated. They should be developed in modular ways to facilitate answering questions across different disciplines (e.g., security, machine learning, control and cyber-physical systems, and computer engineering). This would enable abstractions to isolate subproblems whose solutions can be integrated into larger, more complete solutions. For example, autonomous vehicles will integrate many AI-enabled subsystems (e.g., object detectors, path planning, and coordination) that feed into the control system. Defenses need to protect against specific attacks but also be integrated and evaluated within the larger system's capabilities and constraints. Furthermore, when individual subsystems are under attack, other systems should be evaluated as to how they can contribute to the overall defense.

Testbeds must be open source, widely available, and developed according to best practices that foster collaboration and reproducibility. They should include simulators, emulators, and datasets that will ultimately enable the development of metrics, benchmarks, and evaluation methodologies that result in the creation of safer and more secure AI systems. In addition, testbeds must facilitate education efforts and evolve as new methodologies and technologies emerge. They must both foster innovation and continuously reevaluate cross-layer interaction (e.g., defenses that span hardware, software, control, and algorithmic techniques).

Education, Job Training, and Public Outreach

The ability of the science and engineering community to address the challenges and reap the benefits of AI is dependent on fostering an informed public. Education and outreach efforts should focus on the usefulness, limitations, best practices, and potential dangers of this technology.

AI could be integrated into both primary and secondary educational curricula as well as into scientific centers of education within university systems. These educational centers need to bring together existing disciplines of computer science, data science, engineering, and statistics to foster the necessary workforce to expand the use of AI into the future. Educational communities nationwide could consider the teaching of AI in the accreditation of programs.

Retraining the existing workforce in the use and practice of AI will also be necessary. Possible strategies include the development of adult education programs, open online courses, and professional certifications, possibly working with universities and private sector professional organizations. These programs would help to educate the public and counter widespread misunderstandings about AI and its benefits, limits, and dangers. Public initiatives could focus on consumer-oriented views of AI and be tailored to reach across demographically diverse communities.

Conclusion

The insights in this document were gathered from a diverse set of scientific and engineering experts and suggest that the future of AI and ML will be influenced by the Nation's balanced stewardship of AI's benefits and challenges, particularly in the area of cybersecurity. The authors hope that Federal organizations find this information helpful.

Please note that these discussions represent viewpoints from a single moment in time. The rapid advances in technology, new application domains, and the interplay between ML, AI, and cybersecurity will introduce new opportunities and challenges. Many of the areas of discussion will remain relevant for years, but it will be important to view them through the lens of evolving circumstances. As such, the national (and global) thinking about these issues is expected to change over time, and these questions and insights will need to be reviewed, revisited, and updated periodically.

Abbreviations

AI	artificial intelligence
BD	big data
HPC	high performance computing
IT	information technology
IWG	interagency working group
ML	machine learning
NITRD	Networking and Information Technology Research and Development Program
R&D	research and development

About the Authors

The NITRD Program is the Nation's primary source of federally funded coordination of pioneering IT R&D in computing, networking, and software. The multiagency NITRD Program, guided by the NITRD Subcommittee of the NSTC Committee on Science and Technology Enterprise, seeks to provide the R&D foundations for ensuring continued U.S. technological leadership and meeting the Nation's needs for advanced IT.[9]

The Artificial Intelligence R&D IWG coordinates Federal R&D in AI. It also supports and coordinates activities tasked by the National Science and Technology Council's Select Committee on AI and Subcommittee on Machine Learning and Artificial Intelligence. This vital work promotes U.S. leadership and global competitiveness in AI R&D.[10]

The Cyber Security and Information Assurance IWG coordinates Federal R&D to protect information and information systems from cyber threats. This R&D supports the security and safety of U.S. information systems that underpin a vast array of capabilities and technologies in many sectors, including power generation, transportation, finance, healthcare, manufacturing, and national security. [11]

Acknowledgments

NITRD's Artificial Intelligence R&D, and Cyber Security and Information Assurance IWGs gratefully acknowledge Patrick McDaniel, Pennsylvania State University; John Launchbury, Galois; Brad Martin, National Security Agency; Cliff Wang, Army Research Office; and Henry Kautz, National Science Foundation who helped plan and implement the workshop and write and review the report. Also, we gratefully acknowledge the workshop participants and the other NSTC Subcommittees for their contributions to the report.

[9] https://www.nitrd.gov

[10] https://www.nitrd.gov/nitrdgroups/index.php?title=AI

[11] https://www.nitrd.gov/nitrdgroups/index.php?title=CSIA

ARTIFICIAL INTELLIGENCE AND CYBERSECURITY: OPPORTUNITIES AND CHALLENGES

TECHNICAL WORKSHOP SUMMARY REPORT

A report by the

NETWORKING & INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE

and the **MACHINE LEARNING & ARTIFICIAL INTELLIGENCE SUBCOMMITTEE**

of the

NATIONAL SCIENCE & TECHNOLOGY COUNCIL

MARCH 2020

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at http://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at http://www.whitehouse.gov/ostp.

About the Networking and Information Technology Research and Development Program

The Networking and Information Technology Research and Development (NITRD) Program is the Nation's primary source of federally funded coordination of pioneering information technology (IT) research and development (R&D) in computing, networking, and software. The multiagency NITRD Program, guided by the NITRD Subcommittee of the NSTC Committee on Science and Technology Enterprise, seeks to provide the R&D foundations for ensuring continued U.S. technological leadership and meeting the Nation's needs for advanced IT. More information is available at https://www.nitrd.gov/about/.

About the Machine Learning and Artificial Intelligence Subcommittee

The Machine Learning and Artificial Intelligence (MLAI) Subcommittee monitors the state of the art in machine learning (ML) and artificial intelligence (AI) within the Federal Government, in the private sector, and internationally to watch for the arrival of important technology milestones in the development of AI, to coordinate the use of and foster the sharing of knowledge and best practices about ML and AI by the Federal Government, and to consult in the development of Federal MLAI R&D priorities. The MLAI Subcommittee reports to the NSTC Committee on Technology and the Select Committee on AI.

About this Document

On June 4-6, 2019, the NSTC NITRD Program, in collaboration with NSTC's MLAI Subcommittee, held a workshop to assess the research challenges and opportunities at the intersection of cybersecurity and artificial intelligence. The workshop brought together senior members of the government, academic, and industrial communities to discuss the current state of the art and future research needs, and to identify key research gaps. This report is a summary of those discussions, framed around research questions and possible topics for future research directions. More Information is available at https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019.

Acknowledgements

The National Science Technology Council's NITRD and MLAI Subcommittees gratefully acknowledge Patrick McDaniel, Pennsylvania State University; John Launchbury, Galois; Brad Martin, National Security Agency; Cliff Wang, Army Research Office; and Henry Kautz, National Science Foundation who helped plan and implement the workshop and write and review the report. Also, we gratefully acknowledge the workshop participants for their contributions to the report.

Table of Contents

Executive Summary

On June 4-6, 2019, the National Science and Technology Council Subcommittees on Networking and Information Technology Research and Development, and Machine Learning and Artificial Intelligence held a workshop[1] to assess the research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). This document summarizes the workshop discussions.

Technology is at an inflection point in history. AI and machine learning (ML) are advancing faster than society's ability to absorb and understand them; at the same time, computing systems that employ AI and ML are becoming more pervasive and critical. These new capabilities can make the world safer and more affordable, just, and environmentally sound; conversely, they introduce security challenges that could imperil public and private life.

Though often used interchangeably, the terms AI and ML refer to two interrelated concepts. Coined in the 1950s, AI is the field of computer science that refers to programs intended to model "intelligence". In practice, this refers to algorithms that can reason or learn given the necessary inputs and base knowledge and are used for tasks such as planning, recognition, and autonomous decision-making (e.g., weather prediction). ML is a specialized branch of AI that uses algorithms to understand models of phenomena from examples (i.e., statistical machine learning) or experience (i.e., reinforcement learning). Throughout this document the term AI will be used to discuss topics that apply to the broad field, and ML will be used when discussing topics specific to machine learning.

The challenges are manifold. AI systems need to be secure, which includes understanding what it means for them to "be secure." Additionally, AI techniques could change the current asymmetric defender-versus-adversary balance in cybersecurity. The speed and accuracy of these advances will enable systems to act autonomously, to react and defend at wire speed,[2] and to detect overt and covert adversarial reconnaissance and attacks. Therefore, securing the Nation's future requires substantial research investment in both AI and cybersecurity.

AI investments must advance the theory and practice of secure AI-enabled system construction and deployment. Considerable efforts in managing AI are needed to produce secure training; defend models from adversarial inputs and reconnaissance; and verify model robustness, fairness, and privacy. This includes secure AI-based decision-making and methods for the trustworthy use of AI-human systems and environments. This will require a science, practice, and engineering discipline for the integration of AI into computational and cyber-physical systems that includes the collection and distribution of an AI corpus—including systems, models and datasets—for education, research, and validation.

For cybersecurity, research investments must apply AI-systems within critical infrastructure to help resolve persistent cybersecurity challenges. Current techniques include network monitoring for detecting anomalies, software analysis techniques to identify vulnerabilities in code, and cyber-reasoning systems to synthesize defensive patches at the first indication of an attack. AI systems can perform these analyses in seconds instead of days or weeks; in principle, cyber-attacks could be observed and defended against as they occur. However, safe deployment will require understanding the multiple dimensions and implications of these AI actions.

[1] https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019

[2] *Wire speed* is the rate of data transfer that a telecommunication technology provides at the physical level (hardware wire, box, or function) and that supports the data transfer rate without slowing it down.

Abbreviations

AI	artificial intelligence
IT	information technology
ML	machine learning
MLAI	Machine Learning and Artificial Intelligence Subcommittee (Subcommittee of the NSTC)
NITRD	Networking and Information Technology Research and Development (Program or Subcommittee of the NSTC)
NSTC	National Science and Technology Council
OSTP	Office of Science and Technology Policy

Introduction

The National Science and Technology Council (NSTC) Networking and Information Technology Research and Development (NITRD) Subcommittee and the NSTC Machine Learning and Artificial Intelligence (MLAI) Subcommittee, held a workshop to assess the research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI). The workshop, held June 4–6, 2019, brought together senior members of the government, academic, and industrial communities. The participants discussed the current state of the art, future research needs, and key research and capability gaps. This document is a summary of those discussions. For more details, including the agenda, please go to: https://www.nitrd.gov/nitrdgroups/index.php?title=AI-CYBER-2019.

The document is divided into three topic areas: Security of AI, AI for Cybersecurity, and Science and Engineering Community Needs. Developing a specific structure or prescriptive task list for this pressing domain is outside the scope of the workshop effort. Such a determination and resulting plan will require substantial effort across many organizations over many years.

Security of AI

Recent advances in AI are transformative and already exceed human-level performance in tasks like image recognition, natural language processing, and data analytics. Economic factors will drive the adoption of new AI applications that disrupt almost every aspect of the enterprise both good and bad.

AI-systems can be manipulated, evaded, and misled resulting in profound security implications for applications such as network monitoring tools, financial systems, or autonomous vehicles. Therefore, secure and resilient techniques and best practices are vitally important.

Specification and Verification of AI Systems

Integrated AI systems involve four components: perception, learning, decisions, and actions. These systems operate in complex environments that require each component to interact and be interdependent (e.g., errors in perception can cause an incorrect decision). Furthermore, there are unique vulnerabilities in each of the components (e.g., perception is prone to training attacks while decisions are susceptible to classic cyber exploits). Finally, the notion of correctness is not a purely logical matter; noise and uncertainty require bounds for each component to protect the system from misbehaving.

There is a pressing need for formal methods to verify AI and ML components, both independently and in concert, as it relates to logical correctness, decision theory, and risk analysis. New techniques are needed that specify what a system is expected to do and how it should respond to attack. In traditional systems, qualities that match the specification are tractable for each component. Because AI systems are so complex, their implementation and configuration are difficult to assess. Research is needed in architectural structures and analysis techniques that allow verification of these components and is part of a larger effort to develop manageable standards, best practices, tools, and methods to reason about the behavior of a system.

A new discipline and science of AI architecture could produce an AI “building code”. Such a code could come from theory and experience, capture best practices, and leverage guidelines from other computer

science areas. Analysis of the building code would lead to a better understanding of AI mechanisms and move the field forward.

Specification and verification must also address aspects such as performance, security, robustness, and fairness. Research is needed to better understand performance tradeoffs, the operating environment, and may require a domain expert on the team. And finally, an engineer must be identified to implement, deploy, and maintain the AI system.

Trustworthy AI Decision Making

As AI systems are deployed in high-value environments, the issue of ensuring that the decision process is trustworthy, particularly in adversarial scenarios, is paramount. While there are numerous illustrations of ML vulnerabilities, science-based techniques to predict trustworthiness are elusive. Research is needed to develop methods and principles for a wide array of AI systems, including ML, planning, reasoning, and knowledge representation. Areas that need to be addressed for trustworthy decision making include defining performance metrics, developing techniques, making AI systems explainable and accountable, improving domain-specific training and reasoning, and managing training data.

Threat model research must identify measurable properties that define trustworthiness so a defender can incorporate robustness, privacy, and fairness into decision-making algorithms. Given a specific threat model, the system will have to reason about adversarial interference and define requisite conditions to achieve these trustworthiness properties. Possibilities include adapting definitions from cryptography or computer security, unifying properties into a single reasoning framework, and treating them as variants of a single notion of (in)stability in ML and AI for both decision making and for security models more broadly.

Research is also needed in methods for understanding the learned reasoning of AI methods, particularly deep learning. How do certain data points influence the optimization procedures, and the reasoning, involved in ML systems? Possibilities include analysis of the optimization procedure, or the AI system outcome, if it captures both the training data and the learning method. Techniques that can estimate a training point's influence on individual predictions could also become the basis to assess the relevance of a model in a decision environment.

In ML, there are approaches emerging that provide decision guarantees using a variety of techniques (e.g., convex relaxation of the adversarial optimization problem and randomized smoothing). However, the approaches are currently focused almost exclusively on supervised learning and are difficult to achieve without degrading system performance. A related area of research, AI systems that request guidance when they are uncertain, can improve trust in the eventual decision and allow the system to obtain information for future decision making.

The accuracy of AI is also domain sensitive. Security vulnerabilities arise when training data is not representative of the given environment. Conversely, overly pessimistic vulnerability assessments can occur if constraints in the application domain are not considered. Research is needed on how input data is acquired, secured, maintained, and evaluated within domain-specific AI environments, and as they become a part of the full-use ecosystem. An autonomous vehicle system is trained with images and situations acquired from realistic environments and maintained constantly as its environment changes. Perception, planning, reinforcement learning, knowledge representation, and reasoning are all domain-specific vulnerabilities that need to be considered. This includes reasoning about streaming data, weighing consequences (e.g., causing a car to crash or go in the wrong direction), and adapting to

unanticipated events (e.g., weather or road construction). Domain specificity research necessitates a rethinking of threat models and helps deploy and maintain AI systems in real-world environments.

Researchers must also evaluate the cost/benefit ratio of collecting, protecting, and storing training data. Datasets are valuable (e.g., large network datasets can reveal everything about network vulnerabilities). Proper collection and storage can protect data and provide information for defense. But what if the data is of higher value for an adversary, should it be collected?

Detection and Mitigation of Adversarial Inputs

While AI performs well on many tasks, it is often vulnerable to corrupt inputs that produce inaccurate responses from the learning, reasoning, or planning systems. There are examples where deep learning methods can be fooled by small amounts of input noise crafted by an adversary.[3] Such capabilities allow adversaries to control the systems with little fear of detection. As systems based on deep networks and other ML and AI algorithms become integrated into operational systems, it is critical to defend against adversarial inputs by considering more robust machine learning methods, AI reconnaissance prevention, the study of adversarial models, model poisoning prevention, secure training procedures, data privacy, and model fairness.

Efforts are needed to harden learning methods against adversarial inputs. This problem is well understood in both the statistics and technical communities. Both theoretical and empirical research are needed to make the same advances for deep learning and modern ML methods without sacrificing performance or accuracy.

Modern AI systems are vulnerable to reconnaissance where adversaries query the systems and learn the internal decision logic, knowledge bases, or the training data. This is often a precursor to an attack to extract security-relevant training data and sources or to acquire the intellectual property embedded in the AI. The following are possible reconnaissance prevention measures that need research:

- Increase the attacker workload and reduce their effectiveness through model inversion.
- Leverage cybersecurity approaches, including rate limiting, access controls, and deception.
- Study the impacts on accuracy and other aspects of algorithms and systems.
- Design reconnaissance-resistant algorithms and techniques.
- Integrate resistance into learning and reasoning optimizations.
- Embed security guarantees into the model using new multistep techniques.
- Expose the presence and goals of the attacker using the cybersecurity honeypot[4] concept.

The vulnerability of an AI system is defined by the adversary's knowledge and capabilities. Research is needed to classify the different types of attacks and develop appropriate defenses. Defenses need to address attacks based on the type of information the attacker has access to. These models should be carefully mapped, attack and defense strategies identified, and special research attention given to security critical domains where ML models are most at risk. (e.g., autonomous vehicles and malware detection).

[3] There are many articles available on this topic, for example: Adversarial Attacks and Defenses: A Survey; https://arxiv.org/abs/1810.00069.

[4] A honeypot is a network-attached system set up as a decoy to lure cyberattacks and to detect, deflect or study hacking attempts in order to gain unauthorized access to information systems.

AI and ML models learn how to characterize expected inputs from training data. If the training instances do not represent all possible and future situations, then the model outputs will be inaccurate. This creates a security scenario where an attacker can manipulate the model and introduce an exploitable backdoor. An adversary can control a fraction of the training set and still influence the behavior of the model (model poisoning). ML requires as much data as possible and it is common, but also risky, to use many data sources. If even one source of data is malicious, the entire model becomes untrustworthy. To both mitigate adversarial poisoning and improve training processes, AI best practices must ensure the end-to-end provenance of training data and the detection of data that falls outside the normal input space.

ML methods work well when they are used with similar data to what they were trained on and fails when the data is different (e.g., a self-driving car trained in sunny, cloudy, rainy, and snowy weather might operate poorly in sleet or hail). These are common problems because it is difficult to acquire data for all possible situations. Systems typically do not recognize abnormal data, even when a human would. The research goal is to increase the detection of anomalies, adopt training methods that amplify rare events, and allow the most effective use of existing training data and algorithms. To remain effective and accurate, ML models must be retrained frequently (e.g., social media terminology used for public sentiment analysis changes over time as vocabulary and topics of interest change). Research is needed to identify what training data to collect, when such training data is no longer relevant, and how often models should be retrained.

Recent attacks have shown that an adversary can determine whether a data item was used in training a model. Because many applications require ML training using private data, this puts sensitive information at risk. Further research is needed, but advances, such as differential privacy, provide new pathways to anonymize data and prevent leaks.

Finally, models will learn whatever biases and discriminatory features are present in training data. If the data reflects discrimination against a given community (e.g., in college admissions or loan approvals), that bias will appear in the outcome. Prevention of outcome bias will require scientific and technical foundations for ML fairness to be developed. Goals must be defined, and algorithmic techniques developed to measure, detect, and diagnose unfair ML training data and methods.

Engineering Trustworthy AI-Augmented Systems

New understanding of how vulnerable AI components are to adversarial action raises concerns about the safety of the entire data processing pipeline in which they are used. AI components defy conventional software analysis and can introduce new attack vectors in environments where the AI algorithms operate, implementations of AI frameworks and applications, ML models, and training data. Due to hidden dependencies in the pipeline, multiple applications can be effected. Research is needed to develop theory, engineering principles, and best practices when using AI as a component of a system. This should include threat modeling, security tools, domain vulnerabilities, and securing human-machine teaming. These models need to enable iterative abstractions of attacks and refinements, be designed in accord with an AI expert, and consider data availability and integrity, access controls, network orchestration and operation, resolution of competing interests, privacy, and a dynamic policy environment.

To make AI-enabled systems more trustworthy, engineering principles should be based on science, community experience, and AI component functionality research that includes redundancy (e.g.,

ensemble), supervisory (e.g., doer-checker[5]), and other frameworks. Understanding the conditions, threats, domains, and constraints are necessary but subsidiary goals.

Once overall system AI vulnerabilities are understood, traditional cybersecurity and robust system design can reduce the impact (e.g., to ensure AI training data is more difficult to poison); allow more redundancy and diversity to be built in (e.g., an autonomous vehicle may use lidar, radar, image processing, *and* map information); develop robust system architectures that can withstand AI component failures and attacks; and explore domain-specific counter measures, bounds, and safety defaults (e.g., self-driving cars with a human-driven back up braking system or an AI-controlled temperature system with upper and lower bounds).

As AI technologies become ubiquitous, humans and machines will work together seamlessly to improve the efficiency and accuracy of critical tasks (e.g., helping doctors diagnose illnesses or teachers adapting to individual students' needs). The challenge is that the machine or the human's functionality can be heightened or degraded by many factors. Further research is needed to help both machine and human to sense, monitor, and assess each other's performance and trustworthiness. What if a human cannot respond fast enough in a critical, time-sensitive, human-in-the-loop application? What if the machine and human's results disagree? Theory, techniques, and metrics are needed to support complex decisions, in real time, where the information is ambiguous or subjective, and when a late response could have grave consequences.

AI for Cybersecurity

Just as AI-systems need innovative cybersecurity tools and methods to improve their trustworthiness and resiliency; cybersecurity can use AI to increase awareness, react in real time, and improve its overall effectiveness. This includes self-adaptation and adjustment in the face of ongoing attacks that alter the current attacker-versus-defender asymmetries. Strategies that identify an adversary's weaknesses, use observation methods, and gather lessons learned, can use AI to categorize various kinds of attacks and inform adaptive responses (e.g., find inconsistencies quickly and know how to repair them) at scale.

It is understood that a small team of expert cyber defenders can effectively protect networks used by thousands. The use of AI could extend that same level of system protection, make it ubiquitous, and also provide the domain knowledge necessary to address aspects such as quality-of-service constraints and degradation-of-system behaviors.

Enhancing the Trustworthiness of Systems

AI technologies can capture and process the enormous amount of data produced by today's technology systems. In turn, this ability provides the training data needed to drive AI-system innovation and development. AI-based reasoning, aligned with cybersecurity priorities, could make both fully automated and human-in-the-loop systems more trustworthy. Two potential areas are the creation and deployment of more reliable software systems and identity management. Promising research involves leveraging AI to detect errors in programs, check best practices, identify security vulnerabilities, and make it easier for software engineers to design security into their systems.

[5] *Doer-checker* means that for each transaction, there must be at least two "individuals", a "doer" and a "checker", necessary for its completion.

In modern development practices, code often evolves quickly. The use of AI-based "coding partners" to assist less-experienced developers and analysts in understanding large, complex software systems, and advise them on the security and robustness of proposed code changes, would be valuable. AI can also assist in securely deploying and operating software systems. Once code is developed, AI can be used to detect low-level attack vectors, inspect for domain and application configuration or logic errors, provide best practices for secure system operation, and monitor networks. Open-source software development offers a unique and high-impact opportunity for AI-based security improvements due to its widespread use by commercial and government organizations. However, due to its public nature, open source is vulnerable to malicious actions by an AI-based adversary.

Another promising area of AI use is identity management and access control. Adversaries can compromise many techniques simply by stealing authorization tokens. An AI-based system could use a method based on a history of interactions and expected behavior that is also lightweight, transparent, and difficult to circumvent. For biometric authentication systems, AI could enhance accuracy and reduce threats. However, AI monitoring of behavioral patterns could lead to privacy violations. Further research is needed to develop methods that consider both the ethical and technical aspects, and the potential for abuse of AI-assisted identity management.

Autonomous and Semiautonomous Cybersecurity

Unlike other successful AI applications (e.g., spam filtering), AI is likely to be used by both attackers and defenders in cyber defensive scenarios. The traditional strategy based on eliminating vulnerabilities or increasing the cost of an attack changes with the addition of AI. Both autonomous (independent of human action) and semiautonomous (human-in-the-loop) systems will need to plan for worst cases and anticipate, respond, and analyze potential and actual threat occurrences. There are multiple stakeholders affected by AI-based decisions, including data owners, service providers, and system operators. How stakeholders are consulted and informed about autonomous operations and how decision making is delegated and constrained are important considerations.

Cyber defenders will likely face autonomous attacks at several levels: in a stable cyber environment, attacks could use classic deterministic planning; where the environment is uncertain, attacks may involve planning under uncertainty; when little is known about the environment, the attacker could use AI to obtain information, learn how to attack, execute reconnaissance, and develop strategies that include a model of the victim network or system (i.e., AI-enabled program synthesis) and the cybersecurity product.

Methods and techniques are needed to make deployed systems resistant to autonomous analysis and attack. Promising techniques include automated isolation (e.g., behavioral restrictions), defensive agility (i.e., using simulations and updates to strengthen defenses), and mission-specific strategies (e.g., use of domain experts to categorize attacks and responses). Mission-driven AI systems must always incorporate the organization leader's intent into any security-related decisions (e.g., access to and operation of the system). A key research question is how to express the leader's intent. AI techniques can translate a mission briefing or operations order into something that is addressable by an autonomous decision system (e.g., dormant attackers may be left alone because rooting them out may be even more disruptive than a possible attack).

AI can also support the mission planning and execution involved in security engineering. AI can be used to identify the cyber assets (i.e., key cyber terrain[6]) that are vital for mission success, and to realize that these can change as the mission purpose or goals change. It can help identify and prioritize relevant aspects of the data, computation, information classification, and other security factors including the ongoing adaptation of the AI itself. One challenge is to orchestrate security measures designed for distinct computing resources so that their decisions do not conflict.

Autonomous Cyber Defense

As adversaries use AI to identify vulnerable systems, amplify points of attack, coordinate resources, and stage attacks at scale, defenders need to respond accordingly. Current practice is often focused on the detection of individual exploits, but sophisticated attacks can involve multiple stages before the ultimate target is compromised. Progress requires a top-down strategic view that reveals the attacker's goals and current status, and helps coordinate, focus, and manage available defensive resources.

Consider the scenario of an attack on a power distribution system. A phishing email is opened on a normal workstation; a malware package is downloaded; credentials of a system administrator who logs in to repair the workstation are acquired; the attacker moves to the power grid's operator console; the entire distribution network is disabled. Any of the individual events can be detected, but the ability to intervene before the network is shut down requires a top-down strategic approach. That strategy would include identification of adversarial goals and strategies, intelligent adaptive sensor deployment, proactive defense and online risk analysis, AI orchestration, and trustworthy AI-based defenses.

AI planning techniques can generate attack plans and a network of goals, subgoals, and actions that disclose an attacker's strategy. Each attack will have a plan recognizer that receives sensor data, predicts events, and posits defensive responses. AI is trained on search heuristics to derive a single optimal plan; however, a complete set of attack plans is required. Managing plan generation is a major challenge that warrants several possible approaches: use Monte Carlo[7] techniques to generate a representative subset of attack plans; interleave plan generation and plan recognition; and effectively represent the attacker's strategies and tactics. Other considerations include the efficient storage and maintenance of hypotheses and heuristics, and the integration of intelligent and adaptive sensors/detectors to help establish the top-down plan-recognition process.

Using a top-down strategic approach to the power distribution scenario means that a plan is generated when the attack is still in its early stages and allows the defender to take actions to prevent the shutdown. These defensive actions might be costly (e.g., shutting down certain machines that provide useful services) or inconvenient (e.g., raising the level of protection in a firewall) and thus require a cost-benefit assessment. Reasoning needs to be automated (with possible human-in-the-loop supervisors) because events are extremely time sensitive.

As ML and AI systems improve the performance of individual cybersecurity tools, coordination and orchestration between multiple tools becomes increasingly important. Successful execution may require that models include interactions with other systems. These systems may involve different goals and objectives, cybersecurity tools, and intent and state of mind of human actors.

---

[6] *Key cyber terrain*, analogous to key terrain in a military sense, refers to systems, devices, protocols, data, software, processes, personas, or other network entities, control of which provides an advantage to an attacker or defender.

[7] Monte Carlo (MC) methods are a subset of computational algorithms that use the process of repeated random sampling to make numerical estimations of unknown parameters.

Predictive Analytics for Security

Cybersecurity will benefit from predictive analytics that process information (both internal and external) to assess the likelihood of a successful attack. Initial work has developed techniques for identifying adversarial operations early in the attack's lifecycle by using data streams (such as dark web traffic) or distributed logs of cyber-relevant activity. Work has also begun to identify patterns and linkages among datasets that tie together the cyber and human domains, taking advantage of *a priori* knowledge (e.g., from classified sources) to augment, discover, and track new activities and campaigns. Further research is needed to uncover adversary intent, capability, and motivation of human operators, especially when a system's defenses are being tracked. Beyond just detection and the success/failure factor, information about attacks can help protect sources and methods and provide new insights to improve resilience over time. Focus areas include data sources, operational security, and successful adaptation.

Obtaining the clean, labeled, real data required for predictive analytics is challenging. Some options include lowering the "labeled" threshold to leverage smaller datasets; capturing and using poisoning-resilient data; identifying new cyber-attack early-warning signals using unconventional data streams; and making synthetic training data more realistic.

When diverse datasets and AI analytics are used to monitor, track, and counter cyberattacks, false flags[8] can lead to misattribution or even collateral damage. Therefore, AI analysis for cyberattacks may require a higher standard of validation than other intelligence problems. Research is needed to perform multimodal analysis; cross-validation; and identify risks, potential flaws, or gaps in the data sets or the reasoning.

AI analysis can also provide new insights that help reduce operator error in both human-in-the-loop and human-on-the -loop[9] contexts, provide more confidence in the outcomes, and help large systems adapt over time. Such analysis might consider the internal state of the system, how regularly patches are applied, what security controls exist (including the human operators), and the level of situational awareness. The analysis would provide scenarios that characterize and prioritize the adversaries' goals, threat level, and likelihood of success and include the prediction's rationale and identify the exploitable weaknesses.

Applications of Game Theory

There has been significant research into game-theory models that can be used to understand attack plans and reason about potential defenses. But because an adversary's actions are still not easily observable, and information is not perfect, more research is needed. In cybersecurity settings, the "game" can change quickly due to adversarial actions (e.g., a new attack tool or capability), a shifting game environment, players with different incentives, or irrational players. Also, equilibrium[10] concepts

[8] A false flag cyberattack is when a hacker or hacking group stages an attack in a way that attempts to fool their victims and the world about who's responsible or what their aims are.

[9] The distinction between "human in" and "human on" the loop is based on whether humans make key decisions ("in the loop") or whether humans ("on the loop") simply guide the overall system direction.

[10] Equilibrium is a concept within game theory where the optimal outcome of a game is where there is no incentive to deviate from their initial strategy

may not make sense, and optimality concepts will need to be derived to apply noncooperative game theory[11] to cybersecurity.

Noncooperative game-theory models are appropriate for modeling many different cybersecurity scenarios; however, there may be instances where different players (e.g., coalition partners) need to cooperate to achieve their goals against an adversary. In some networks it may make sense to treat collections of assets as coalitions, or to consider cooperative orchestration of multiple AI systems (e.g., among different Internet service providers) and teams of AI experts.

Additional research is needed on uncertainty planning in a mixture of cooperative and noncooperative environments. This should also address, in the context of human-machine teaming, how multimodal information is incorporated for more effective decision support. Conversely, game-theory models must assume certain attacker capabilities and incentives. By analyzing data related to attacker tools, AI could provide adversarial modeling including capabilities and incentives. Probabilistic modeling using AI tools may help assess the security of a system (i.e., the extent to which defenses will protect the system against a specific set of threats).

Game-theory models can be dual use. It is possible that a model can be used for cyber offense and cyber defense. More research is needed to model offense and defense scenarios where there is significant uncertainty, equilibrium is not optimal, attacker action visibility is poor, and the game's action space and assumptions are constantly evolving.

Human-AI Interfaces

As threats grow more complex and severe, not only is coordination between AI-cybersecurity systems important, but coordination and trust between human-AI interfaces becomes critical. From enterprise IT to self-driving cars, problems arise when individual system components maximize their own goals without consideration of system-level objectives. Attackers can induce a module to behave in a manner that is locally optimal but globally pathological. Moreover, in an era where information can be misinformed, misattributed, or manipulated, good decision making requires hybrid approaches that leverage and orchestrate the unique human and AI capabilities and perspectives. Human-machine teaming, building trust between systems and humans, and providing decision-making assistance are three important research areas to consider.

Human-machine teaming needs to be designed so humans can understand, trust, and explain the outcomes. Users must be trained to supply goals, feedback, and well-formatted and relevant data, and to know where they fit in the decision-making process. Research is needed on how to incorporate humans to maximize outcomes and minimize latency and negative consequences. AI is often used to automatically shut down suspicious activity to allow time for human decision making. Will this still work as AI is applied to critical systems such as the electrical utility grid, where even a short shutdown could be extremely widespread, disruptive, or dangerous? One solution would be to slow AI systems to accommodate humans in the loop. This would reduce agility, but it could also allow humans to intervene and replace failing components.[12] In a diverse human-AI system environment, interactions must be managed with a goal to reduce human error, increase safety, and provide accountability.

[11] https://www.sciencedirect.com/topics/computer-science/noncooperative-game

[12] Note: some decisions that do not involve conscious processing can be faster than current machine processing.

Stakeholders who adopt and use an AI system must understand and trust its operation. The right level of trust requires that humans can identify a system's state and predict its behavior under various circumstances. Over trust could lead to a reluctance to overrule a misbehaving system; under trust could lead to the abandonment of an otherwise effective system. Determining the right level of trust requires human-readable, rule-based specifications based on approximating system behavior, and consideration of cognitive and other biases.

Research literature cites AI systems that can generate extremely convincing fake video and audio that humans will trust. Research must include decision-making assistance such as training human operators to withstand data falsification attacks, and AI-models that can predict failure modes and adapt when humans make erroneous decisions.

Science and Engineering Community Needs

Research Testbeds, Datasets, and Tools

To establish the AI community standards and metrics required to safely deploy future AI systems, more investment is needed in research testbeds and datasets. Threat detection mechanisms must be tested and evaluated for critical AI application domains (e.g., autonomous vehicles, medical diagnosis) to incentivize adoption. Possibilities include the creation and maintenance of realistic simulation environments and diverse domain-specific datasets.

The complexity of both the AI system and the AI-threat landscape require testbeds and datasets that evaluate capabilities and defenses in a comprehensive, principled, and sustainable manner. They should be modular (to facilitate use across different disciplines) and open source; foster innovation, collaboration, and reproducibility; and continually reevaluate cross-layer interaction.

Education, Job Training, and Public Outreach

Education and outreach efforts should focus on fostering the necessary workforce and developing an informed public that understands the usefulness, limitations, best practices, and potential dangers of AI technology. AI should be integrated into primary, secondary, and university education that brings together the disciplines of computer science, data science, engineering, and statistics. The teaching of AI should be considered as part of the accreditation process.

Conclusion

This document reflects information gathered from a diverse set of scientific and engineering experts and suggests that the future of AI rests on the Nation's ability to balance AI's benefits and challenges, particularly in the area of cybersecurity.

Please note that these discussions represent viewpoints from a single moment in time. The rapid advances in technology, new application domains, and the interplay between ML, AI, and cybersecurity will continue to introduce new opportunities and challenges. As such, the national (and global) thinking about these issues is expected to change over time, and these questions and insights will need to be reviewed, revisited, and updated periodically.