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.