AI-Cybersecurity Workshop Briefing to the NITRD and MLAI Subcommittees

Patrick McDaniel, Penn State University
John Launchbury, Galois

July 31, 2019

Executive overview …

- On June 4-6 2019, the National Science and Technology Council's (NSTC) Networking and Information Technology Research and Development (NITRD) Subcommittee, in collaboration with the NSTC's Machine Learning and Artificial Intelligence (MLAI) and the Special Cyber Operations Research & Engineering (SCORE) Subcommittees, held a workshop to:
  - Assess the current and future research challenges and opportunities at the intersection of cybersecurity and artificial intelligence (AI).
  - Bring together senior members of the government, academic, and industry.
  - Discuss the current state of the art and future research needs in this space and identified key research gaps.
- This briefing is a summary of those discussions and the refinement of the broader themes framed as a collection of discussions around research questions and future research directions.

Workshop Objectives

Focus: Assess the key research challenges and opportunities in the interplay between cybersecurity and artificial intelligence

1. The use of AI/ML to improve cybersecurity
2. The improvement of security of AI/ML

Goal: Produce a **specific** and **actionable** written report, refined from participant discussions

1. Highlight the applications and opportunities for research between AI and cybersecurity

Day 1 AM	Opening
	Keynote: Yan Shoshitaishvili (Arizona State University)
	Breakout: AI for Secrity
Day 1 PM	Keynote: Michael Kearns (University of Pennsylvania)
	Breakout: Security of AI
Day 2 AM	Keynote: Úlfar Erlingsson (Google))
	Breakout: Broader questions
Day 2 PM	Keynote: David Wagner (University California, Berkeley)
	Breakout: Security of AI
Day 3 AM	Keynote: Dean Souleles (Office of the Director of National Intelligence)
	Writing Sessions / Poster Session

AI for Security

- AI and ML provide unique new opportunities for improving security, enhancing our situational awareness, and protecting citizens rights to privacy, fairness, ….
- Why now? AI/ML* can ….
  1. Improve speed of reasoning
  2. Improve speed of reaction, human latency vs system latency
  3. Scale in data: Reasoning over large amounts of data
  4. Scale personnel: force multiplier. For Cyber security defenders, thousands of people may be defended by 2 skilled cyber security defenders.
  5. See patterns, connections, and where people can't ..

*We will use the term AI when referring to the broad field of techniques and systems employing machine reasoning, and ML when referring to more narrow field of algorithms and practice of statistical or other machine learning.

AI4Sec Research Area 1: Enhancing Trustworthiness of Systems

- AI for Creating and Deploying Trustworthy Systems
  - Enhance or replace expensive systems analysis to detect errors and vulnerabilities and check best practices, and evaluate security coverage.
  - Aid the design or coding of systems using AI/ML based “coding partners”
  - Aid or automate the deployment and operation of security systems
  - Continuous automated ethical red-teaming of deployment environments

- AI for Identity Management
  - Harden identity mgmt. by integrating other biometrics and historical profiles
  - Support the detection and mitigation of insider threats

AI4Sec Research Area 2: Resilient and Autonomous Cyber Action

- AI for autonomous attacks and defenses
  - Development of automated offensive capabilities to perform reconnaissance, plan and execute attacks, and to yield advantage from attacks
  - Development of methods to recognize automated (and non-automated) reconnaissance and attacks, and to counter in optimal manner. This has connections to real operational security as well as game theory.
- Mission-specific resilience and
  - Autonomy can develop means to connect the high level mission goals with the actions of the agent (i.e., understanding the leader's intent)
  - All autonomy can also consider other missions' goal and resource requirements, and develop means to reason about conflicting goals and resource needs

AI4Sec Research Area 3: Attacker-oriented Cyber Defense

- Discovering an Attacker Mission and Techniques
  - Develop techniques to identify attacker's goals and strategies from multiple actions and historical data – serve as an input to defensive planning
  - Countering deception and search-space explosion in complex environments
- Designing, Deploying, and Managing Taskable Recognizers
  - Develop techniques to iteratively refine hypothesis attacker goals and strategies through refinement and adaptive sensor deployment
- Proactive Defense and Online Risk Analysis
  - Countermeasures must be orchestrated through complex set of AI/ML components, working adaptatively in real time
  - Defenses must provide human-interpretable intelligence and command capabilities

AI4Sec Research Area 4: Predictive Analytics for Security

- Predictive Analytics
  - AI/ML offers the ability to probabilistically predict adversarial action and intent in ways that may fundamentally alter current security asymmetries
  - Develop methods to identify and counter deception or misdirection operations that can lead to misattribution or even collateral damage
  - Develop and refine understandability techniques that will explain the “how” and the “why” of predicted adversarial actions and campaigns
- Analytic Data Source Management
  - Develop methods and tools to continuously capture and curate training data that is resilient to adversary attempts to introduce bias will help to mitigate against potential counter-attack by data poisoning

Future Research Challenge – Security of AI

- NITRD National AI R&D Strategic Plan

  "Since the 2016 release of the *National AI R&D Strategic Plan*, there has been rapid growth in scientific and societal understanding of AI security and safety. Much of this new knowledge has helped identify new problems: it is more evident now how AI systems can be made to do the wrong thing, learn the wrong thing, or reveal the wrong thing, for example, through adversarial examples, data poisoning, and model inversion, respectively. Unfortunately, technical solutions for these AI safety and security problems remain elusive."

- Why now?
  - Recent work has shown that AI/ML is fragile in adversarial settings.
  - Almost all defenses have failed or show to be weak?
  - AI/ML can increasingly be used to by pass human and cyber-defenses or subvert existing systems.
  - Economic and capability drivers forcing technology adoption faster than our ability to understand risks.

Sec4AI Research Area 1: Specification and Verification of AI Systems

- Techniques for AI System Specification and Validation
  - Pressing needs for metrics of AI security (not just accuracy)
  - Must be based on sound formal guarantees, but may be statistical in nature
  - Practical, unambiguous, reproducible, and understandable specifications that can be tested at acceptance and deployment time

- Verification of AI Systems
  - Standards, tools, and methods to tractably reason about the correctness and behavior of a system
  - Clear practices that enable operational verification of systems after training and/or deployment by non-experts

Sec4AI Research Area 2: Trustworthy AI Decision-Making*

- Techniques for Trustworthy Decision-Making
  - Metrics and measurement techniques for identifying the quality of a decision or decision process is needed to weigh risks, e.g., robustness, fairness, privacy …
  - Rigorous understanding of trustworthiness in in both traditional AI (planning, path-finding) as well as advanced ML such as deep learning is essential
- Explainability / accountability
  - Future AI systems must be able to identify the structural reasoning or training inputs that led to outputs and behaviors
  - Failure to provide explainability will result in a crisis of accountability in future deployments -- making testing, debugging, and defending AI difficult or impossible
- Domain Specific Training and Reasoning
  - A new science and engineering discipline is require to allow AI/ML to be deployed and calibrated to new problem domains

*Ensuring that future AI decision processes driving autonomous or semi-autonomous action are trustworthy is paramount.

Sec4AI Research Area 3: Trustworthy Machine Learning

- Robust Machine Learning Methods
  - AI and ML algorithmic refinements must aim to limit adversarial manipulations and ensure training data privacy, model fairness
  - Preventing AI Reconnaissance – new science is needed to understand how to control the amount of information leakage from a AI system while maintaining accuracy
- Detection and Mitigation of Adversarial Inputs
  - A new science of detecting or limiting the effectiveness of these adversarial inputs
  - Detection and mitigation may happen within the AI (model) or within the system or environment in which it is used; thus defenses for the entire AI/ML pipeline are needed
- Secure Training
  - A new science of AI/ML training is needed to preventing adversarial model poisoning
  - Science and best practices must be developed to ensure secure AI/ML training calibration, training confidence, and retraining (in non-stationary phenomenon)

Sec4AI Research Area 4: Engineering Trustworthy AM/ML-Augmented Systems

- AI/ML Engineering/Design principles
  - A new science and set of engineering practices and principles are needed to ensure the secure integration of new methods
  - Design patterns, testing strategies and architectural, resilience, and recovery patterns that involve positioning additional components around the AI functionality
  - Develop domain centric threat modeling, security tools are required for future safe systems
- Securing Human-Machine Teaming
  - A new (multi-discipline) science is needed to understand how joint AI/ML systems can maximize performance, accuracy, and speed
  - Fostering human/computation trust and conflict resolution in teaming is essential

Science and Engineering Community Needs

- Testbeds, datasets and tools
  - Support the development and foster the widespread sharing of tools, datasets, and test environments for AI/ML broadly and security in specific
- Education
  - AI is not part of the basic curriculum in primary schools or many university
  - Encourage AI/ML to be part of accreditation of schools
  - Incentivize schools to adopt AI/ML based curriculum
- Job Training
  - Existing engineers and scientists need training, we need a national strategy to retrain the community to best practices

Status

- Workshop: June 4-6 2019
  - Attendees 72 total: 21 government, 35 academia (includes 12 students), 8 industry, with 8 Committee/Staff.
- Report: In final revision
- Availability: Public, expected mid-August 2019
- Contact: Alex Thai, NITRD/NCO (thai@nitrd.gov)

Organizational Leadership

- Program Chairs
  - Patrick McDaniel, Penn State University
  - John Launchbury, Galois
- Planning Committee Co-Chairs
  - Brad Martin, NSA
  - Cliff Wang, ARL
  - Henry Kautz, NSF
- Advisory Committee
  - Kamie Roberts, NITRD
  - Chuck Romine, MLAI
- National Science and Technology Council (NSTC) subcommittees:
  - NITRD (Networking & Information Technology Research and Development)
  - MLAI (Machine Learning & Artificial Intelligence)
  - SCORE (Special Cyber Operations Research & Engineering)

Contact: Patrick McDaniel (mcdaniel@cse.psu.edu)

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov