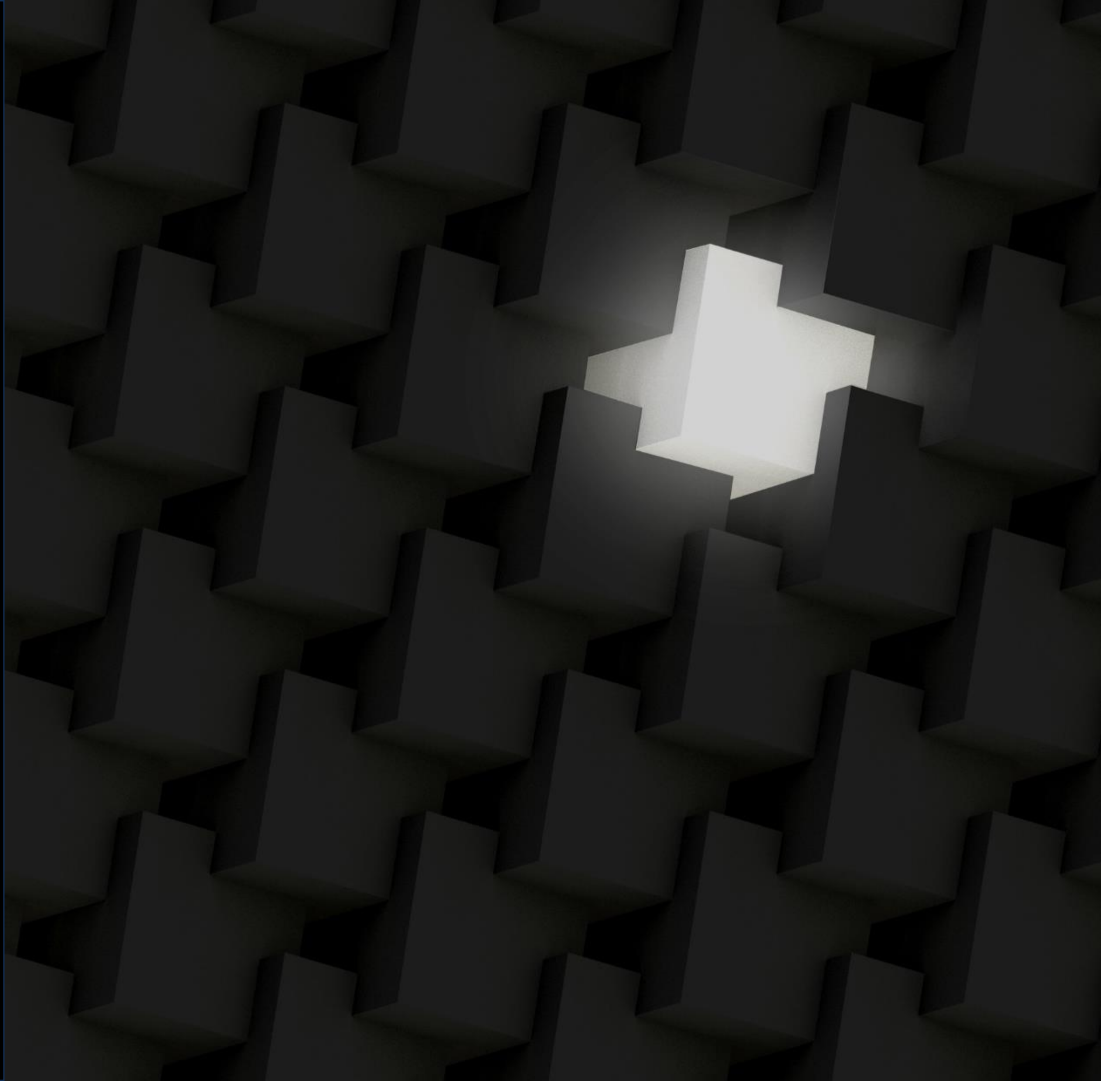
Carnegie Mellon University
Software Engineering Institute

RESEARCH REVIEW 2020

AIDE: Artificial Intelligence Defense Evaluation

Dr. Shing-hon Lau

Dr. Grant Deffenbaugh



This material is based upon work funded and supported by the Department of Homeland Security under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM20-0909

Acknowledgements

The Software Engineering Institute (SEI) CERT Division conducted the the Artificial Intelligence Defense Evaluation (AIDE) project as a funded program, not as LSI or LENS work.

The SEI team extends its thanks to the Department of Homeland Security (DHS) Cybersecurity and Infrastructure Security Agency (CISA) for its generous funding of this important project.

Why Are Organizations Turning to AI Defenses?

Cyber Risks, Speed of Attacks Increasing

—Association of the United States Army, February 25, 2018

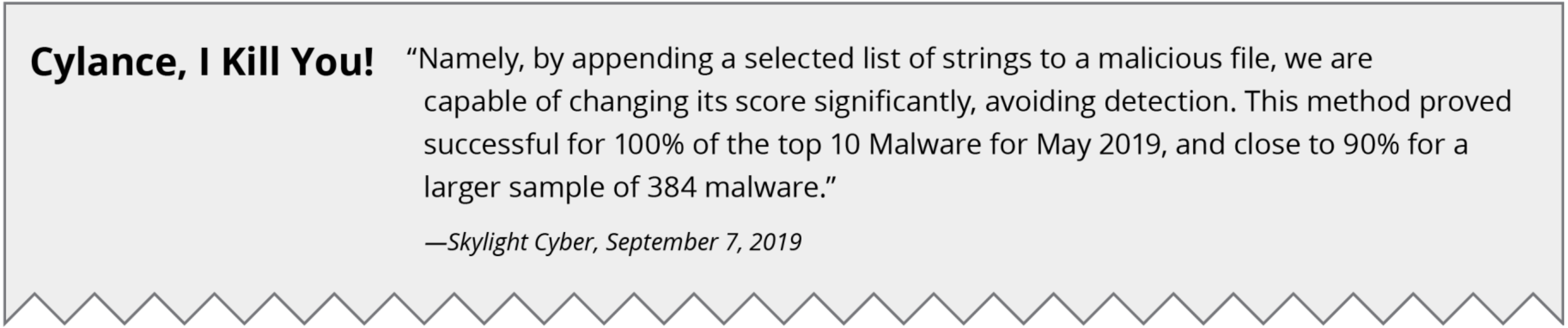
- To augment analyst pool
  - Expected shortfall of 3.5M cybersecurity staff by 2021 (Varonis blog, March 29, 2020)
  - AI can act as significant force multiplier
  - AI can address “easy” alerts, freeing human analysts to handle harder problems
  - AI may be able to catch threats an analyst may not

The Untold Story of NotPetya, the Most Devastating Cyberattack in History

—Wired, August 22, 2018

- To counter the speed of operation of cyberattacks
  - NotPetya attack took down an entire Ukrainian bank in 45 seconds
  - Human reaction to the threat is slow: the damage can be irreversible

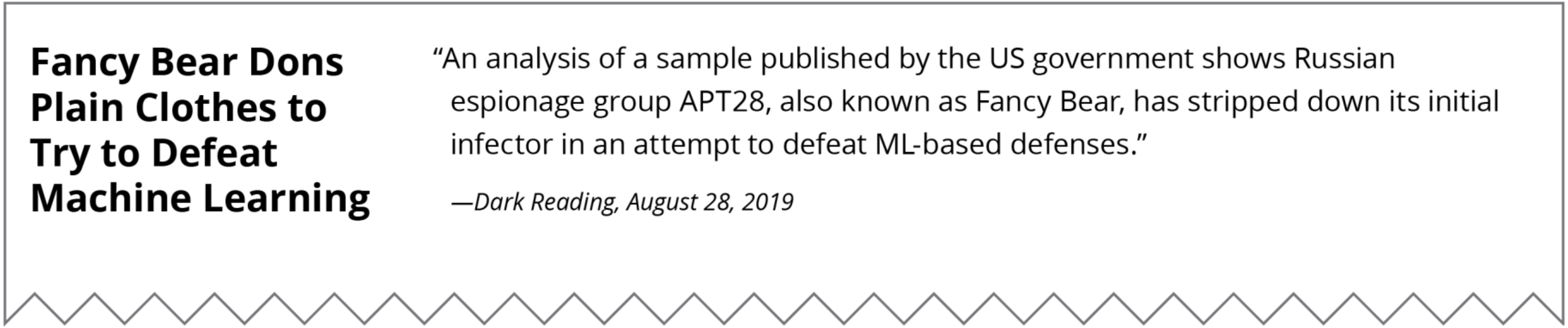
What Happens When AI Goes Wrong?



Cylance, I Kill You!

"Namely, by appending a selected list of strings to a malicious file, we are capable of changing its score significantly, avoiding detection. This method proved successful for 100% of the top 10 Malware for May 2019, and close to 90% for a larger sample of 384 malware."

—Skylight Cyber, September 7, 2019



Fancy Bear Dons Plain Clothes to Try to Defeat Machine Learning

"An analysis of a sample published by the US government shows Russian espionage group APT28, also known as Fancy Bear, has stripped down its initial infector in an attempt to defeat ML-based defenses."

—Dark Reading, August 28, 2019

Purpose of AIDE Project

- Assessment capabilities dictate the testing of our own defenses
- It is public knowledge that our adversaries are working to evade AI defenses
- Our assessors also need this information to accurately gauge our own defenses

Problem

How can we understand the capabilities of AI defenses to detect malicious network activity and how can those defensive capabilities be bypassed?

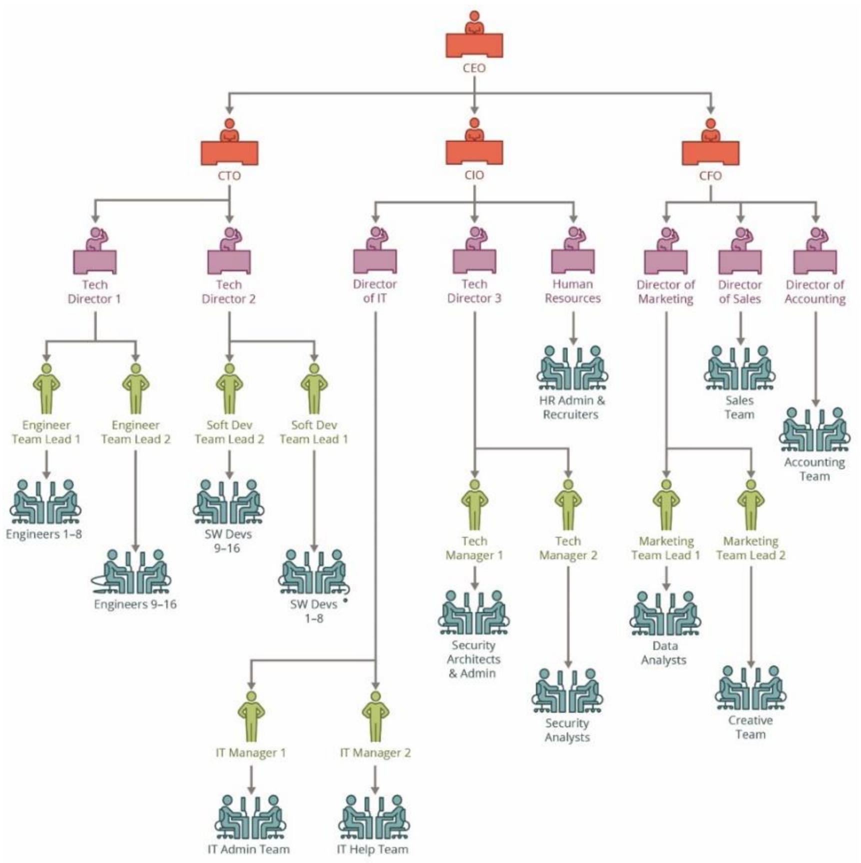
Approach

Develop a comprehensive testing methodology for AI defenses to identify their capabilities and the ways they can be bypassed

High-Level Methodology

1. Create a test network environment
2. Procure and install commercial-off-the-shelf AI network defenses
3. Train defenses on normal network traffic
4. Run baseline test against an AI-based network defense
5. Analyze activities caught by AI-based network defense
6. Conduct obfuscation test
7. Conduct data poisoning test
8. Conduct combined data poisoning and obfuscation test

Fictional Institution Organization Chart



- 99 employees
- 5 divisions
- 3 levels of management

- Each user is provided a unique behavior
  - Customized work schedules
  - Role-specific work tasks
  - Hobbies that influence personal use
- Privileges and access set by role

- SEI GHOSTS software used to simulate user behavior

Static vs Learned Rules

Static rule example:

If <HOST> send <Web Traffic> to <Outside of Company> then [ALERT]

Learned rule examples:

If <HOST> sends <Any Traffic> to <hosts not communicated in last 10 days> then [ALERT]

If <HOST> sends <Unusual Traffic> to <hosts it rarely communicates with> AND <hosts are not considered “safe”> then [ALERT]

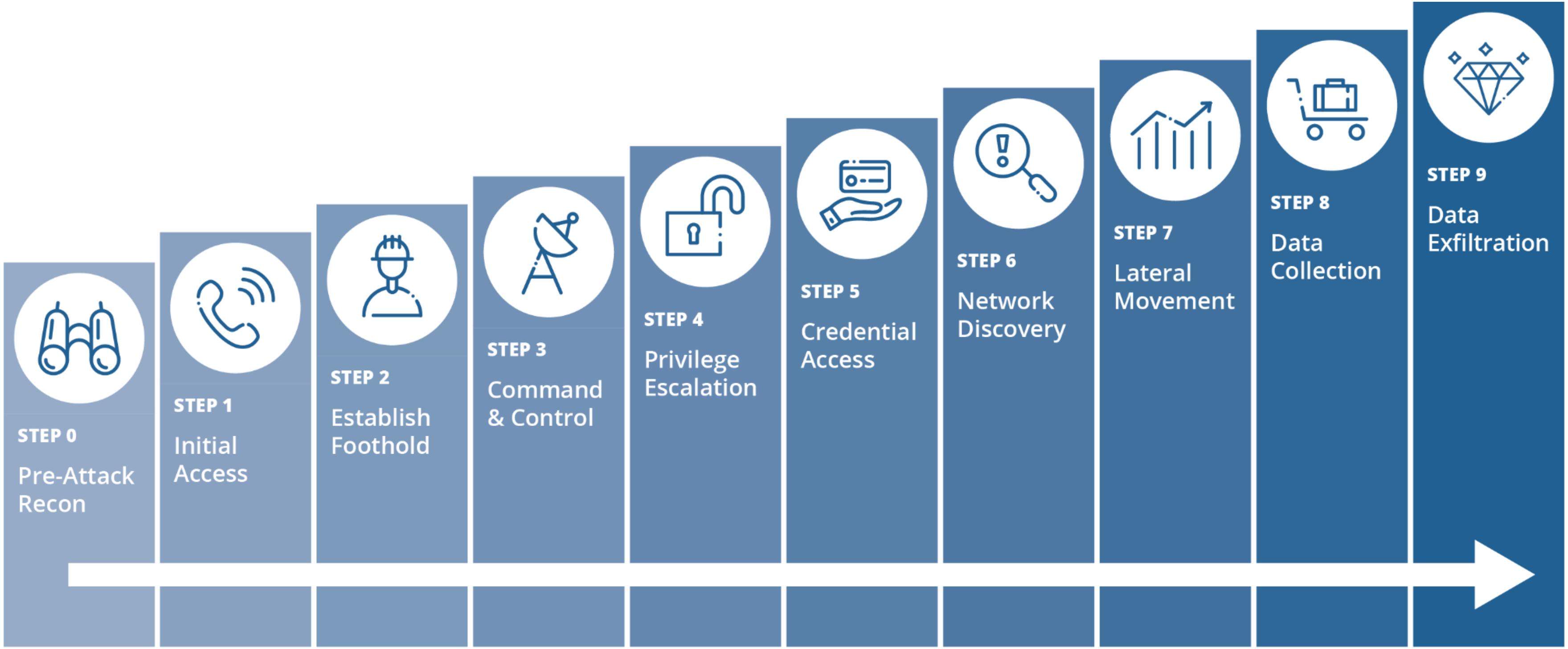
Training and Testing AIs

- Trained AIs on normal background network traffic for 1 month, the amount of time the vendors claim is required to fully train our AIs

- Took a snapshot of the AIs after the end of training and used that snapshot for all evaluations

- Continued the same normal background traffic during testing, with the addition of the traffic generated by the test

Baseline Evaluation Attack


STEP 0

Pre-Attack Recon

STEP 1

Initial Access

STEP 2

Establish Foothold

STEP 3

Command & Control

STEP 4

Privilege Escalation

STEP 5

Credential Access

STEP 6

Network Discovery

STEP 7

Lateral Movement

STEP 8

Data Collection

STEP 9

Data Exfiltration

Finding: AIs were capable of detecting baseline attack

Obfuscation of Attacks

- Identify possible obfuscations based on domain knowledge:
  - Utilizing different tools or techniques that may not generate the same type(s) of traffic
  - Going sufficiently “low and slow”
  - Performing attack steps using machines of users that may perform similar activities as part of their job duties
- Execute modified attack path to determine whether AI is still capable of detecting it
- Finding: Obfuscating attack path allowed bypass of AI defense

Data Poisoning

- Introduce “poisoned data” slowly by performing benign activities that generate traffic similar to the traffic generated during the attack
- Cause AI to “mis-learn” and think that attack traffic is normal background traffic
- Must take care to ramp up poisoning slowly to avoid detection
- Finding: Data poisoning enabled attack path to bypass AI defense

Summary of Findings

- The AIs we examined were capable of detecting our baseline attack path
- We were able to evade detection through both obfuscation and data poisoning techniques

Future directions for AIDE

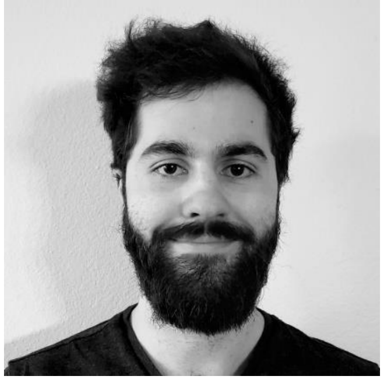
- Increase realism of traffic and system environment
- Evaluate using replay traffic in addition to generated traffic
- Expand the range of attacks considered in our methodology and increase the number of AIs evaluated
- Evaluations over time for assurance of performance
- Extension of testing and evaluation methodology to other types of AI systems
- We are actively seeking collaboration opportunities

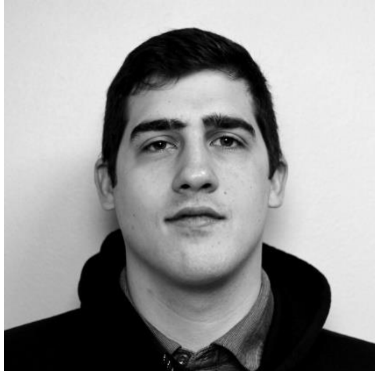
AIDE: Artificial Intelligence Defense Evaluation Team

Dr. Shing-hon Lau

Dr. Grant Deffenbaugh

Chesleah Kribs





Brandon Marzik

Reggie Savoy

Alec Woods