Carnegie Mellon University
Software Engineering Institute

Test and Evaluation of AI Cyber Defense Systems

APRIL 26, 2023

Shing-hon Lau

This material is based upon work funded and supported by the Department of Defense 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.

DM23-0403

Agenda

- **Intersection of AI and cyber**
- Methodology for test and evaluation of AI cyber defenses

Intersection of AI and cyber

Carnegie Mellon University Software Engineering Institute

AI for defending cyber systems

Network defense, endpoint protection, malware detection

AI for attacking cyber systems

Decision support, automating of portions of attacks

Cyber defense for AI systems

AI is ultimately just software on a computer

Cyber offense against AI systems

Identifying code, models, data, etc., after owning AI box

Adversarial AI / ML

Attack an AI system through cleverly selected valid inputs

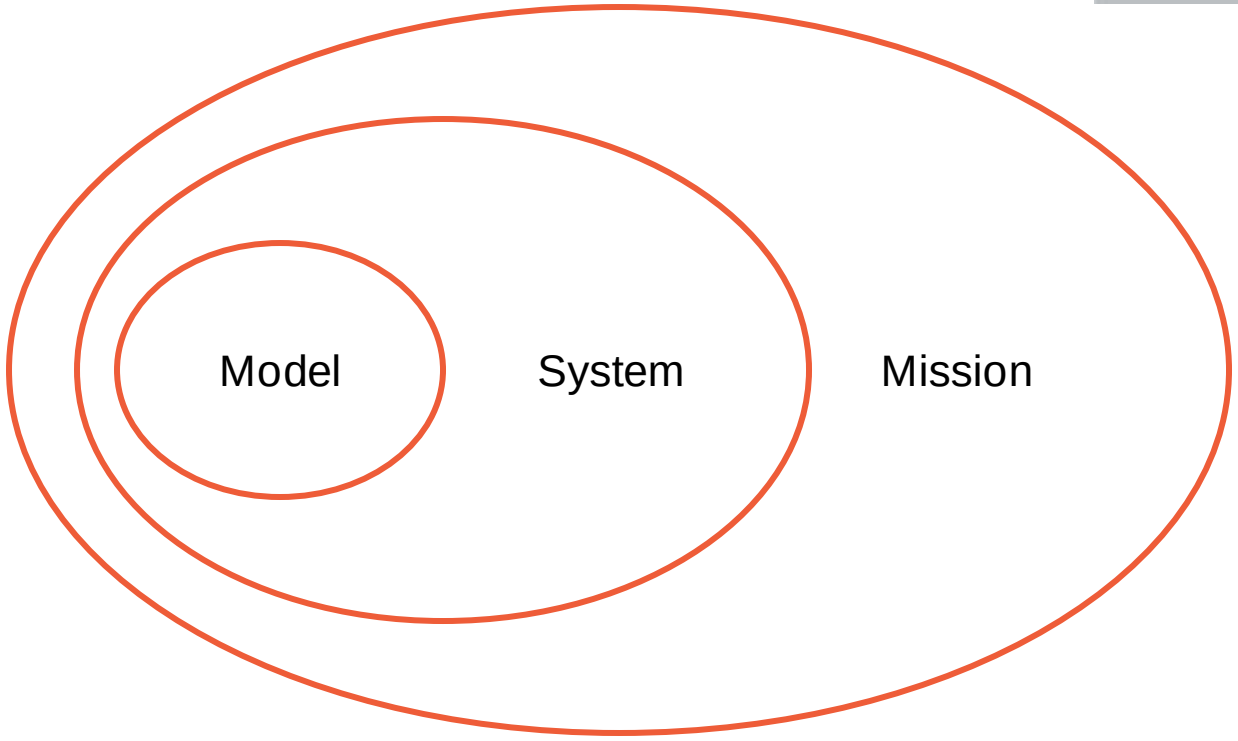
Cyber attacks for adversarial AI effects

E.g., buffer overflows to realize adversarial AI attacks

Thinking about AI systems

- Consider AI at several levels:
  - the model level
  - the system level
  - the mission level

- Gaining an understanding of AI at a system or mission level provides more accurate insights, but is considerably more difficult



AI for defending cyber systems

Carnegie Mellon University Software Engineering Institute

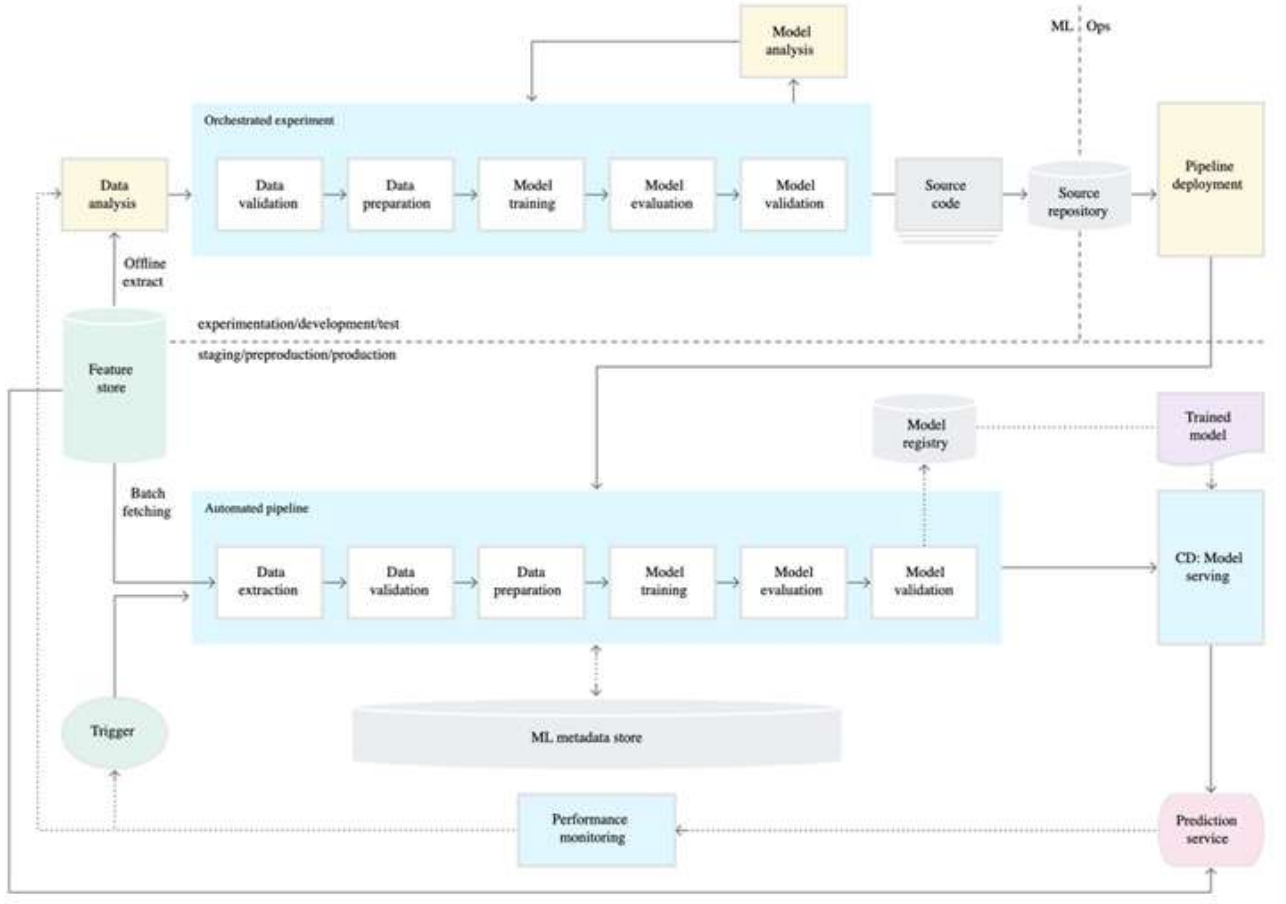
- Network defense
  - Use network traffic to determine malicious activity
- Endpoint behavior detection
  - Observe local execution to determine malicious activity
- Malware detection
  - E.g., Heuristic-based anti-virus or anti-malware

- More on this later…

AI for attacking cyber systems

- Decision support for red teams
- Automated and semi-automated tools for well-defined activities
- Continuous automated red teaming
- Researchers have developed proofs of concept of machine learning or AI-based malware

Cyber defense for AI systems



https://insights.sei.cmu.edu/blog/a-hitchhikers-guide-to-ml-training-infrastructure/

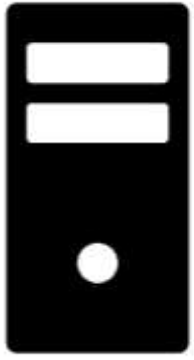
- AI systems are software running on computers
- Data is a first-order concept in machine learning systems
- Control of data can allow attacker to have arbitrary control of system behavior
- Use of AI expands the attack surface

Cyber offense against AI systems

What do you steal when you own the AI box?

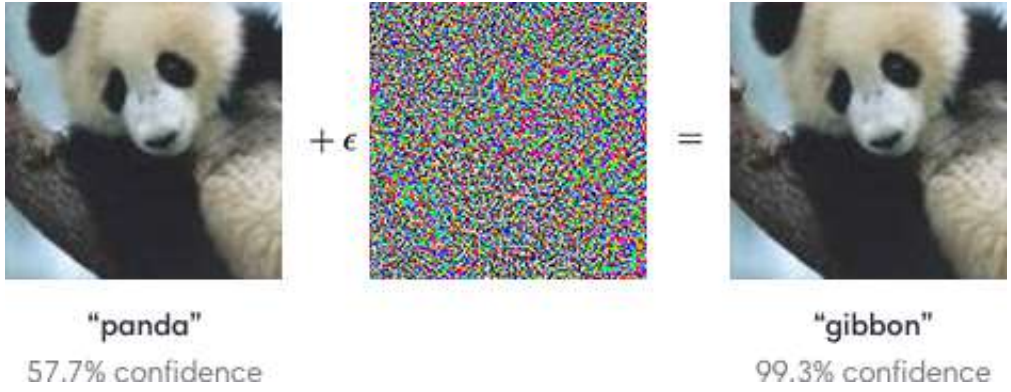
How do you identify key AI components?

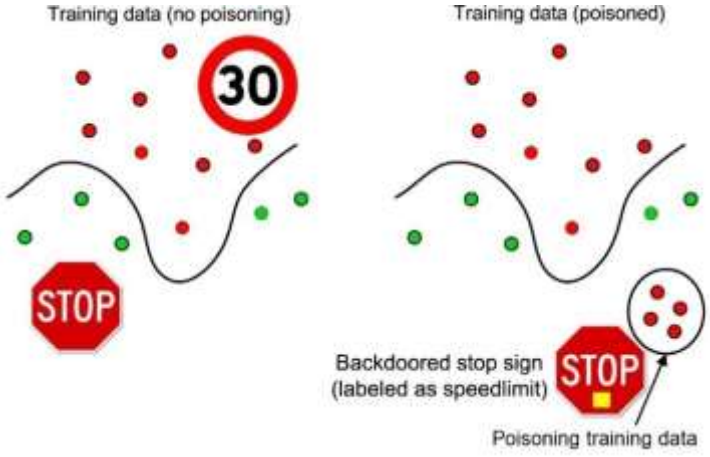
What information is revealed by stolen materials?



Neutral network architecture	Pre-trained models
Fully trained models	Training parameters
Training data	Testing data
Code	Documentation

Adversarial AI and ML





Do the wrong thing

https://openai.com/blog/adversarial-example-research/

Learn the wrong thing

https://smartcities.ieee.org/newsletter/june-2021/explainable-machine-learning-for-secure-smart-vehicles

Reveal the wrong thing

https://gab41.lab41.org/robust-or-private-model-inversion-part-ii-94d54fd8d4a5

Adversarial AI and ML

Carnegie Mellon University Software Engineering Institute

- What does adversarial AI and ML look like for network data, static binary analysis, dynamic binary analysis, firmware, etc.?
- What are the data sources?
- What are the right set of features and architectures for cyber domains?
- How can red teams or adversaries map ML feature level changes back to a cyber domain while respecting constraints, e.g.,
  - Cyber attacks must still be effective
  - Executables must be valid and must execute with correct effect
  - Firmware must not brick devices

Cyber attacks for adversarial AI effects

- Cyber attacks are often treated as orthogonal to adversarial AI

- An adversarial AI attack can be the objective of a cyber attack:
  - Insert a backdoor into an ML model by introducing poisoned data into the ML dataloader process during training
  - Cause misclassifications of test samples by adding adversarial noise during testing

- Common AI software is typically research-quality code, with plenty of security flaws

Agenda

- Intersection of AI and cyber
- **Methodology for test and evaluation of AI cyber defenses**

AI for defending cyber systems

- **Network defense**
  - **Use network traffic to determine malicious activity**
- Endpoint behavior detection
  - Observe local execution to determine malicious activity
- Malware detection
  - E.g., Heuristic-based anti-virus or anti-malware

Organizations are turning to AI-powered network defenses

Carnegie Mellon University Software Engineering Institute

- There is a significant shortage of qualified cybersecurity staff
  - The US has a shortfall of 700k+ cybersecurity staff (https://www.cyberseek.org/heatmap.html)
  - 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 complex threats that may elude analyst detection (e.g., SolarWinds)

- Cyber attacks can be so rapid that human response is impractical
  - NotPetya attack took down an entire Ukrainian bank in 45 seconds
  - Human reaction to the threat is slow; the damage can be irreversible

Test and evaluation of AI defenses

Carnegie Mellon University Software Engineering Institute

- AI defenses pose a test and evaluation challenge unlike those posed by traditional cybersecurity defenses
  - Organizations might need to evaluate tools in a black-box or gray-box environment, without direct access to the innards of the defense
  - AI defense designers intend for systems to learn from their network environment, necessitating creation of a realistic testbed
  - Designers intend for defenses to learn and change over time, so a singular evaluation is insufficient
  - Adversarial manipulation can fool AI defenses, creating vulnerabilities an adversary may exploit

Creating a testing and evaluation methodology

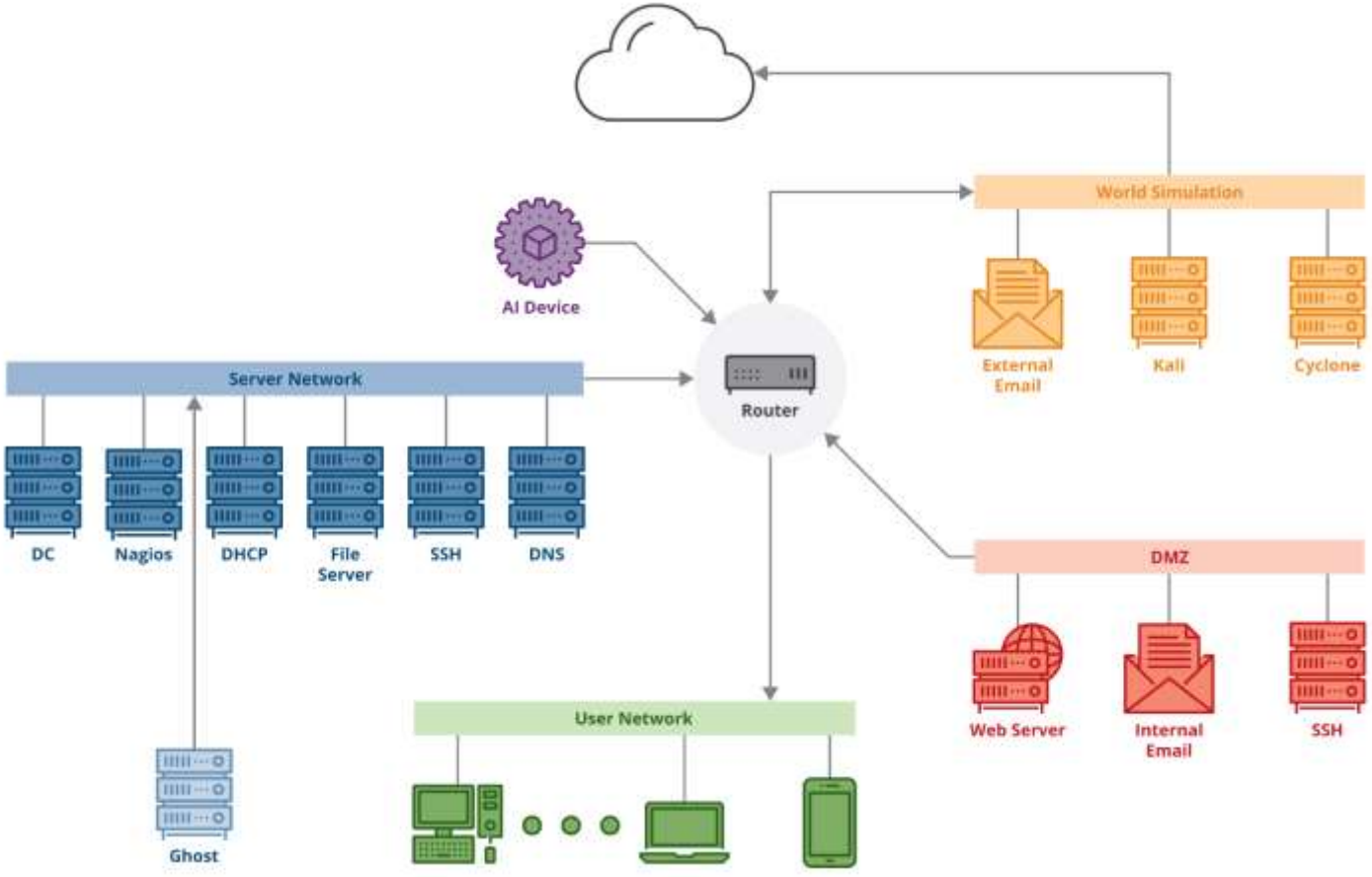
Carnegie Mellon University Software Engineering Institute

- Based on the identified challenges, our methodological approach:

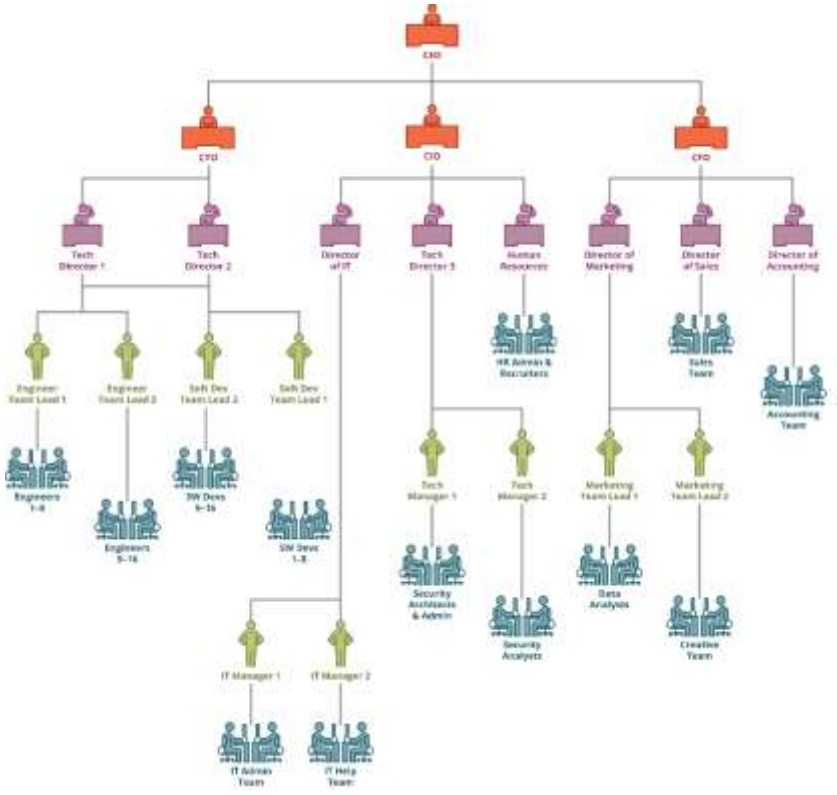
  - Creates a realistic network environment where an AI defense can be deployed
  - Populates the network environment with sufficiently realistic background traffic to allow the AI to learn
  - Tests AI defense performance against realistic cyber attacks
  - Tests AI defense performance when exposed to adversarial manipulation

- Tooling development to expedite and partially automate testing

Creating a network environment



Simulating realistic user behavior (background traffic)

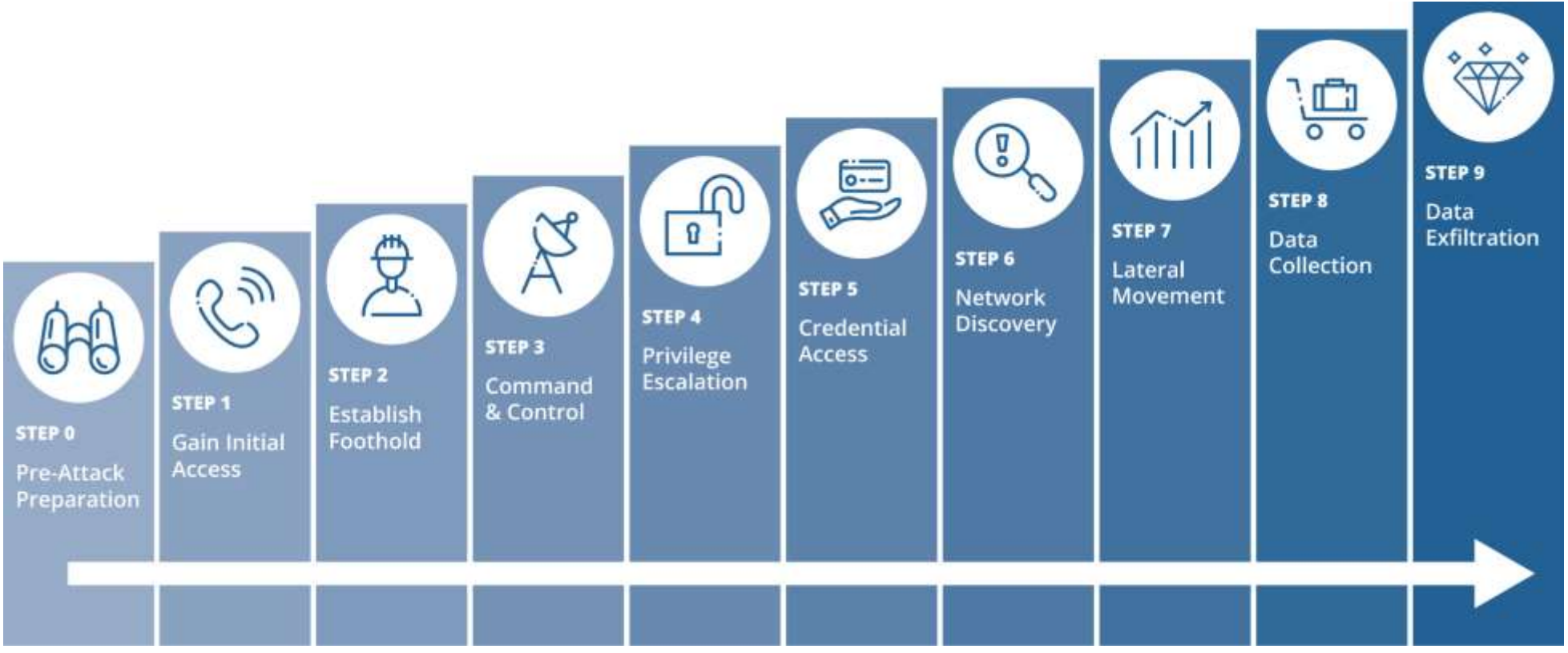


- 99 employees split across 5 divisions
- We provided a unique behavior for each user
  - Customized work schedules
  - Role-specific work tasks
  - Hobbies that influence personal use
- We set privileges and access by role
- Traffic results from simulated behavior–it is not directly simulated

Realistic network user behaviors

- SEI GHOSTS software used as the foundation of network traffic generation (available on public GitHub repository)
- GHOSTS was initially designed to create realistic background traffic to support cyber training exercises
- We repurposed it to run indefinitely to generate background traffic on our network
- GHOSTS permits definition of a user's behavior throughout the day using "timelines"
  - Developed timeline generator that permits fine-grained control of user behavior on a day-to-day basis. This can be customized for each user

Testing with meaningful cyberattacks (malicious traffic)



Automated testing (MITRE CALDERA)

Cyber attack test coverage

- We have mapped our cyber attack test suite to the MITRE ATT&CK framework
- Not all attacks with the ATT&CK framework are detectable by the types of AI defenses we consider
- A total of 70 techniques are covered so far in our methodology

Test cases

- Four meaningful cyberattacks were defined and implemented using the functionality in CALDERA and using publicly-available knowledge and tools:
  - Creation of a domain administrator account
  - Creation of a local administrator account
  - Disabling a public-facing webserver
  - Exfiltration of user files
- Tests were performed in a baseline condition and obfuscated conditions
- An initial test was performed, followed by a test after one month

Looking ahead

- Testbed improvements
  - Collect sensor data to facilitate other avenues of research, particularly involving adversarial machine learning
  - Create a more “modern” test network that is likely closer to what was envisioned when designing AI defenses
- Improved automation
  - Automation at the “above-VM” level to facilitate more types of cyberattacks with minimal human intervention and to offer coordination with blue-team results
- Accommodating replay traffic
  - Results on the best simulated data are no substitute for results using real data
- Increasing the set of test attacks
  - The broader the set of test attacks, the more we understand AI defense capabilities

Dr. Shing-hon Lau
slau@sei.cmu.edu

U.S. Mail
Software Engineering Institute
4500 Fifth Avenue
Pittsburgh, PA 15213-2612 USA

Website
https://www.sei.cmu.edu/contact-us/index.cfm