CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

# Cybersecurity and Artificial Intelligence

***The future of offense, defense, and cyber stability***

AUGUST 12, 2021

**Wyatt Hoffman** | Research Fellow

wyatt.hoffman@georgetown.edu | cset.georgetown.edu

# Overview

## Outline:

- The Basics (AI vs ML)
- AI for Cyber Offense
- AI for Cyber Defense
- Hacking AI
- Strategic Implications
- Recommendations for Cooperation

## Key Questions:

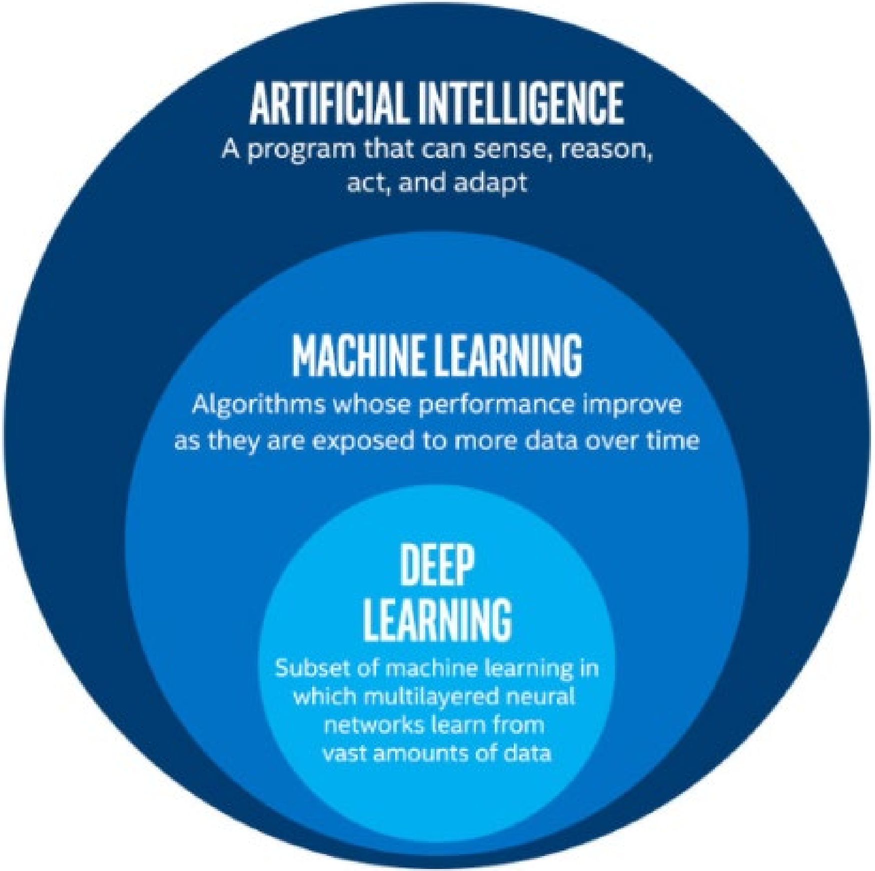
- What does AI have to offer for cybersecurity? What are its limitations?
- How might AI reshape the cyber threat landscape?
- How might AI change the strategic dynamics of cyber competition?

# The Basics

*1: AI vs Machine Learning*

“Machine learning systems use computing power to execute algorithms that learn from data.”

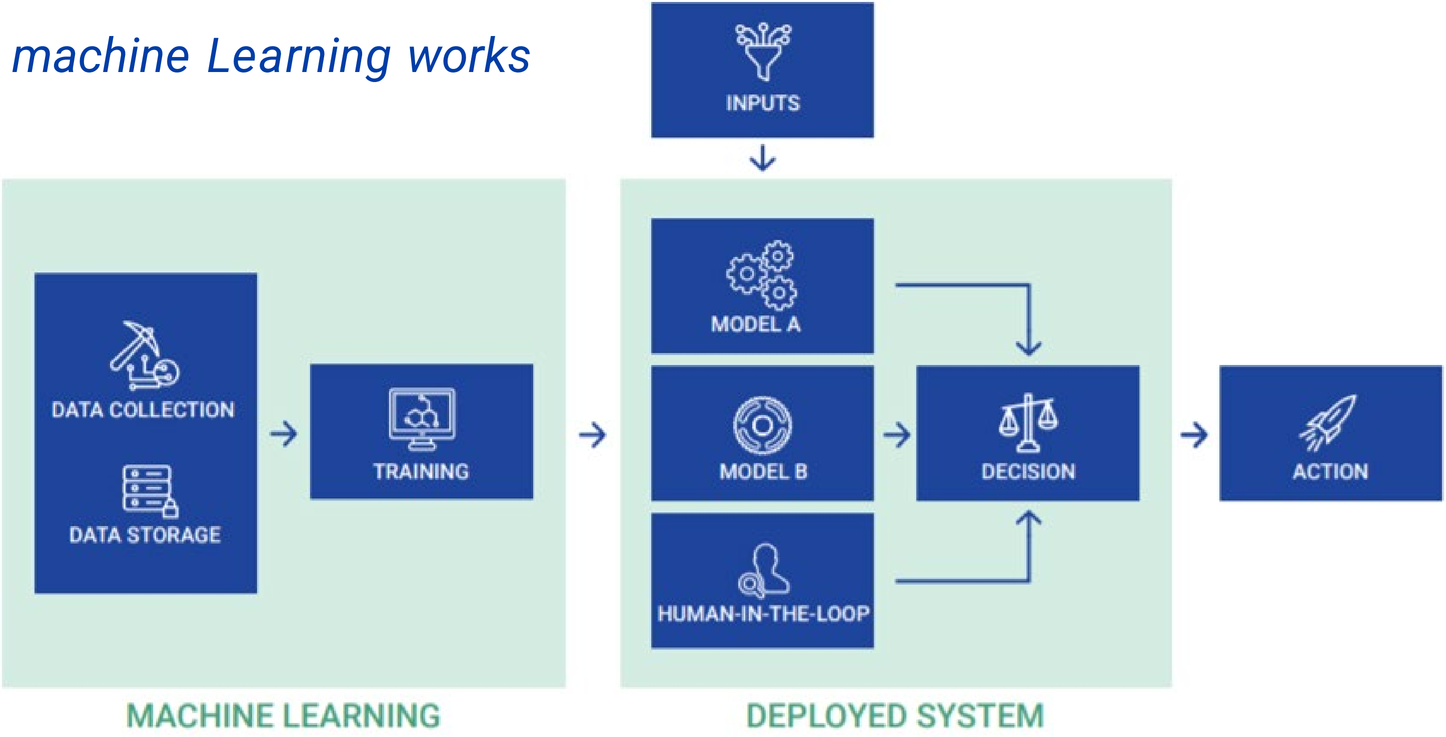
-Ben Buchanan, “The AI Triad and What It Means for National Security Strategy”


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Source: Artem Oppermann, “Artificial Intelligence vs. Machine Learning vs. Deep Learning,” Toward Data Science, Oct 29, 2019

# The Basics

*2: How machine Learning works*



Source: Andrew Lohn, “Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity” CSET, December 2020

# The Basics

## *3: Machine learning strengths and limitations*

### Strengths

- **Superhuman performance:** ML can discover patterns imperceptible to humans, useful for making predictions
- **Adaptivity:** ML systems can continue to learn while deployed
- **Automation:** ML systems can perform tasks that would otherwise require human expertise

### Limitations

- **Data-dependent:** Success hinges crucially on high quality and quantity training data
- **Resource-intensive:** Training and operation demand significant computing power
- **Brittle:** ML systems cannot cope well with environmental changes or adversarial inputs that violate assumptions learned in training
- **Explainability:** ML systems are ‘black boxes’ whose decisions are difficult to understand

***Machine learning is no panacea.***

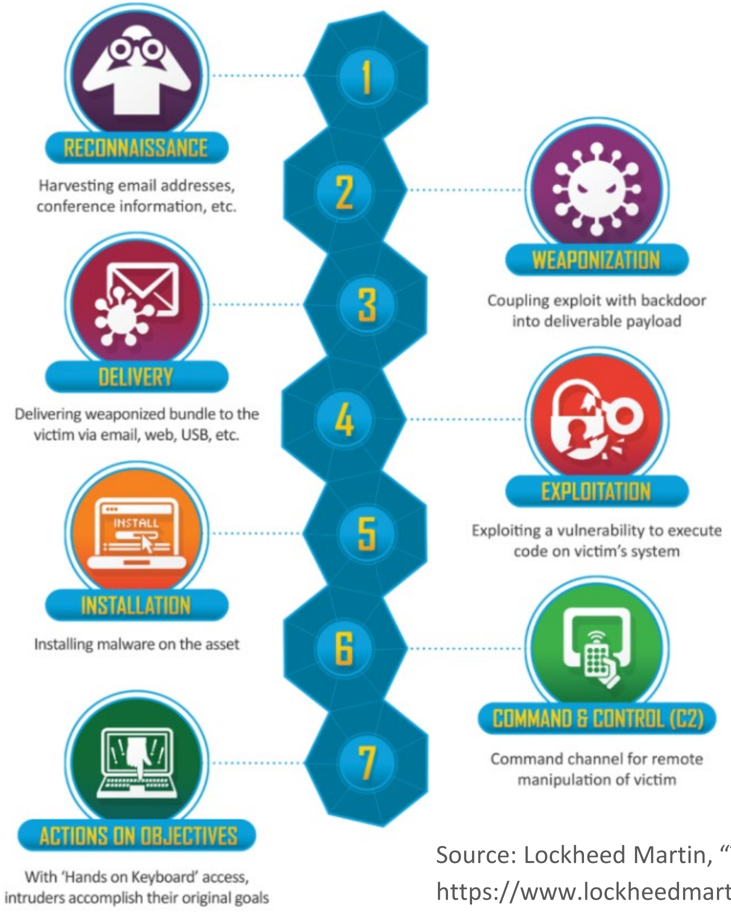
# AI for Cyber Offense

**Near-term applications:**

- Automated vulnerability hunting
- Highly targeted spearphishing and social engineering

**More speculative:**

- Smarter propagation
- Stealthier, evasive malicious capabilities
- More powerful offensive operations


1
RECONNAISSANCE
Harvesting email addresses, conference information, etc.
2
WEAPONIZATION
Coupling exploit with backdoor into deliverable payload
3
DELIVERY
Delivering weaponized bundle to the victim via email, web, USB, etc.
4
EXPLOITATION
Exploiting a vulnerability to execute code on victim's system
5
INSTALLATION
INSTALL
Installing malware on the asset
6
COMMAND & CONTROL (C2)
Command channel for remote manipulation of victim
7
ACTIONS ON OBJECTIVES
With 'Hands on Keyboard' access, intruders accomplish their original goals

Source: Lockheed Martin, “The Cyber Kill Chain” https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

# AI for Cyber Defense

**Near-term applications:**

- Automated vulnerability hunting
- ML-enabled malware and intrusion detection

**More speculative:**

- Active defense measures (e.g adaptive honeypots)
- Moving target defenses

Source: DARPA, “Cyber Grand Challenge” https://www.darpa.mil/program/cyber-grand-challenge

# Hacking AI

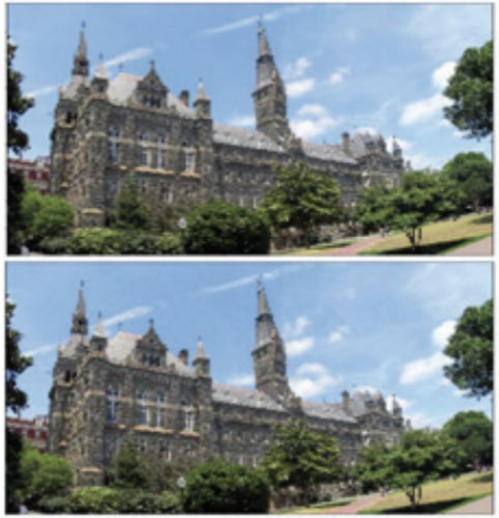
## *1: Adversarial machine learning*

Two main approaches:

- ***Evasion***: craft inputs that violate the assumptions of the model
- ***Poisoning***: tamper with training data to mistrain a system or insert a backdoor

FIGURE 3

Classification of Georgetown's Healy Hall unperturbed on top and attacked to appear to a machine learning system to be a triceratops on bottom. To human eyes, the two images look identical.



ORIGINAL IMAGE

Castle: 85.8%

Palace: 3.17%

Monastery: 2.4%

ATTACKED IMAGE

Triceratops: 99.9%

Barrow: 0.005%

Sundial: 0.005%

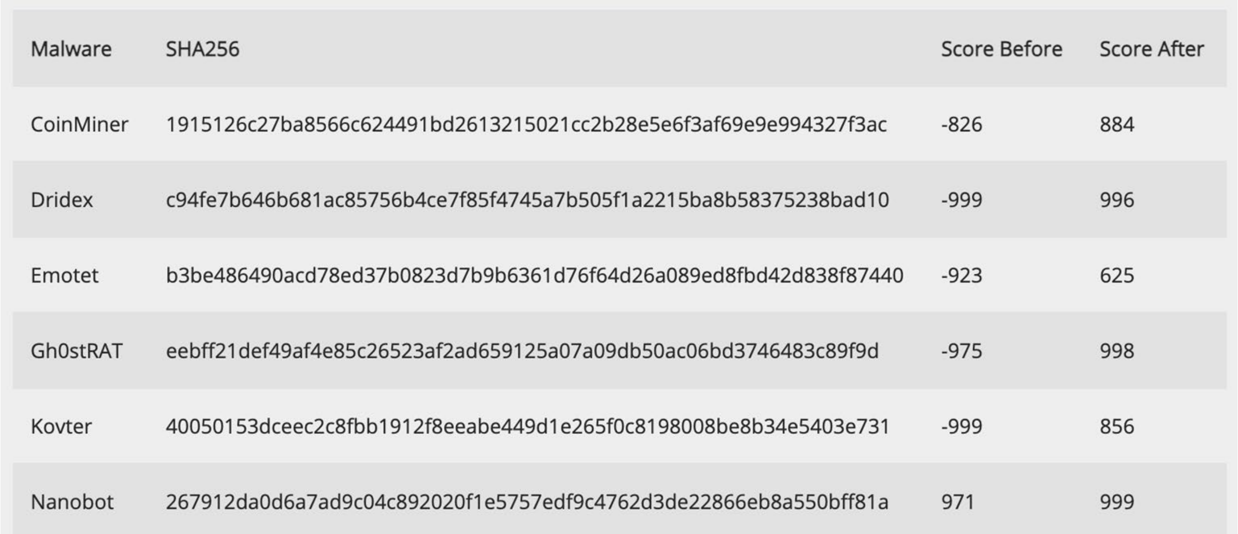
Source: Lohn, "Hacking AI"

# Hacking AI

## *2: Hacking ML-based cyber defenses*

Ex) “Universal bypass” discovered in Cylance ML-based antivirus engine

Skylight Cyber, “Cylance, I Kill You!” https://skylightcyber.com/2019/07/18/cylance-i-kill-you/



| Malware | SHA256 | Score Before | Score After |
|---|---|---|---|
| CoinMiner | 1915126c27ba8566c624491bd2613215021cc2b28e5e6f3af69e9e994327f3ac | -826 | 884 |
| Dridex | c94fe7b646b681ac85756b4ce7f85f4745a7b505f1a2215ba8b58375238bad10 | -999 | 996 |
| Emotet | b3be486490acd78ed37b0823d7b9b6361d76f64d26a089ed8fbd42d838f87440 | -923 | 625 |
| Gh0stRAT | eebff21def49af4e85c26523af2ad659125a07a09db50ac06bd3746483c89f9d | -975 | 998 |
| Kovter | 40050153dceec2c8fbb1912f8eeabe449d1e265f0c8198008be8b34e5403e731 | -999 | 856 |
| Nanobot | 267912da0d6a7ad9c04c892020f1e5757edf9c4762d3de22866eb8a550bff81a | 971 | 999 |

# Strategic Implications

*1: How might AI reshape the cyber threat landscape?*

ML could empower attackers, or level the playing field for defenders:

- Countervailing offensive and defensive applications across the ‘kill chain’
- ML could unlock unrealized defensive advantages: control over the “playing field,” access to vast data on network activity
- BUT defense faces unique challenges: greater concerns over reliability, attack vectors targeting ML itself
- In the worst case, ML might fuel more dangerous, destructive attacks

***Whether AI helps attackers or defenders more depends in part on making AI defensible.***

# Strategic Implications

*2: How might AI shape the strategic dynamics of cyber competition?*

AI could be destabilizing for several reasons:

- Introduce new risks of unintended impacts or collateral damage from autonomous capabilities
- Incentivize more aggressive cyber campaigns to compromise or sabotage ML systems (e.g. targeting supply chains) or target trust in ML itself
- Increase the escalation risks of cyber engagements (e.g. misinterpretation of an espionage operation as an attack)
- Expand the scope of possible impacts from cyber operations targeting AI capabilities in general

# Recommendations for Cooperation

***Maximize potential defensive benefits***

- Share best practices for AI safety and security
- Collaborate on *adversarial robustness*
- Secure the foundation for AI development (supply chains, data sources)

***Limit potential harm from offensive use***

- Information sharing on common threats (e.g. emerging threats to industrial control systems)
- Counter the proliferation of offensive capabilities
- International norms for offensive cyber operations

# Further Reading

- **Automating Cyber Attacks: Hype and Reality** by Ben Buchanan, John Bansemer, Dakota Cary, Jack Lucas and Micah Musser
- **Destructive Cyber Operations and Machine Learning** by Dakota Cary and Daniel Cebul
- **Machine Learning and Cybersecurity: Hype and Reality** by Micah Musser and Ashton Garriott
- **Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity** by Andrew Lohn
- **AI and the Future of Cyber Competition** by Wyatt Hoffman

*Available at: cset.georgetown.edu*

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

- Research at https://cset.georgetown.edu/research/
- Sign up to receive research the day it's issued, subscribe to our biweekly newsletter, and get invited to our events at https://cset.georgetown.edu/sign-up/
- Watch CSET webinars and request briefings, if needed
- Share your questions and knowledge gaps

**cset.georgetown.edu | policy.ai | @csetgeorgetown | cset@georgetown.edu**