Welcome
2017 Faculty Summit
Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI
AI and Security: Lessons, Challenges and Future Directions

Taesoo Kim
Taesoo Kim
About Myself

Taesoo Kim (taesoo@gatech.edu)

• 14- : Assistant Professor at Gatech
• 11-14: Ph.D. from MIT in CS

Research interests:

Operating Systems, Systems Security, Distributed Systems, Programming Languages, Architecture

https://taesoo.kim
Clarification: Security and AI

• Security → Software or Computer Security
  • In particular, attacker’s perspectives
  • Excluding the security issues that involved human (e.g., fraud, phishing ...)

• AI → ML or Deep Learning
  • In particular, training-based, stochastic approaches
  • It works well in practice, but too complex to understand why? or how?
Three Key Points

• Part 1. What AI can learn from Security?
  → Thinking like an adversary

• Part 2. What Security can learn from AI?
  → Measuring the progress of research

• Part 3. Security after AI?
  → New Era for Advanced Persistent Threats (APT)
Part 1. What AI can learn from Security?
Thinking Like an Adversary

How to hijack this self-driving car?

- Putting wall?
- Attacking sensors?
- Put STOP signs?
Part 1. What AI can learn from Security? Thinking Like an Adversary

Laying a trap for self-driving cars

Posted Mar 17, 2017 by Devin Coldewey
Adversary? Meeting with the Best Hacker!

Full-chain exploitation on all major browsers and platforms!

$225,000 in Pwn2Own’15
$300,000 in PwnFest’16
...

Now in Google’s Project Zero Team
First Public Talk @Zer0Con’17

Conference for Exploit Developers & Bug Hunters

A medley of modern web browser exploits

This talk introduces the various web browser vulnerabilities I've found and reported, and how I exploited those vulnerabilities. I will discuss not only just web browser vulnerabilities, but also various logical bugs and kernel bugs.
Lots of (even) Hackers are Curious ..

Could you explain how you found bugs in Pwn2Own’16?
Lots of (even) Hackers are Curious..

Could you explain how you found bugs in Pwn2Own’16?

Umm .. what?
(his friend translated ..)
Lots of (even) Hackers are Curious..

Could you explain how you found bugs in Pwn2Own’16?

“Intuition ...”
Lots of (even) Hackers are Curious..

Could you explain how you found bugs in Pwn2Own’16?

Intuition... except one bug that I had to open IDA for reverse engineering.
Approaches to Security vs. ML

• **Security:**
  (Translating) Intuition $\rightarrow$ Methodologies

  VS.

• **ML:**
  (Inferring) Training data $\rightarrow$ Parameters
Take-away Messages from Security

• Attackers target a single, weakest component
• Rethinking of your assumption (aka, threat model)
• Increasing #features → larger attack surface
• Focusing on *directly* translating intuition to models
• Making the design iteration comprehensive (ie., explainable)
Part 2. What Security can learn from AI? Measuring the Progress of Research

- Turing Test: 1950
- Chess: 1985-1996
  - Deep Blue: 1996
- Jeopardy: 2011
  - Watson: 2011
- Go!: 2016
  - AlphaGo: 2016
Electronic Frontier Foundation (EFF) announces:

https://www.eff.org/ai/metrics (e.g., handwritten digit recognition)
What happens to Security (and Privacy)?

• Perhaps, too subject?
• What do you even mean by measuring “security”?

• In terms of exploit/defenses:

  CTF games (human vs. human)

  DARPA Cyber Grand Challenge (computer vs. computer)
Take-away Messages from AI

- AI fields drive research as various landmark competitions
- Public resources for quantifying the progress (e.g., data sets)
- Perhaps, people tend to “hide” security-related data
- Too subjective, but we might be able to tackle subfields of security?
- So we can objectively measure pros/cons of security mechanisms

It works, but I don’t know why?

• AI takes off → “unknown” software everywhere!
• In particular, when Security relies on AI-based approaches

(APT = Advanced persistent threat, or targeted attack)
Take-away Messages (once AI takes off)

• More attack surface for attackers: impl, algorithm, data, etc.

• What if attackers understand more deeply than you?
• What if attackers can influence your data set?
• What if we don’t even observe attacks (i.e., accountability)?
On-going Efforts at Georgia Tech

• Intel Science and Technology Center (ISTC) for Adversary-Resilient Security Analytics (MLsploit)
• NSF/SaTC: CORE: Medium: Understanding and Fortifying Machine Learning Based Security Analytics

• Security: Wenke Lee, Taesoo Kim
• ML: Polo Chau, and Le Song
Mike Walker
“What can AI learn from security”? 
1996

.oO Phrack 49 Oo.

Volume Seven, Issue Forty-Nine

File 14 of 16

BugTraq, r00t, and Underground.Org bring you

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Smashing The Stack For Fun And Profit
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

by Aleph One
aleph1@underground.org

2011

<table>
<thead>
<tr>
<th>Memory Corruption (19)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Defeated by DEP</td>
<td>14</td>
</tr>
<tr>
<td>Defeated by ASLR</td>
<td>17</td>
</tr>
<tr>
<td>Defeated by EMET</td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logic Flaws (8)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Java in Internet Zone</td>
<td>4</td>
</tr>
<tr>
<td>No EXEs in PDFs</td>
<td>1</td>
</tr>
<tr>
<td>No Firefox or FoxIt Reader</td>
<td>2</td>
</tr>
</tbody>
</table>

$114B

Right: Dan Guido, Exploit Intelligence Project
Left: Aleph One, Phrack 49
The Honeymoon Effect

Bugs: Starts fast, then *slows down*

Vulnerabilities: Starts slow, then *speeds up!*

Frei, Clark, Blaze, Smith: Familiarity Breeds Contempt
No Reports of Attacks

Mandiant, APT1 Report
Golden Opportunities in AI Security

• Any software that serves as a gatekeeper to valuable IP, wealth, or life safety must consider the eventual arrival of an expert adversary

• Attack detection is not free; it requires active research & sensors

• No reports of attacks != no attacks

• Techniques to defeat security properties must be discovered & published in the open first (Fun) or be exploited (Profit)
“What can security learn from AI?”
Machine Learning versus Sensors

SandPrint:

“we can use those inherent features to detect sandboxes using supervised machine learning techniques [...] an attacker can reveal characteristics of publicly available sandboxes and use the gathered information to build a classifier that can perfectly distinguish between a user PC and an appliance”

AFL vs. djpeg

Michal Zalewski, Google, on AFL
“Security & AI”
Discussion
“What can AI learn from security”? 
“What can security learn from AI?”
“What does security look like after AI ‘happens’?”
Wrap-up and next steps

• What can AI learn from security?
• What can security learn from AI?
• What does security look like after AI happens?

New techniques, new problems to solve, new collaborations
Find someone to work with today!
Thank you