Industrial Practitioners’ Mental Models of Adversarial Machine Learning

Lukas Bieringer, Kathrin Grosse, Battista Biggio, Michael Backes, Katharina Krombholz

Department of Electrical and Electronic Engineering
University of Cagliari, Italy
Outline

Recap ML & AML

Sample

Results
Outline

Recap ML & AML

Sample

Results

Kathrin Grosse (kathrin.grosse@unica.it) – SOUPS 2022
Machine Learning
Machine Learning
Machine Learning
Machine Learning
Adversarial Machine Learning
Outline

Recap ML & AML

Sample

Results

Kathrin Grosse (kathrin.grosse@unica.it) – SOUPS 2022
Qualitative Sample – 15 Participants (2020)

- 14 male / 1 female
- Age: 34 (+/- 4.27)
- Employer: European start-ups (<200 employees)
- Application areas:
  - Cybersecurity, healthcare, vision, human resources...
Interview Procedure

Demographics

Outline

Recap ML & AML  Sample  Results
Key findings – AML versus Non-AML Security

- AML
- Non AML Security

Expected

Key findings – AML versus Non-AML Security

Expected

- AML Security

- Non AML Security

Found

- AML Security

- Non AML Security

Details – AML versus Non-AML Security

• AML mitigations** vs security defenses
Details – AML versus Non-AML Security

• AML mitigations vs. security defenses

• Threats in AML are doubted
  • Externalized responsibility (4)
  • Have not encountered threat
  • Doubt attacker's motivation (7)
  • Believe have a working mitigation (9)
Details – AML versus Non-AML Security

• AML mitigations** vs security defenses

• Threats in AML are doubted
  • Externalized responsibility (4)
  • Have not encountered threat
  • Doubt attacker’s motivation (7)
  • Believe have a working mitigation (9)

Model reverse engineering
Model Stealing
Code breach

Membership Inference
Data Breach

DoS Attacks
Details – AML versus Non-AML Security

- AML mitigations** vs security defenses
- Threats in AML are doubted
- Externalized responsibility (4)
- Have not encountered threat
- Doubt attacker's motivation (7)
- Believe have a working mitigation (9)

Poisoning
Evasion

Model reverse engineering
Model Stealing
Code breach

Membership Inference
Data Breach

DoS Attacks

Kathrin Grosse (kathrin.grosse@unica.it) – SOUPS 2022
Key findings – Model versus Workflows

Key findings – Model versus Workflows

Research

Model

Practice

Model

Details – Model versus Workflows

DATA → API → HD SFT → ML → OUTPUT (REPORT)

1/ No continuous learning
2/ No continuous learning
3/ passwords

Kathrin Grosse (kathrin.grosse@unica.it) – SOUPS 2022
Open questions

Application vs perceived Relevance vs Education

Kathrin Grosse (kathrin.grosse@unica.it) – SOUPS 2022

AML attacks in practice

- ‘What we experienced is not so much AML – but semi-automated fraud’

Implications

- Enforce that both ML and non-ML security are taken care of
- Provide reasonable data so that research can be practical
- There are AML attacks in practice
- Educating practitioners on AML seems crucial