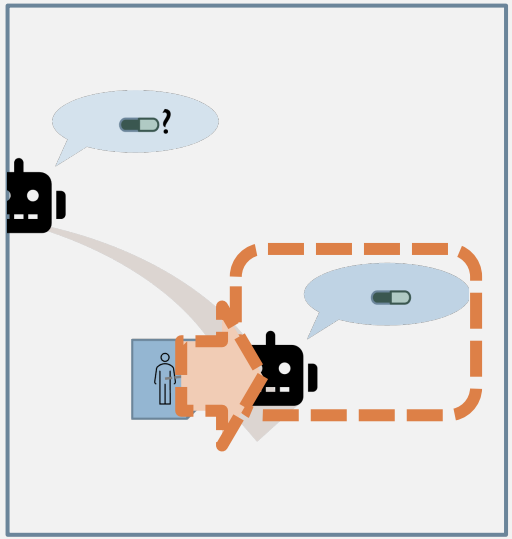
Industrial Practitioners' Mental Models of Adversarial Machine Learning

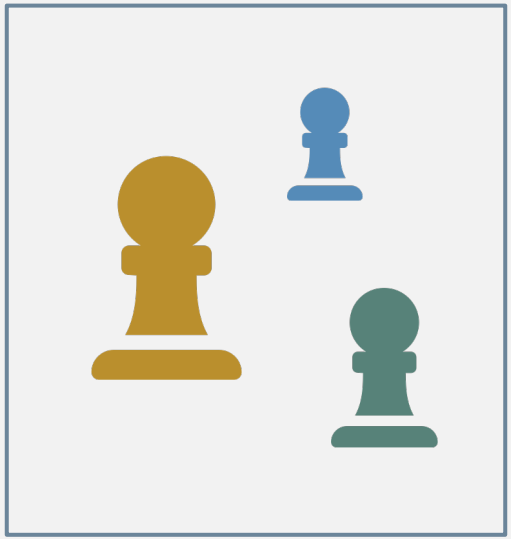
Lukas Bieringer, **Kathrin Grosse**, Battista Biggio, Michael Backes, Katharina Krombholz

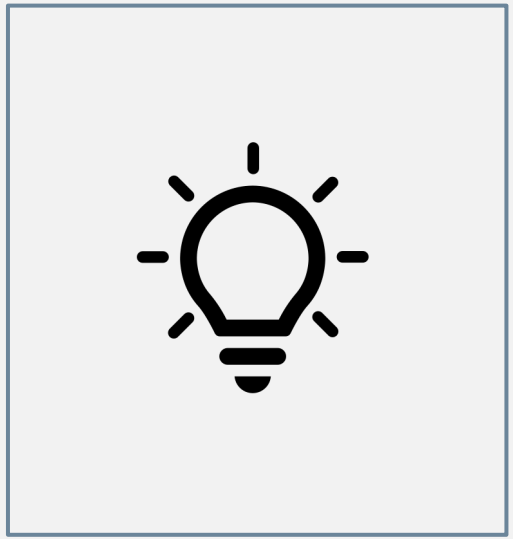
Department of Electrical and Electronic Engineering

University of Cagliari, Italy

Outline





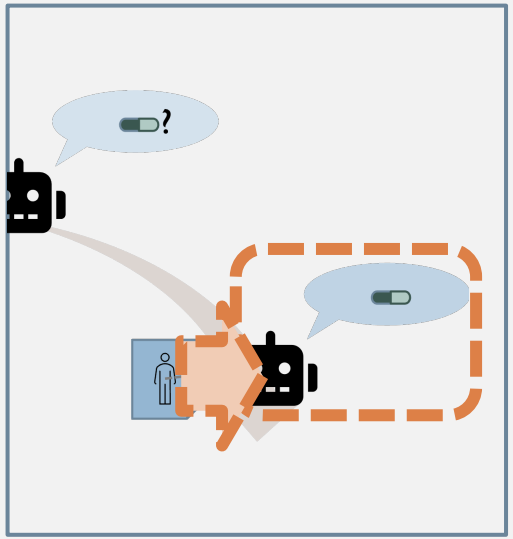


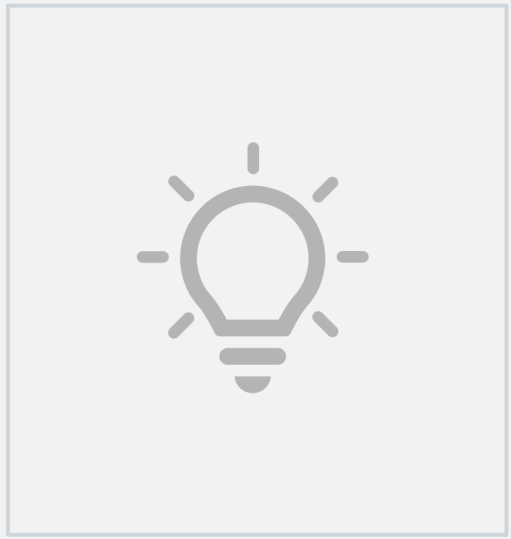
Recap ML & AML

Sample

Results

Outline





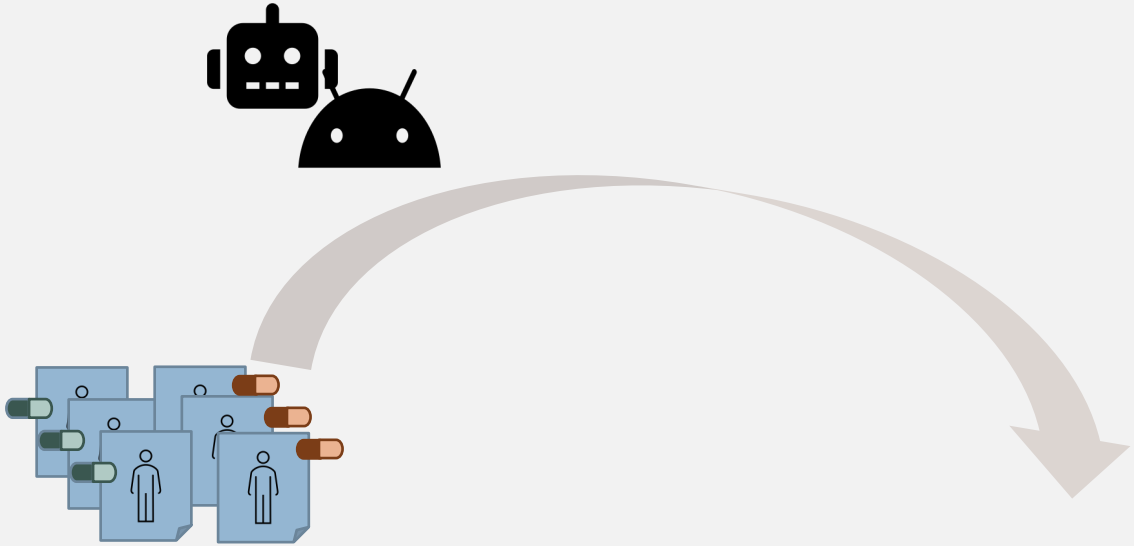
Recap ML & AML

Sample

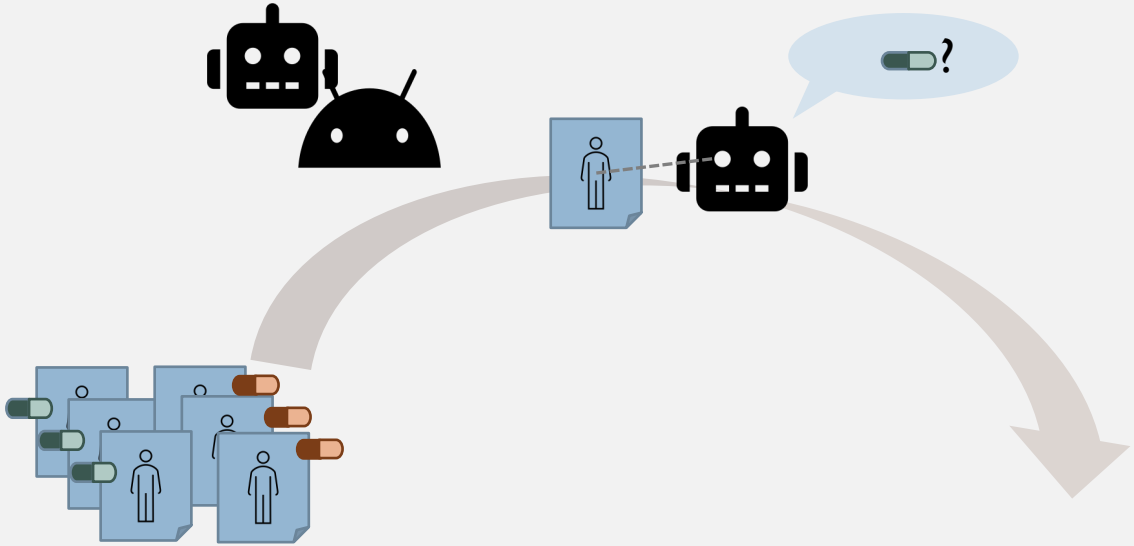
Results

Machine Learning

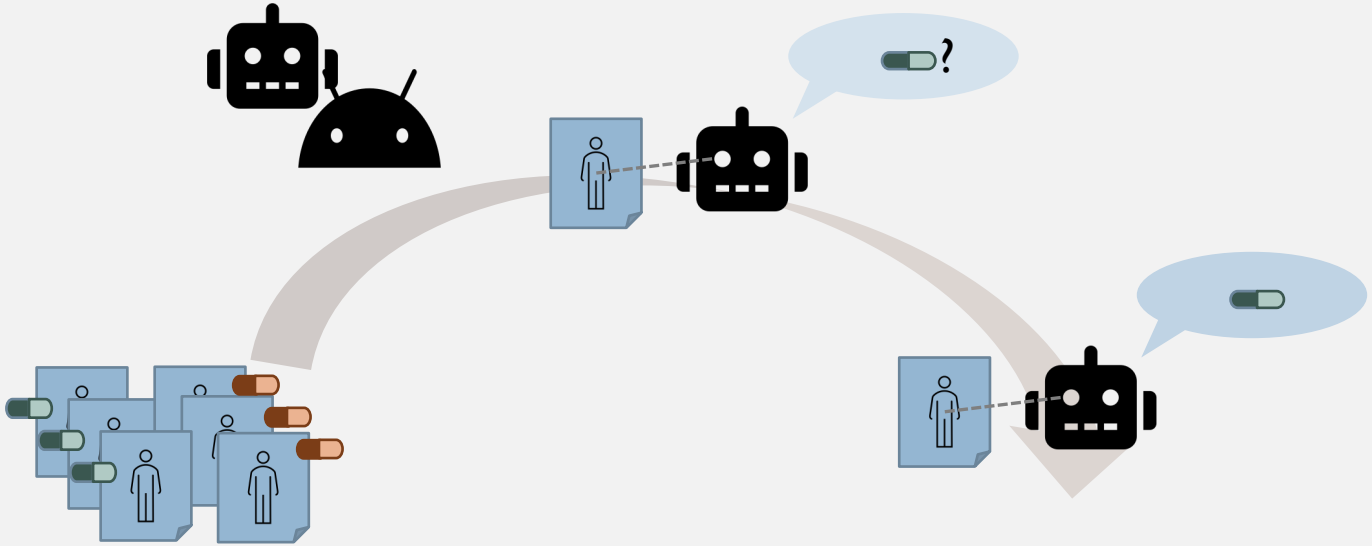
Machine Learning



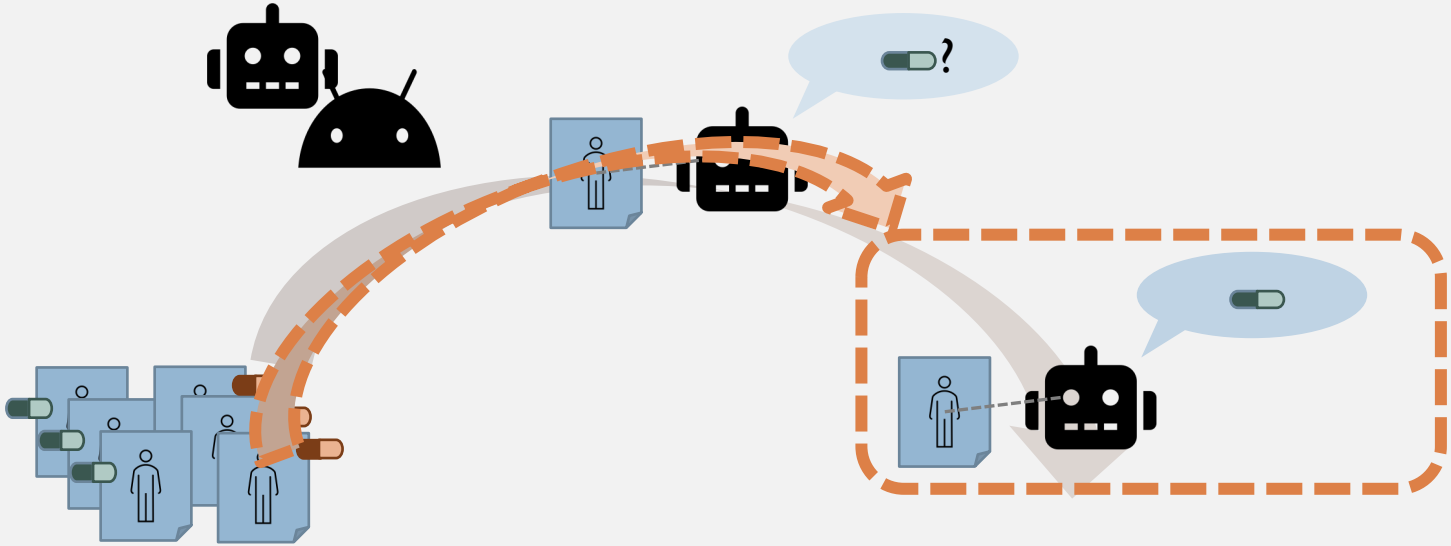
Machine Learning



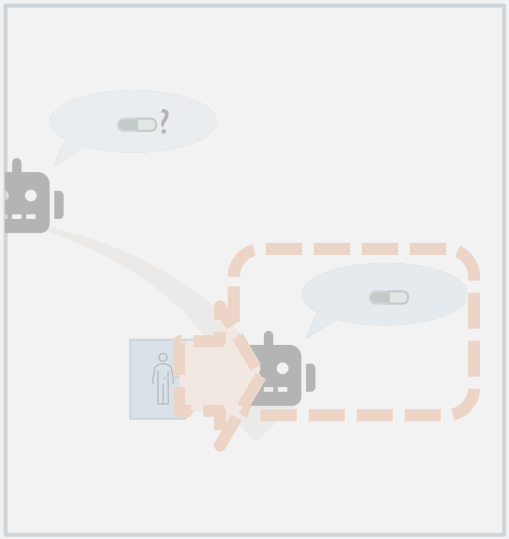
Machine Learning

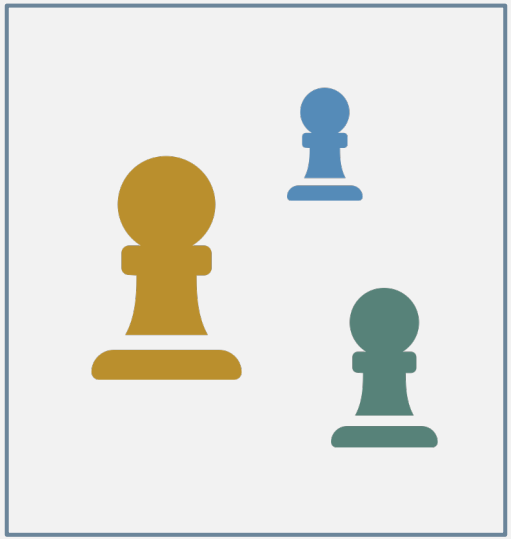


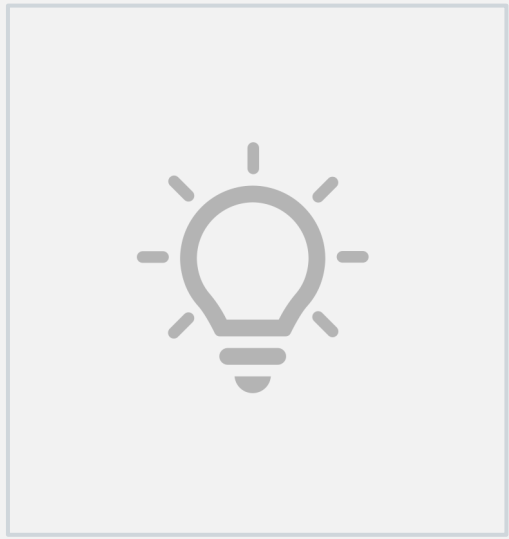
Adversarial Machine Learning



Outline







Recap ML & AML

Sample

Results

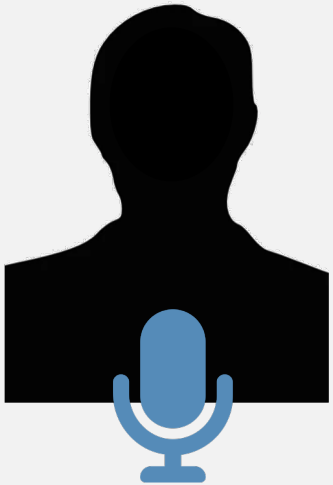
Qualitative Sample – 15 Participants (2020)

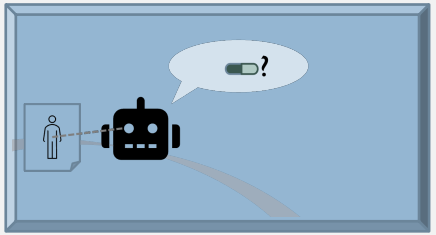
- 14 male / 1 female
- Age: 34 (+/- 4.27)
- Employer: European start-ups (<200 employees)
- Application areas:
  - Cybersecurity, healthcare, vision, human resources…

Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

Interview Procedure

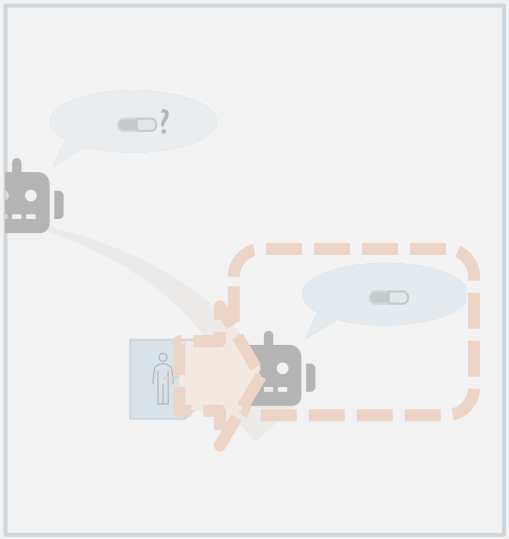
Demographics

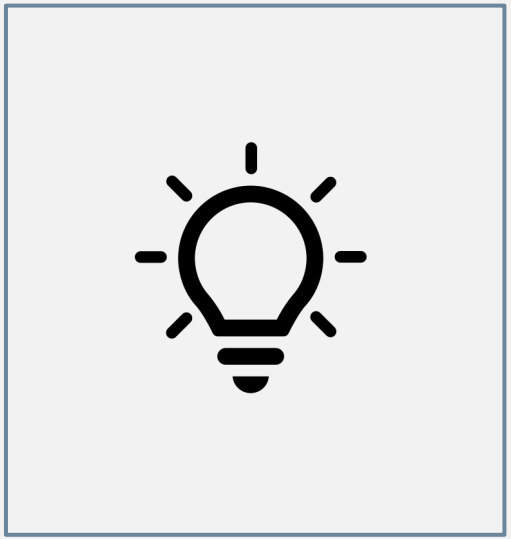




Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

Outline



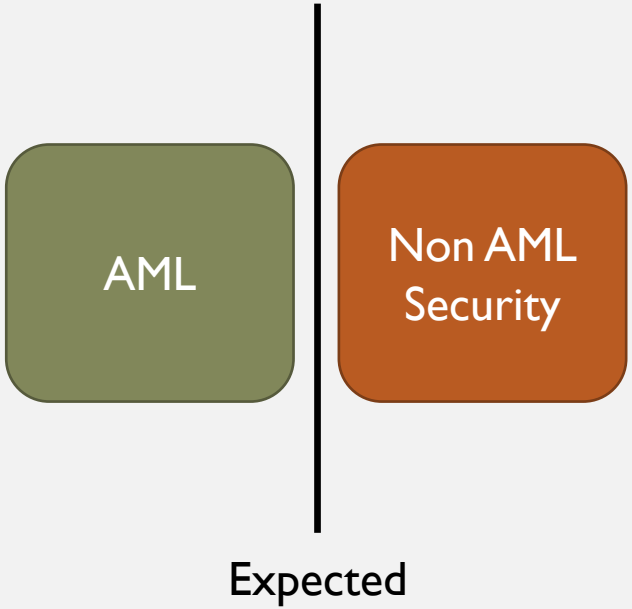


Recap ML & AML

Sample

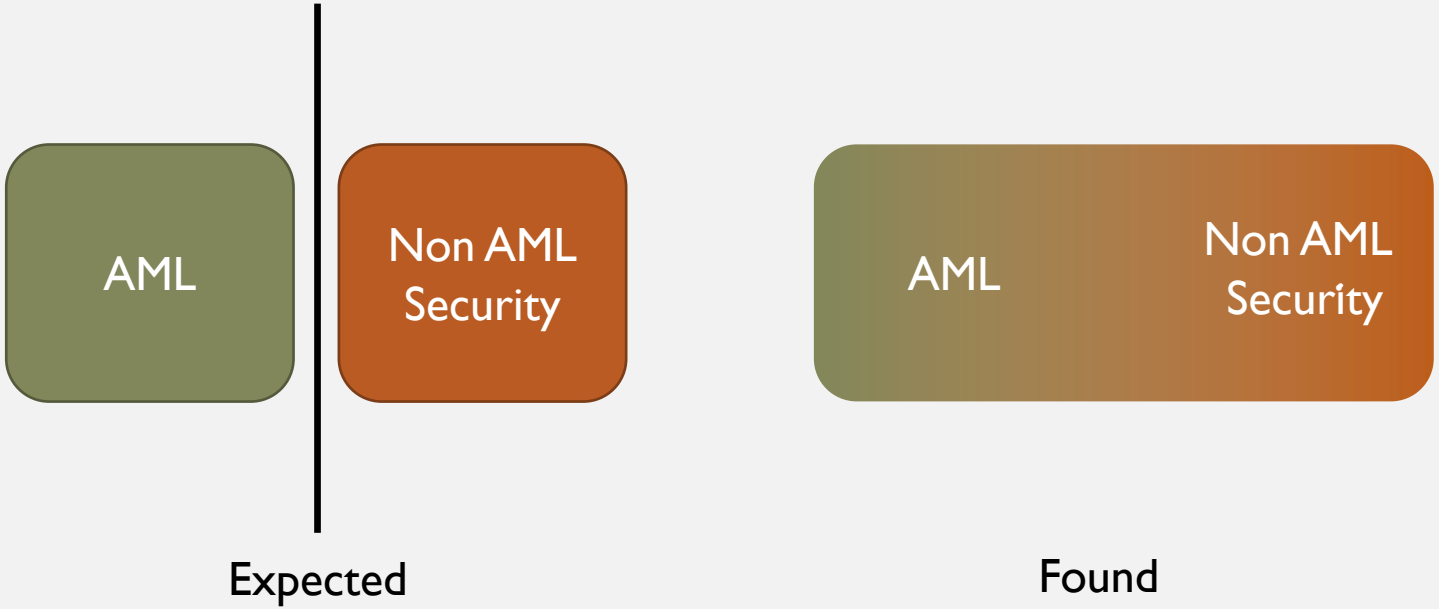
Results

Key findings – AML versus Non-AML Security



Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

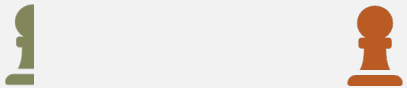
Key findings – AML versus Non-AML Security



Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

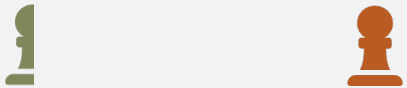
Details – AML versus Non-AML Security

- AML mitigations** vs security defenses



Details – AML versus Non-AML Security

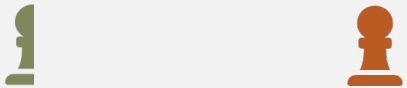
- AML mitigations** vs security defenses



- 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

Poisoning

Evasion

ML are doubted

re

nc

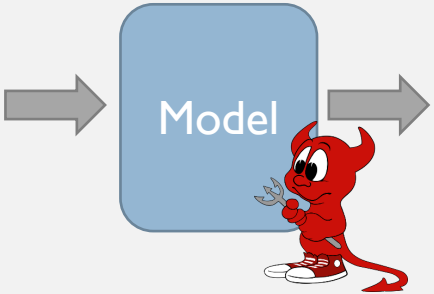
- Doubt attack
- Believe have c on (9)

Model reverse engineering
Model Stealing
Code breach

Membership Inference
Data Breach

DoS Attacks

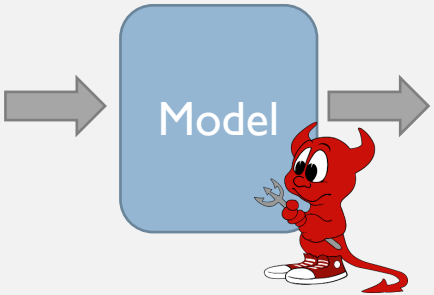
Key findings – Model versus Workflows



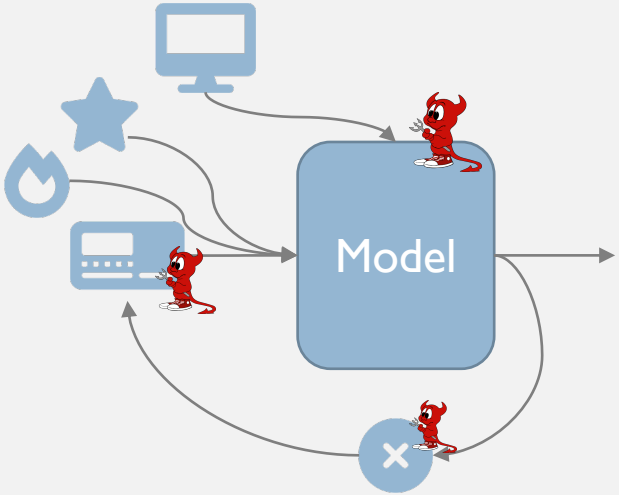
Research

Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

Key findings – Model versus Workflows



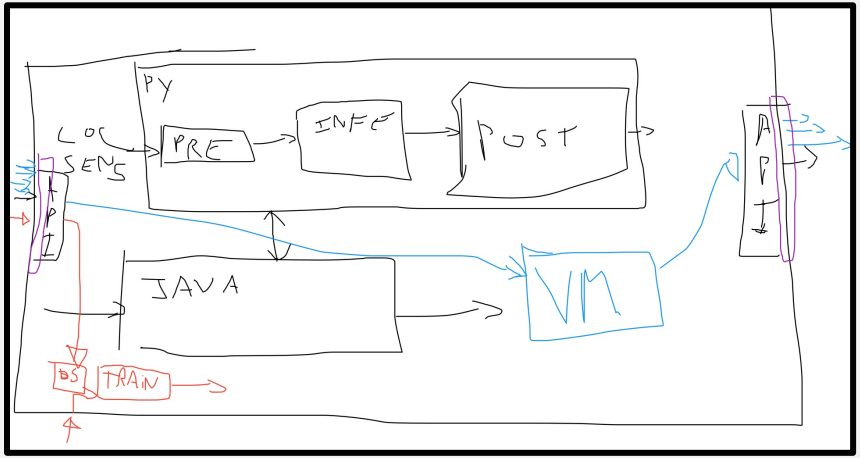
Research



Practice

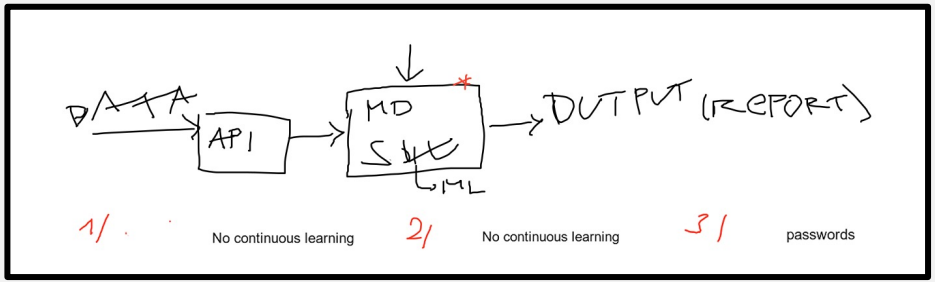
Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

Details – Model versus Workflows



S16

S18



Open questions

Application

perceived Relevance

Education

Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

AML attacks in practice

- 'What we **experienced** is not so much AML – **but semi-automated fraud**'

Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

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