University of
Cagliari, Italy

AI Security and Safety: The PRAlab Research Experience

Ambra Demontis, Maura Pintor, Luca Demetrio, Angelo Sotgiu, Daniele Angioni, Giorgio Piras, Srishti Gupta, ***Battista Biggio****, Fabio Roli*

Ital-IA, Pisa, May 29, 2022

PRALab – Dept. of Electrical and Electronic Engineering

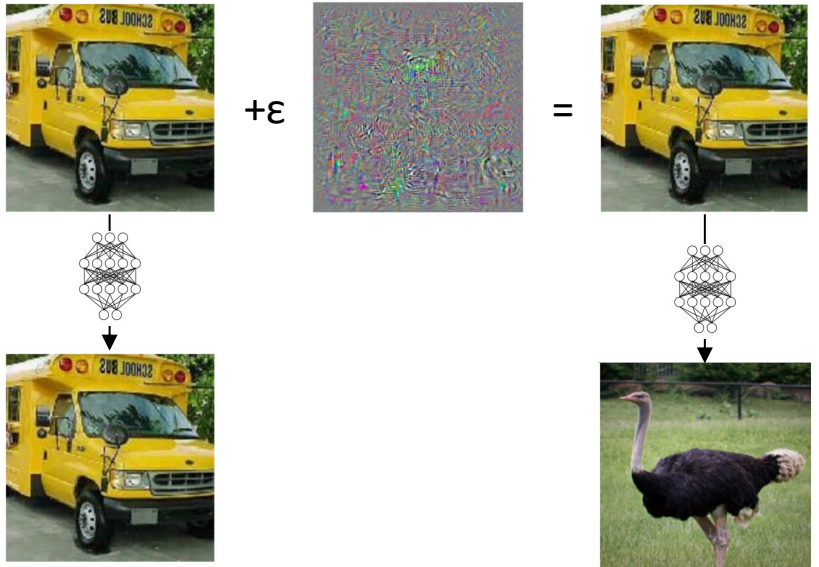
- **Pattern Recognition and Application Laboratory**
  - DIEE, University of Cagliari, Italy
- ~30 people working mainly on:
  - Biometric Recognition
  - Video Surveillance
  - Cybersecurity
  - **AI/ML Security**
- **Recent projects on AI Security**
  - *HE Sec4AI4Sec 2023-2025*
  - *HE ELSA 2022-2024*
  - PRIN 2017 RexLearn
  - FFG Comet Module S3AI

- 25+ research projects (last 10 years)
- 8 EU projects (2 coordinated)
- 1.5 M€ EU funding
- More than 3M€ overall funding
- 400k€ yearly turnover

The Elephant in the Room: *Adversarial Examples*

- AI/ML successful in many applications
  - Computer Vision
  - Speech Recognition
  - Cybersecurity
  - Healthcare

- ... but extremely *fragile* against *adversarial examples*
  - Carefully-perturbed inputs that mislead classification



Biggio et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**

Attacks against AI are Pervasive!

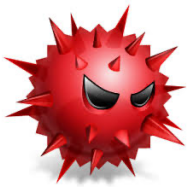
Sharif et al., *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,* ACM CCS 2016

"without the dataset the article is useless"

"okay google browse to evil dot com"

Carlini and Wagner, *Audio adversarial examples: Targeted attacks on speech-to-text*, DLS 2018 https://nicholas.carlini.com/code/audio_adversarial_examples/

Eykholt et al., *Robust physical-world attacks on deep learning visual classification*, CVPR 2018

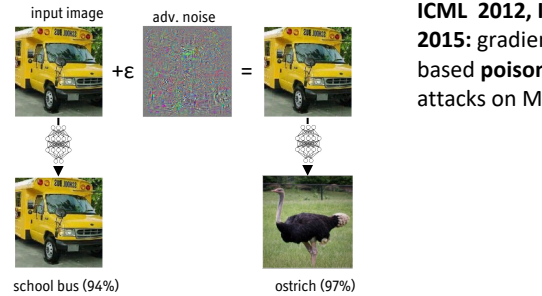


- Demetrio, Biggio, Roli et al., *Adversarial EXEmples: ...*, ACM TOPS 2021
- Demetrio, Biggio, Roli et al., *Functionality-preserving black-box optimization of adversarial windows malware*, IEEE TIFS 2021
- Demontis, Biggio, Roli et al., *Yes, Machine Learning Can Be More Secure!...*, IEEE TDSC 2019

Pioneers of AI/ML Security

- Our team is internationally recognized among the pioneers of AI/ML security
  - we have been the first to discover the impact of gradient-based attacks on ML models
  - we have been the first to discover and systematize adversarial attacks on AI/ML, prior to their application to deep learning

ECML-PKDD '13: gradient-based **evasion** attacks on ML (**one year before adversarial examples**)



ICML 2012, ICML 2015: gradient-based **poisoning** attacks on ML

	Attacker's Goal: Misclassifications that do not compromise normal system operation	Attacker's Goal: Misclassifications that compromise normal system operation	Attacker's Goal: Querying strategies that reveal confidential information on the learning model or its users
Attacker's Capability	**Integrity**	**Availability**	**Privacy / Confidentiality**
Test data	Evasion (a.k.a. adversarial examples)	Sponge attacks	Model extraction / stealing Model inversion (hill climbing) Membership inference
Training data	Backdoor poisoning (to allow subsequent intrusions) – e.g., backdoors or neural trojans	DoS poisoning (to maximize classification error)	-

B. Biggio and F. Roli, *Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning*, Pattern Recognition, 2018 - ***2021 Best Paper Award and Pattern Recognition Medal***

B. Biggio, B. Nelson, and P. Laskov, *Poisoning Attacks against SVMs*, ICML 2012 - **ICML 2022 Test of Time Award**

Main Research Directions

Attacks on Machine Learning

ECML '13 / ICML '12, '15: Pioneering work on gradient-based evasion and poisoning attacks

USENIX Sec. '19: Transferability of evasion and poisoning attacks

IEEE TDSC '19, IEEE TIFS/ACM TOPS '21: Adversarial perturbations on Android and Windows malware

ECML '20: Poisoning attacks on algorithmic fairness

NeurIPS '21: Fast Minimum-Norm attacks

NeurIPS '22: Indicators of Attack Failures

WACV '23: Phantom Sponges

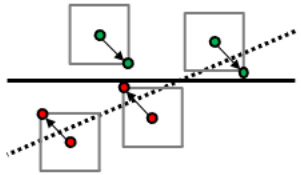
Robust Learning and Detection Mechanisms

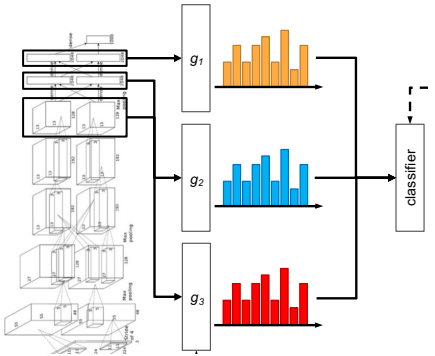
IEEE Symp. S&P '18: Robust learning against training data poisoning

IEEE TDSC '19: Optimal/robust linear SVM against adversarial attacks (use case on Android malware)

NEUCOM '21: Fast adversarial example rejection

IEEE TPAMI '21: Learning with domain knowledge to improve robustness against adversarial examples

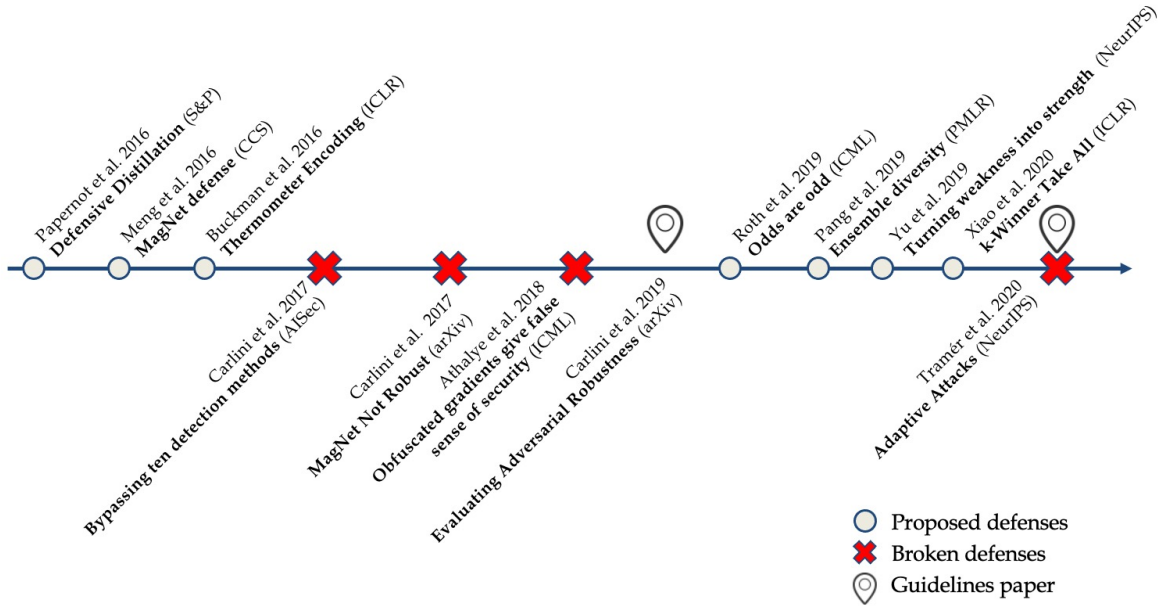




Ineffective Defenses and Flawed Evaluations

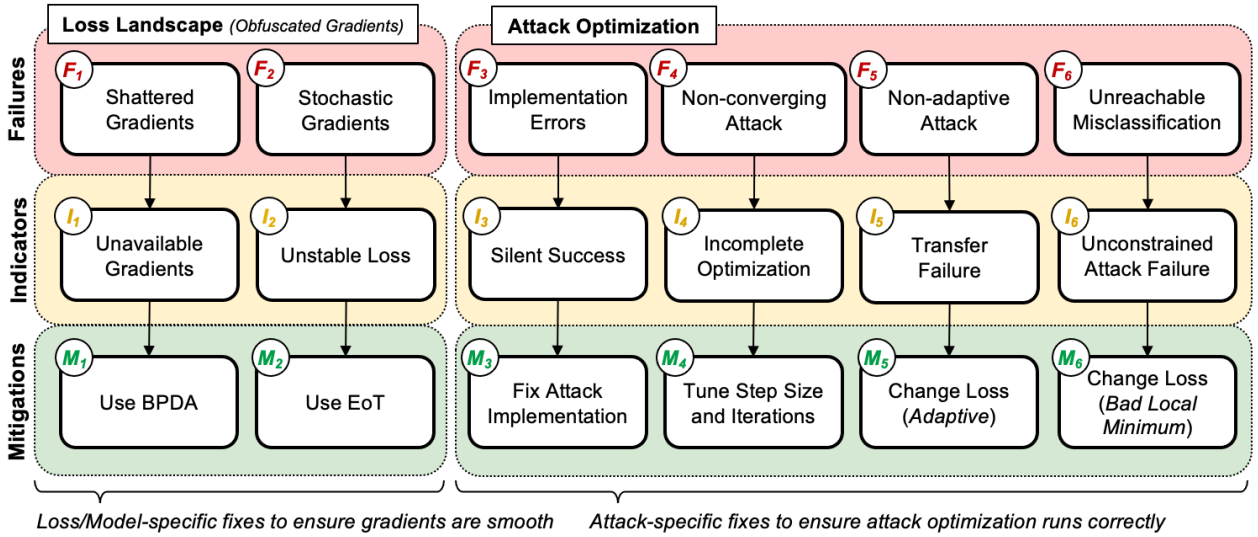
Detect and Avoid Flawed Evaluations

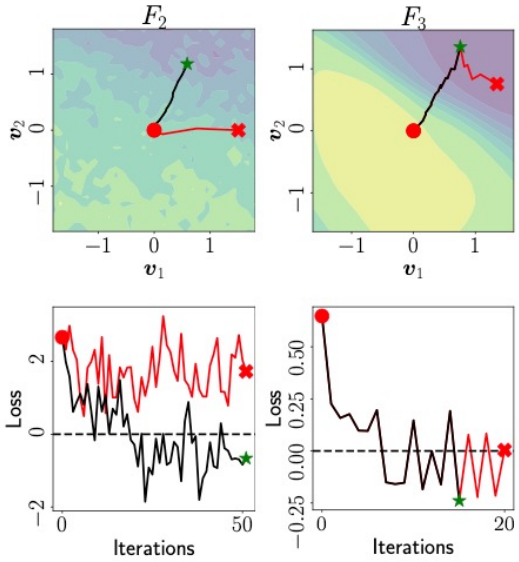
- **Problem**: formal evaluations do not scale, adversarial robustness evaluated mostly empirically, via gradient-based attacks
- **Gradient-based attacks can fail:** many flawed evaluations have been reported, with defenses easily broken by adjusting/fixing the attack algorithms



Pintor, Biggio, et al., *Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples*, NeurIPS 2022

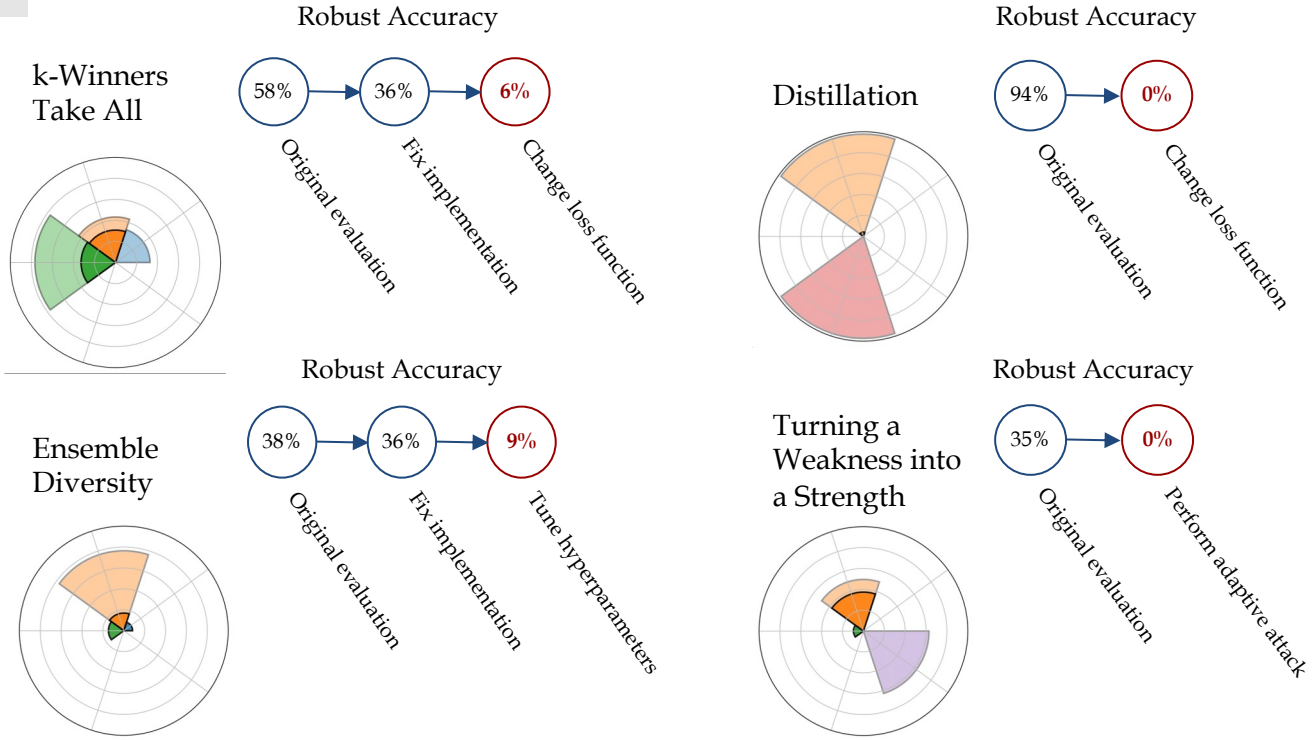
Indicators of Attack Failure





Pintor, Biggio, et al., *Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples*, NeurIPS 2022

Experiments



Pintor, Biggio, et al., *Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples*, NeurIPS 2022

What's Next?

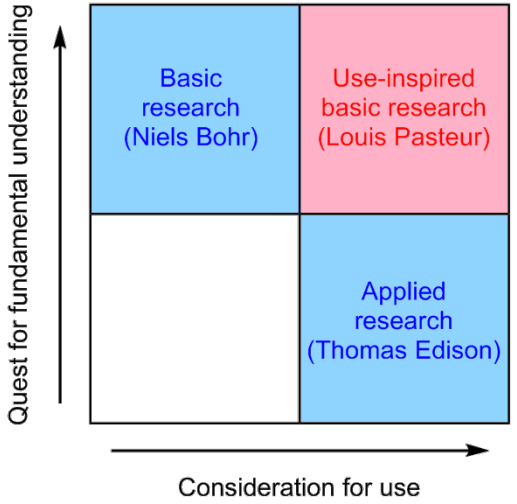
What's Next?

Use-Inspired Basic Research Questions from the Pasteur's Quadrant

- Studying ML Security may help understand and debug ML models... but

- ... can we use MLSec to help solve some of today's industrial challenges?
  - To improve robustness/accuracy over time, requiring less frequent retraining
  - To detect OOD examples and provide reliable predictions (confidence values)
  - To improve maintainability and interpretability of deployed models (update procedures)
  - To learn reliably from noisy/incomplete labeled datasets
  - ...

- **Challenge:** to build more reliable and practical ML models using MLSec / AdvML



Open Course on MLSec

https://github.com/unica-mlsec/mlsec

Software Tools

https://github.com/pralab

Machine Learning Security Seminars

https://www.youtube.com/c/MLSec

YouTube IT

The ELSA Project

elsa
European Lighthouse on Secure and Safe AI

elsa-ai.eu

ellis

Thanks!

Battista Biggio
battista.biggio@unica.it
@biggiobattista