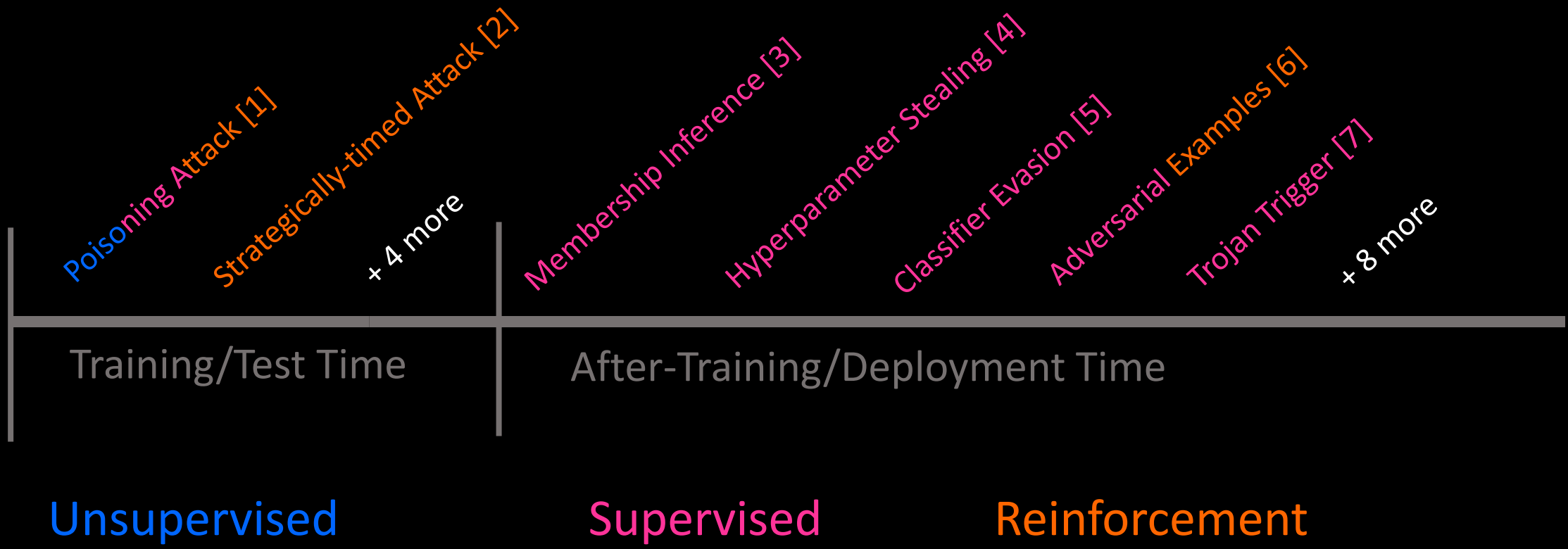
TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Attacks on Machine Learning: Lurking Danger for Accountability

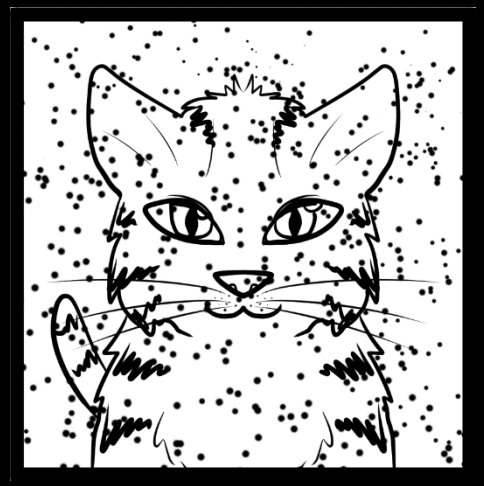
Katja Auernhammer, Ramin Tavakoli Kolagari, Markus Zoppelt
Nuremberg Institute of Technology, Germany

SafeAI
January 27, 2019

Attacks on Machine Learning



Adversarial Examples [6,8]



Manipulated Data

ML Model

Output: Dog

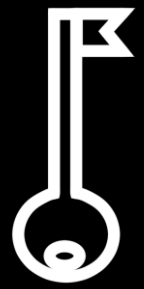
Wrong Output

Integrity

Availability

Confidentiality

Reliability



Authenticity

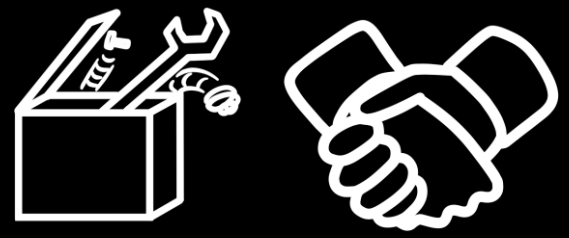
Accountability

Attacks and Security Goals

Poisoning Attack [1]
Strategically-timed Attack [2]
+ 4 more
Membership Inference [3]
Hyperparameter Stealing [4]
Classifier Evasion [5]
Adversarial Examples [6]
Trojan Trigger [7]
+ 8 more

Training/Test Time
After-Training/Deployment Time

Manipulate Model



Leak Information

Evade Model

Peculiarity of Accountability

- Problems:
  - There is no single distinct definition for Accountability
  - Existing definitions are difficult to apply to ML
- Working Definition: „An executed action is clearly traceable to the entity performing said action.“

➢But: What is an „entity“ in ML?

- More explainable AI will lead to a better understanding of Accountability

➢But: It might also open the door for more elaborated attacks

  - More understanding will ultimately also serve the attacker
  - Attackers can create targeted and more efficient attacks

Future Research

- ...should define Accountability distinctively
- ...should apply the definition to the area of ML
- ...has to examine if Accountability will enable new types of attacks
- ...has to find defenses against „misuse of Accountability“

Conclusion

- There might be a danger lurking in the dark for Accountability

References

1. Benjamin I.P. Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J. D. Tygar. “ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors”. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference - IMC ’09. New York, New York, USA: ACM Press, Nov. 2009. ISBN: 9781605587714. DOI: 10.1145/1644893.1644895.
2. Yen Chen Lin, Zhang Wei Hong, Yuan Hong Liao, Meng Li Shih, Ming Yu Liu, and Min Sun. “Tactics of adversarial attack on deep reinforcement learning agents”. In: IJCAI International Joint Conference on Artificial Intelligence (2017), pp. 3756–3762. ISSN: 10450823. arXiv: 1703.06748.
3. Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks Against Machine Learning Models”. In: Proceedings - IEEE Symposium on Security and Privacy (2017), pp. 3–18. ISSN: 10816011. DOI: 10.1109/SP.2017.41. arXiv: 1610.05820.
4. Binghui Wang and Neil Zhenqiang Gong. “Stealing Hyperparameters in Machine Learning”. In: Proceedings - IEEE Symposium on Security and Privacy 2018-May.May (2018), pp. 36–52. ISSN: 10816011. DOI: 10.1109/SP.2018.00038. arXiv: 1802.05351.
5. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. “Evasion Attacks against Machine Learning at Test Time”. In: ECML PKDD (2013), pp. 387–402. DOI: 10.1007/978-3-642-40994-3_25.
6. Christian Szegedy,Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks”. In: (Dec. 2013), pp. 1–10. ISSN: 15499618. DOI: 10.1021/ct2009208.arXiv: 1312.6199.
7. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Authors Yingqi Liu, Weihang Wang, and Xiangyu Zhang. “Trojaning Attack on Neural Networks”. In: NDSS 2018 (Network and Distributed System Security Symposium) (Feb. 2018). DOI: 10.14722/ndss.2018.23291.
8. Fabiano Dalpiaz, Elda Paja, and Paolo Giorgini. Security Requirements Engineering: Designing Secure Socio-Technical Systems. MIT Press, 2016, p. 224. ISBN: 0262034212.