AICA Agents' stealth and resilience capabilities

Adversarial Attacks against ML Agents

Giovanni Apruzzese, PhD

Post-doc Research Assistant

✉ giovanni.apruzzese@uni.li

UNIVERSITÄT LIECHTENSTEIN

Hilti Chair of Data and Application Security

University of Liechtenstein

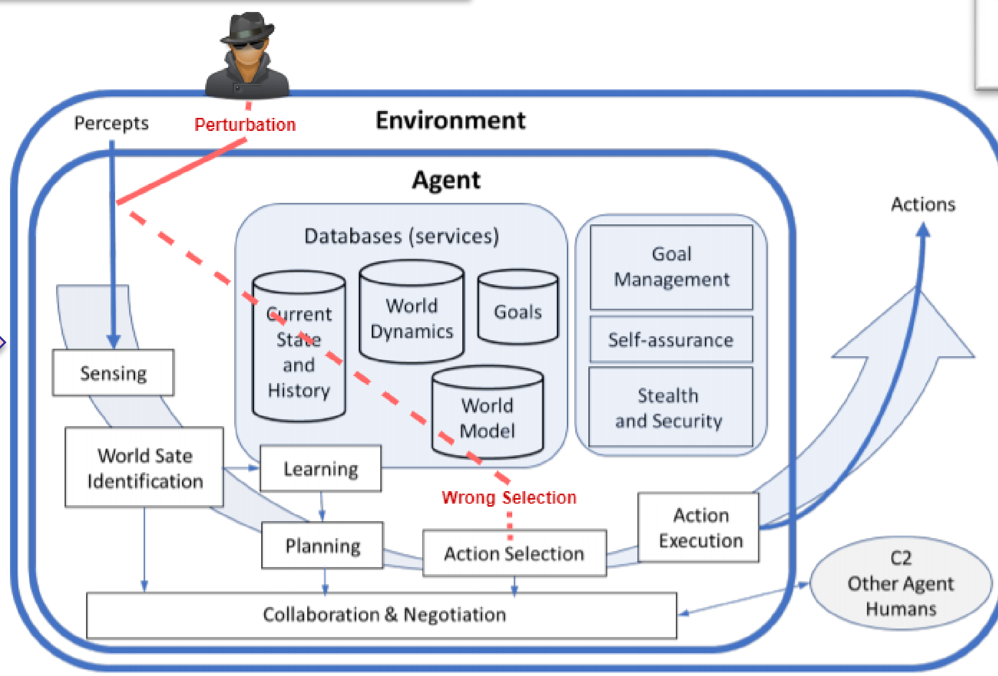
Problem – What's the problem to solve in this research area?

- Machine Learning (ML) is becoming increasingly popular to develop autonomous systems
- **Even the future AICA agents will make ample use of ML techniques**
- However, the application of ML also creates new security risks, e.g.: *adversarial attacks*

Adversarial attacks involve the creation of specific samples with the goal of thwarting the machine learning algorithm.

Even **tiny perturbations** can **greatly affect** the prediction performance [1]

In the case of an AICA agent, an attacker could craft samples that induce the model to select a **wrong** action.



UNIVERSITÄT LIECHTENSTEIN
Giovanni Apruzzese, PhD
giovanni.apruzzese@uni.li

[1]: Su Jiawei et al. "*One pixel attack for fooling deep neural networks.*" IEEE Transactions on Evolutionary Computation (2019).

Scenario – How is it done today, and what are the limits of current solutions?

- The problem is that the source of data used to train the ML model is assumed to be *neutral*
  - However, this is not the case if the model is to be deployed in adversarial environments!
  - **REMEMBER: attackers are attracted by "sensitive" targets!**

- Today, when applying Machine Learning algorithms to solve a problem, <u>the only focus is maximizing performance.</u>
  - Considerations on the security and safety of these approaches are often neglected [2].
  - Although there are no confirmed cases of successful adversarial attacks against real world targets, the situation is likely to change as ML methods become commonplace.
  - **REMEMBER: attackers are attracted by what is "popular"!**

- ML models represent just a component within a system, and they **can** (and they **will**) be compromised

Takeaway: adversarial attacks will be exploited by expert attackers when ML becomes embedded into autonomous systems!

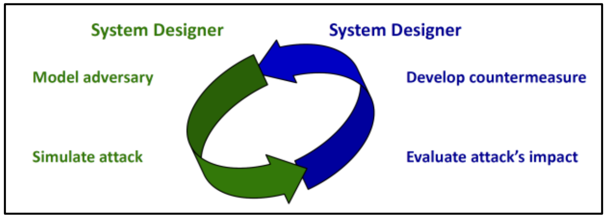
Future AICA agents represent an enticing target for next-generation attackers, who will resort also to Adversarial ML approaches.

UNIVERSITÄT LIECHTENSTEIN
Giovanni Apruzzese, PhD
giovanni.apruzzese@uni.li

[2]: Ram Shankar Siva Kumar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann and Sharon Xia. "*Adversarial Machine Learning--Industry Perspectives*." Proc. IEEE Secur. Privacy Workshops (2020).

Solution – What new approach should be adopted?

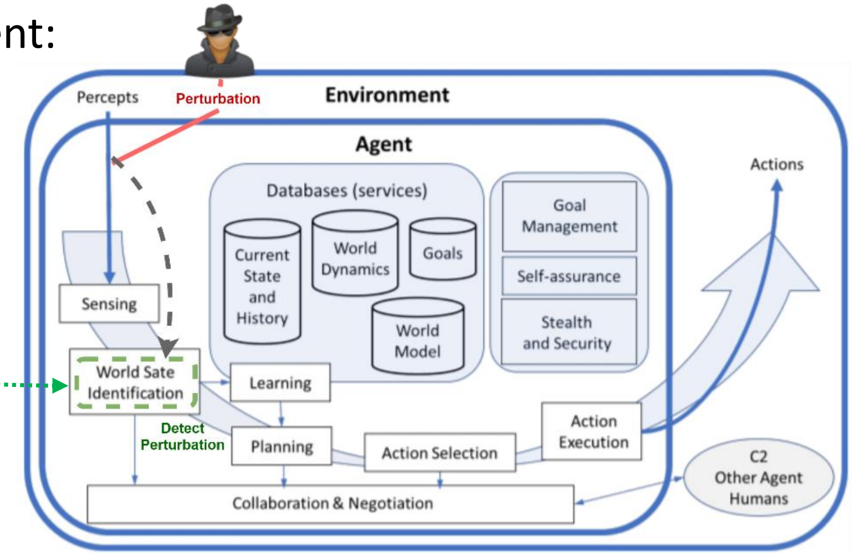
- When applying ML techniques to solve *any* task, it is important to adopt a <u>proactive</u> defensive approach [3].



- The goal is developing AICA agents that:
  - are capable of detecting novel and evasive attacks (e.g., autonomous malware),
  - but that are also <u>resilient</u> against adversaries that aim to thwart the ML model integrated and leveraged by the agent.

We should devise agents that use "Secure-by-design" ML models.

- Imperatives when deploying a ML-based agent:
  - Model an adversary
  - Simulate the attack and evaluate its impact
  - Devise a suitable countermeasure



UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, PhD
giovanni.apruzzese@uni.li

[3] Biggio, Battista, and Fabio Roli. "*Wild patterns: Ten years after the rise of adversarial machine learning*." Pattern Recognition (2018)

Obstacles – What risks or uncertainties might this approach create?

- Existing countermeasures against adversarial attacks present some limitations [4]:
  - Re-training with adversarial samples (*adversarial learning*):

    Requires the availability and mainteance of (multiple) adversarial datasets.

  - Use feature sets that cannot be leveraged by attackers:

    Decreases the performance of the baseline ML component against non-adversarial samples

- Devising threat models against *any* possible attack variant is impossible. An attacker could potentially affect:
  - The capacity of the AICA to detect attacks
  - The response executed by the AICA to a given input
  - The process of data-collection for continuous retraining of the AICA
  - The reporting process of the AICA to the human operators
  - …

    All of the above can be affected in different ways, which can result in different outcomes

UNIVERSITÄT LIECHTENSTEIN
Giovanni Apruzzese, PhD
giovanni.apruzzese@uni.li

[4]: Giovanni Apruzzese, Michele Colajanni, Luca Ferretti and Mirco Marchetti "*Addressing Adversarial Attacks Against Machine Learning Security Systems*" Proc. NATO International Conference on Cyber Conflicts (2019).

Course of Action – What’s the roadmap to success?

- **Key point**: do not aim to fight <u>all</u> attacks
  - Development of *realistic* threat models
  - Evaluation of proposed ML methods for AICA agents in *realistically feasible* adversarial environments

- When devising countermeasures, ensure that the baseline performance of the ML-component does not degrade excessively
  - In case of degradation, consider and evaluate the tradeoff

- Even if it is not possible to consider all possible adversarial scenarios and even if no countermeasure is effective, <u>at least:</u>
  - Identify the potential weaknesses
  - Evaluate how they could be exploited
  - Notify the users of these risks

> The worst scenario is having a *rogue* AICA that makes "smart" incorrect decisions, without suspecting that an opponent may have compromised or taken control of it through adversarial attacks.

UNIVERSITÄT LIECHTENSTEIN
Giovanni Apruzzese, PhD
giovanni.apruzzese@uni.li