SSF CLAS
Cyber-security for learning and control systems

Alexandre Proutiere
KTH Royal institute of Technology
January 2023

Machine-Learning based systems under attack

1. Assess the vulnerability of ML-based systems

2. Detect attacks, and devise secure ML algorithms

3. Illustrate the concepts in a smart building testbed

K.H. Johansson

A. Proutiere

G. Dan

H. Sandberg

M. Molinari

V. Ctekovic

Machine Learning under attack?

At training time

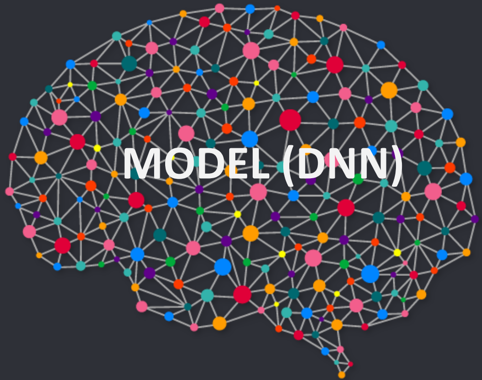
Machine Learning under attack?

At training time

Slightly modify the data in an ***adversarial*** manner

Machine Learning under attack?

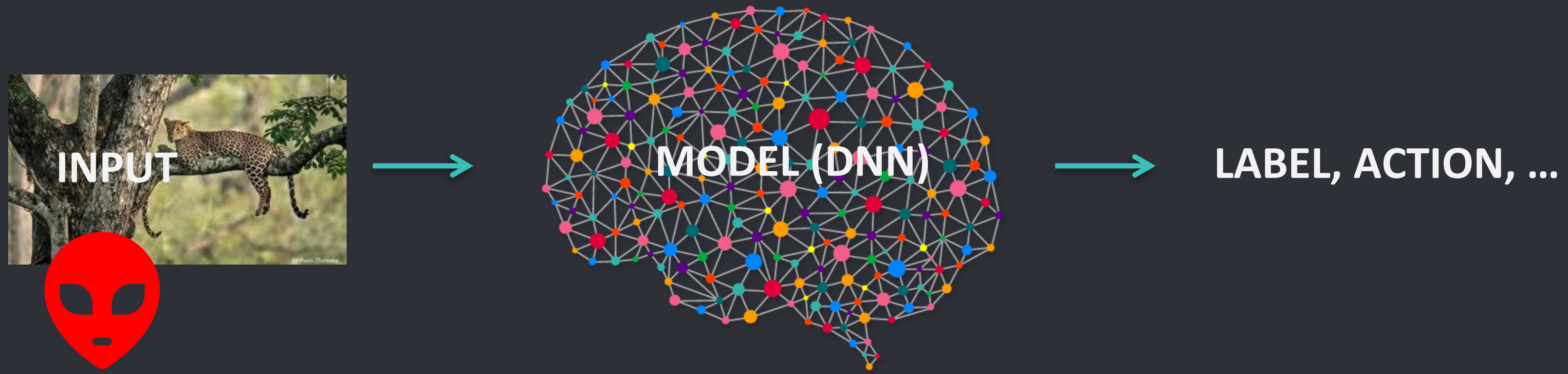
At test time



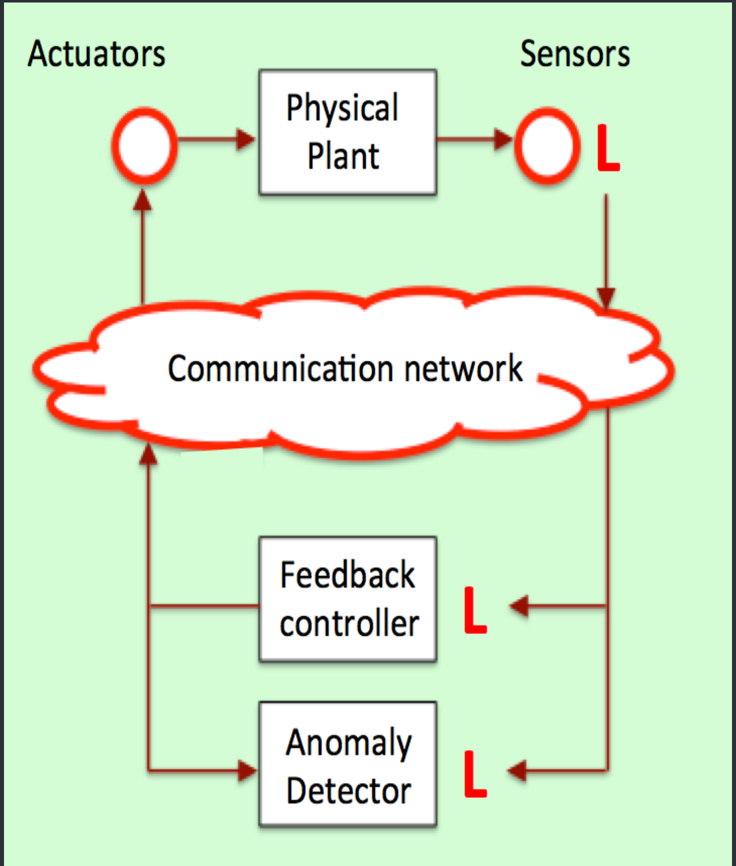
LABEL, ACTION, …

Machine Learning under attack?

At test time



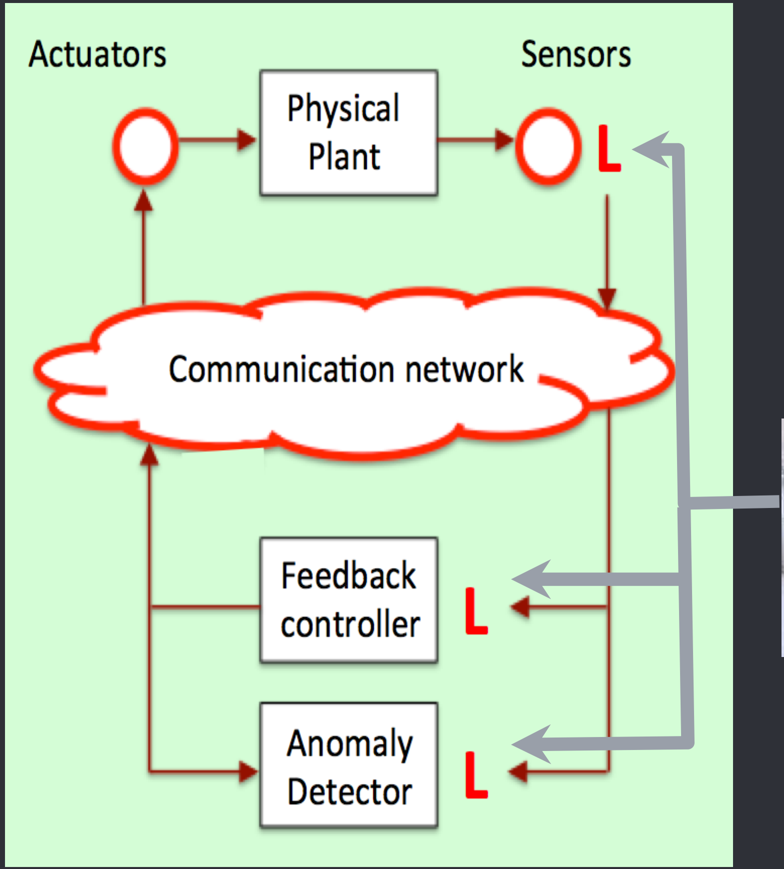
Slightly modify the input to the model in an ***adversarial*** manner



A generic ML-aided control system

L indicates ML components

a. ML helps interpreting sensor measurements
b. ML helps adapting the control (decisions taken) to a (partially) unknown system
c. ML helps building anomaly detectors

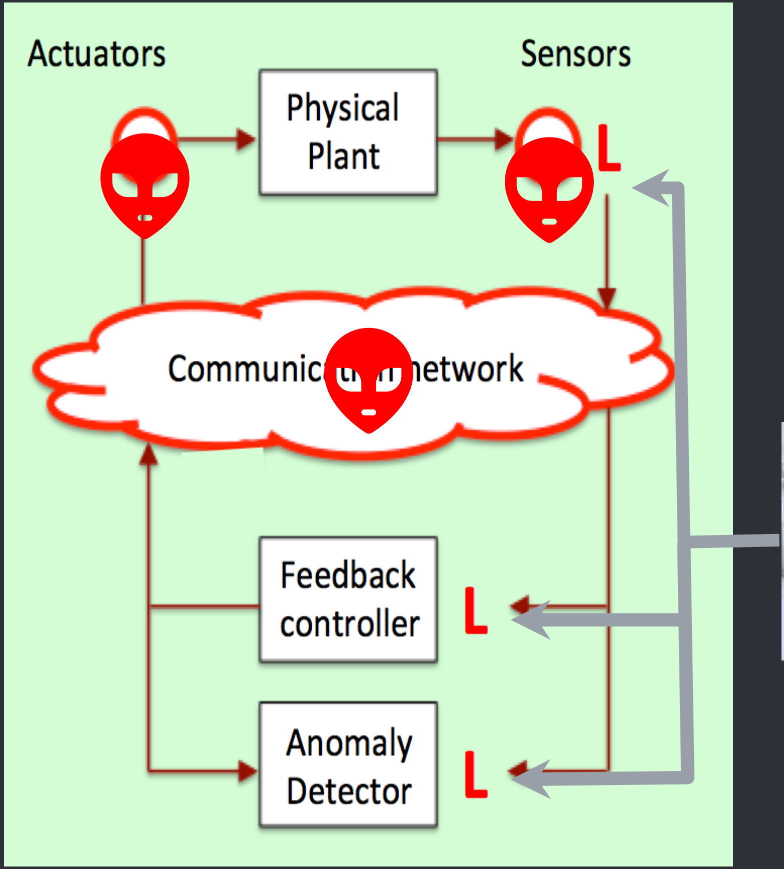


A generic ML-aided control system

ML methods most often comes with external datasets

A generic ML-aided control system

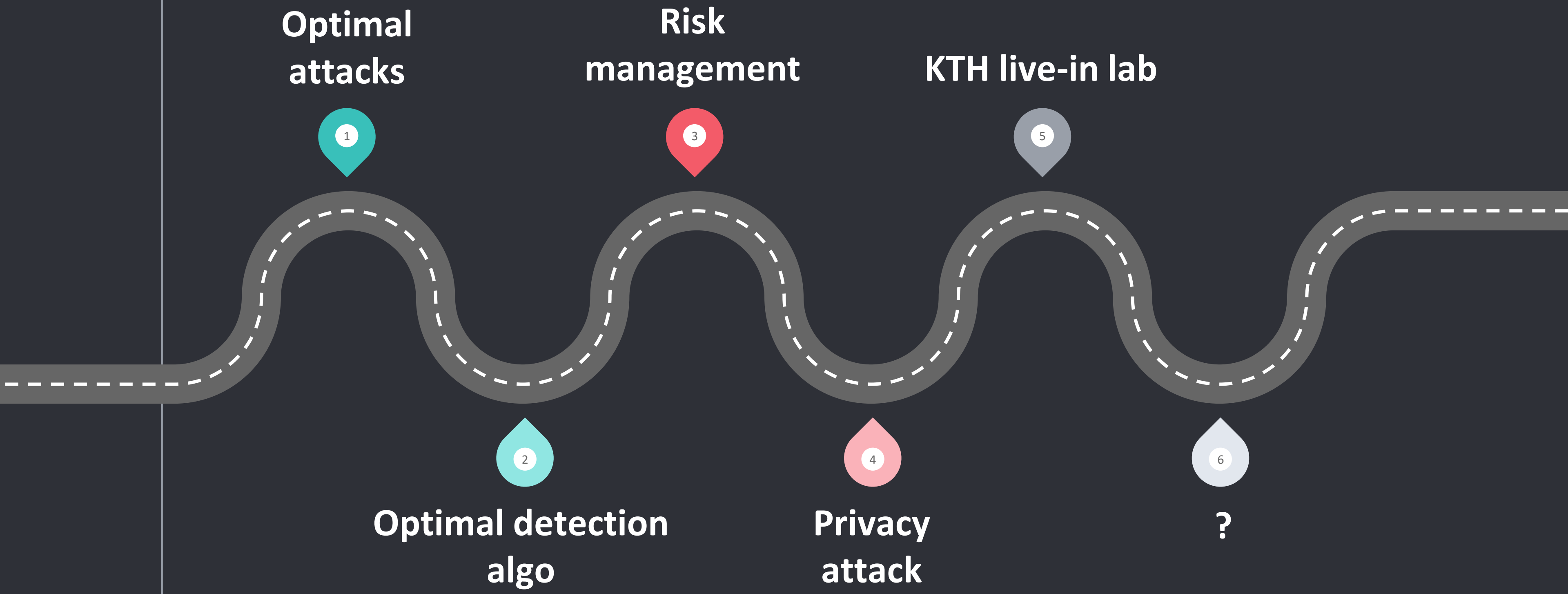
Possible points of attack



Attack at test time

Attack at training time

CLAS results in ML-based controlled systems



RESULTS SO FAR

Achievement snapshots

A - What are the threats and their potential impact?

- Optimal (or worse) attacks and their impact on
  - Reinforcement Learning policy at test time
  - Data-driven control policy at test and / or training time
  - Remote and distributed state estimation

B - Securing learning algorithms

- Secure multi-sensor estimation mechanisms
- Worst-case (adversarial) ML algorithms (e.g. regression)

C - Securing ML-aided control systems

- Secure Reinforcement Learning algorithms
- Secure platooning

Achievement snapshots

D - How can we evaluate risks and allocate defense resources accordingly?

- Game theoretical framework for risk management in advanced persistent threat

The Live-in Lab

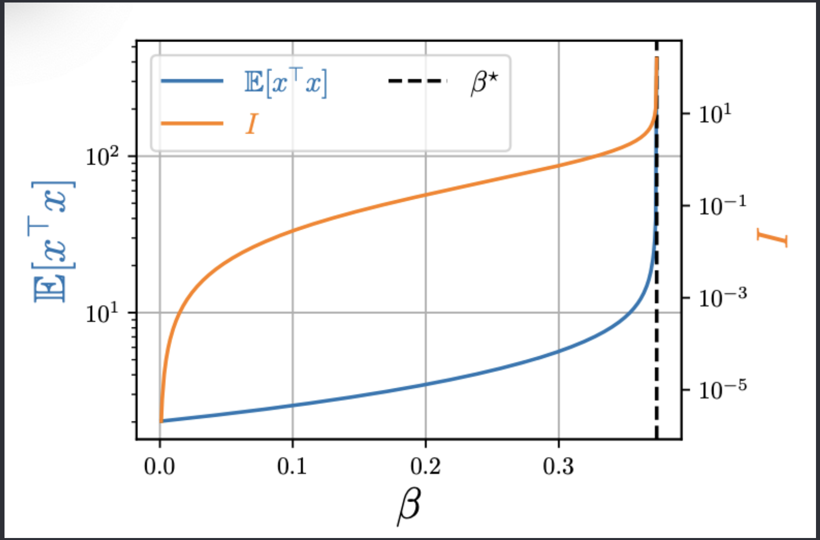
- Analysis of privacy attacks with data generated from the virtual testbed
- Co-simulation Environment to test state-of-the-art ML libraries for control
- Analysis of weak links in the data flows in the Live-In Lab

> 60 published papers

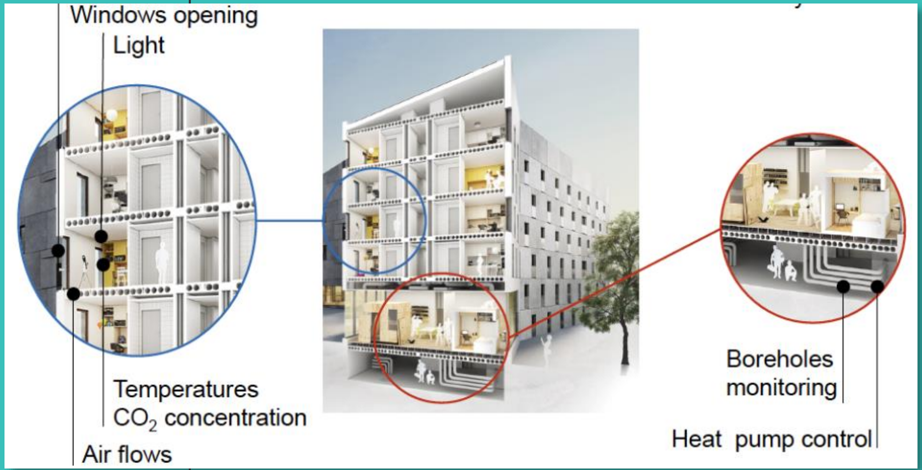
Example 1: Optimal attack / detection of RL policies

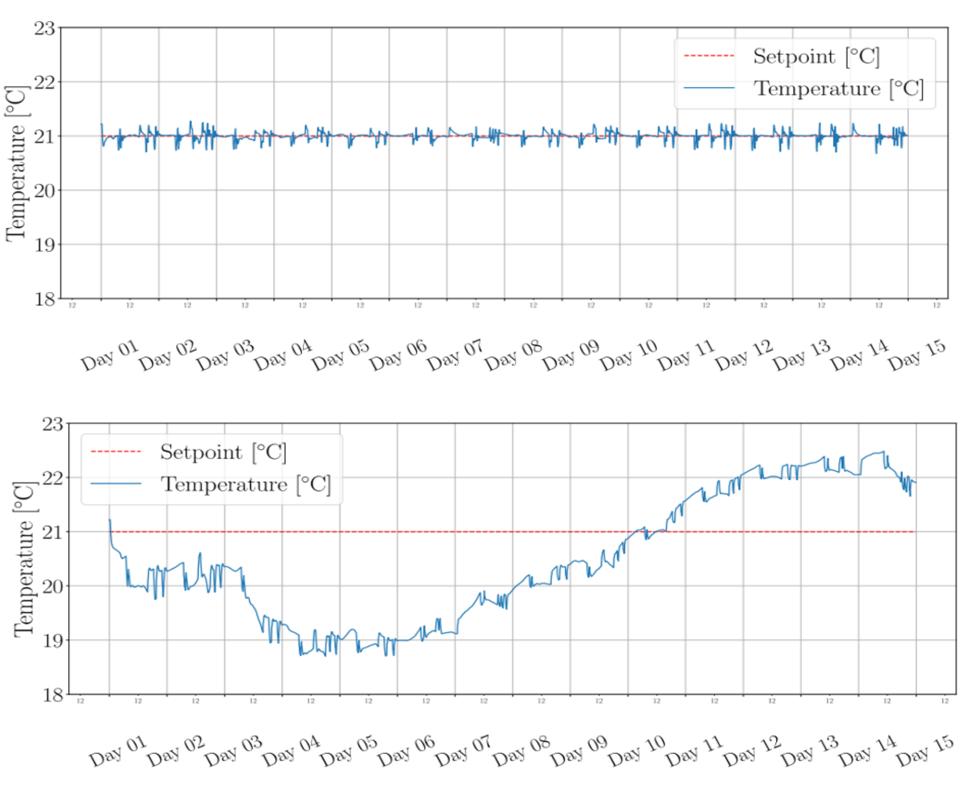
Attacking at test-time a control policy obtained through deep RL

1. Maximal detection rate of an attack π: $\mathbb{P}[\text{detect}] = e^{-I(\pi)}$
2. Optimal attack : $\min_{\pi} R(\pi) \; s.t. \; I(\pi) \geq \gamma$



Example 2: **KTH Live-in lab** (privacy and security)





Desired collaborations

1. How secure and robust is your ML system?

1. Contact: alepro@kth.se