ADTM 2026 Conference

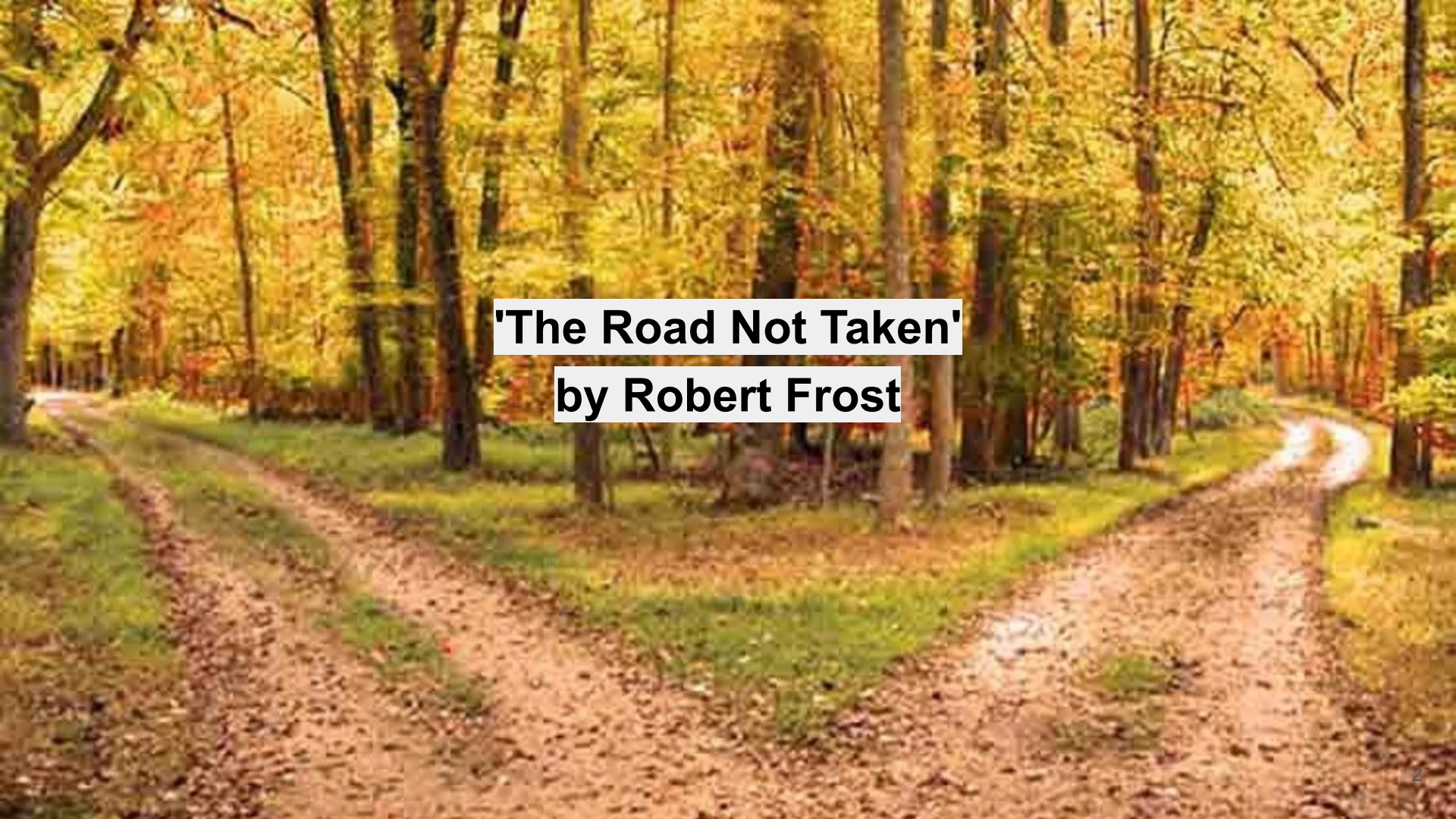
Digital Twin as an additional security protection layer in Industrial Control Systems (ICS)

Dr. Guy Tel-Zur
Ben-Gurion University of the Negev

gtelzur@bgu.ac.il

15.6.2026

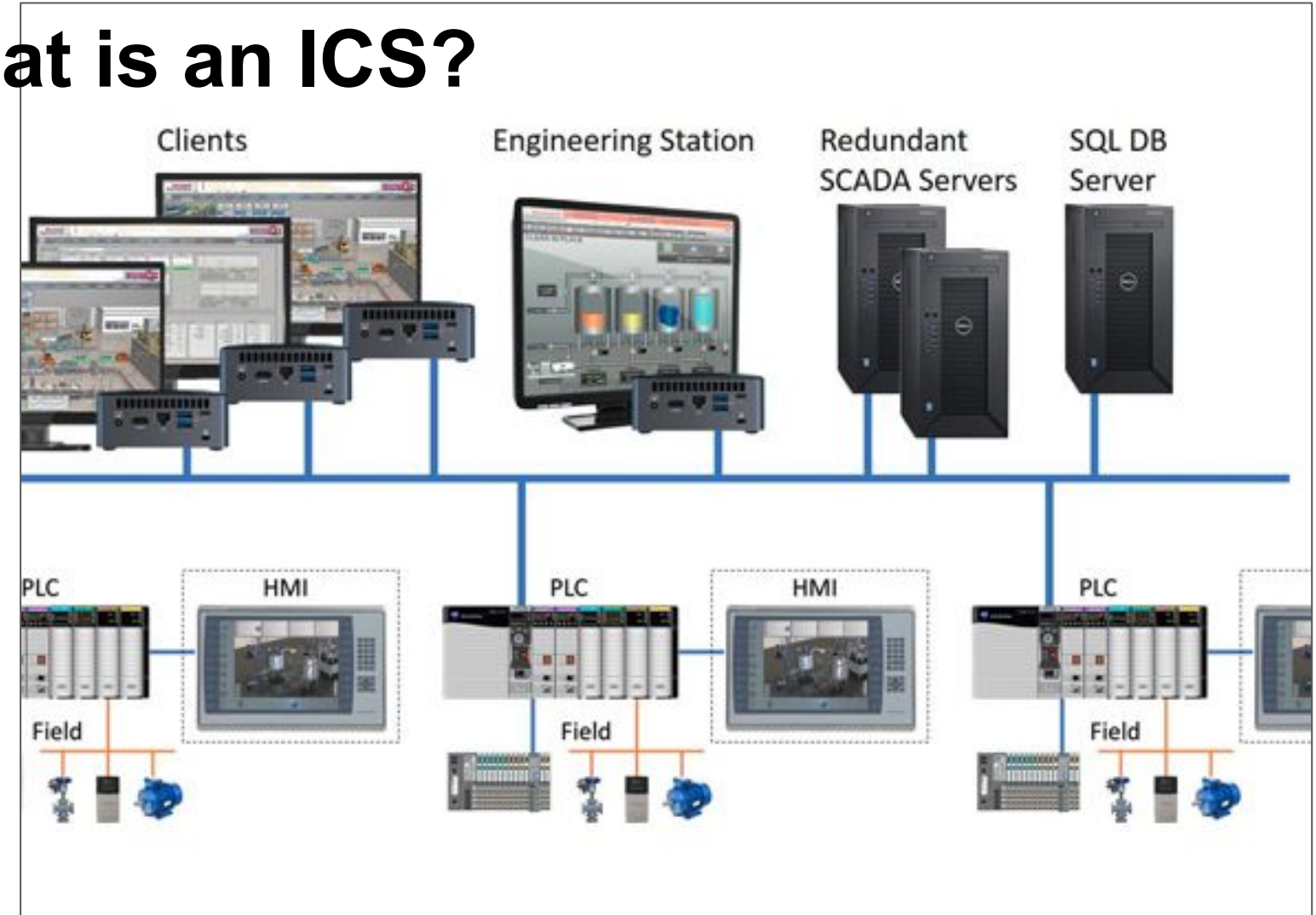
V7


'The Road Not Taken'
by Robert Frost

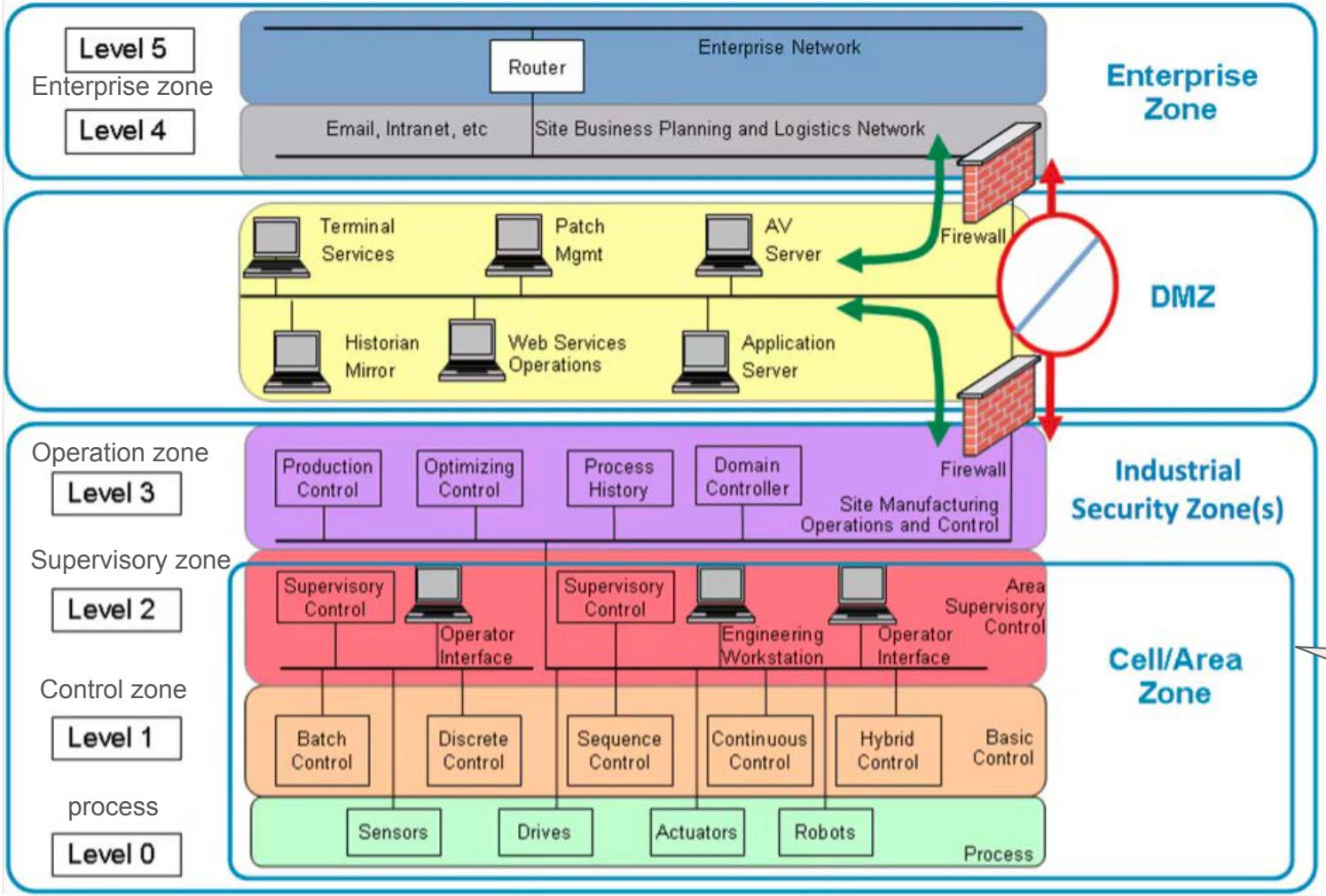
Talk agenda

- What is an ICS?
- Cybersecurity Risks
- What is a Digital Twin (DT)?
- How a Digital Twin can help protecting an ICS?
- An example: Reinforcement Learning AI-assisted DT
- Project: 4th year engineering students final project
- Outlook and Summary

What is an ICS?



Purdue Model



This work

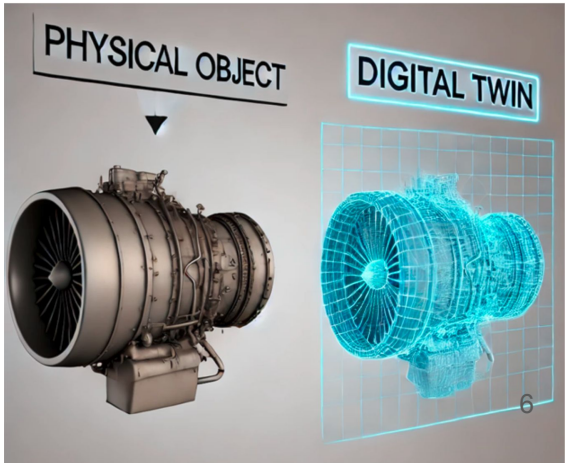
Credit: https://www.cisco.com/site/us/en/learn/topics/security/what-is-ot-security.html

Digital Twin

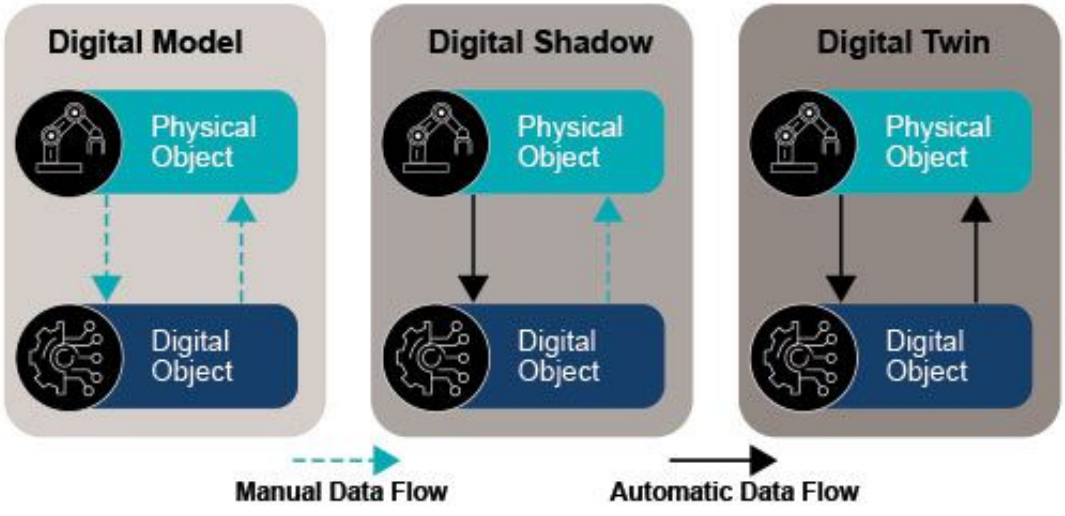
A virtual replica of a physical system that updates in real-time

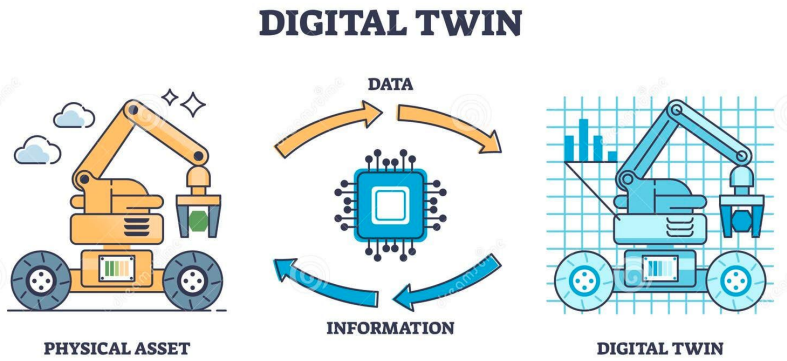
ISO/IEC 30173: *“Digital representation of a target entity with data connections that enable convergence between the physical and digital states at an appropriate rate of synchronization”*

The Digital Twin mirrors the real object's:
- Current state
- Behavior
- Performance
- Even predicts future states



Digital Twin





Digital Twin ≠ Simulation

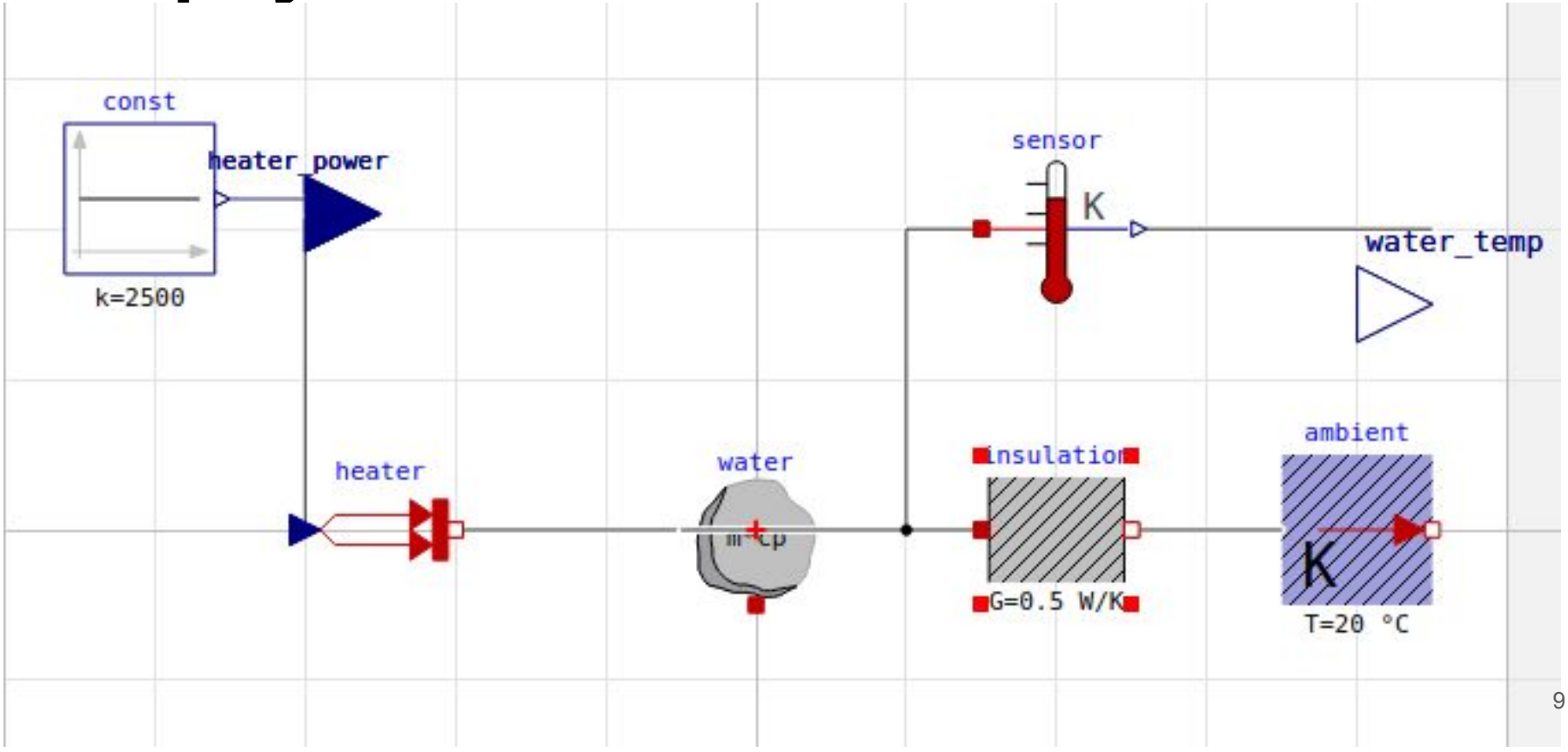
2 issues:

- How to use the Digital Twin for protection of physical system?
- How to protect the Digital Twin and the access to its models to make sure its main functions?

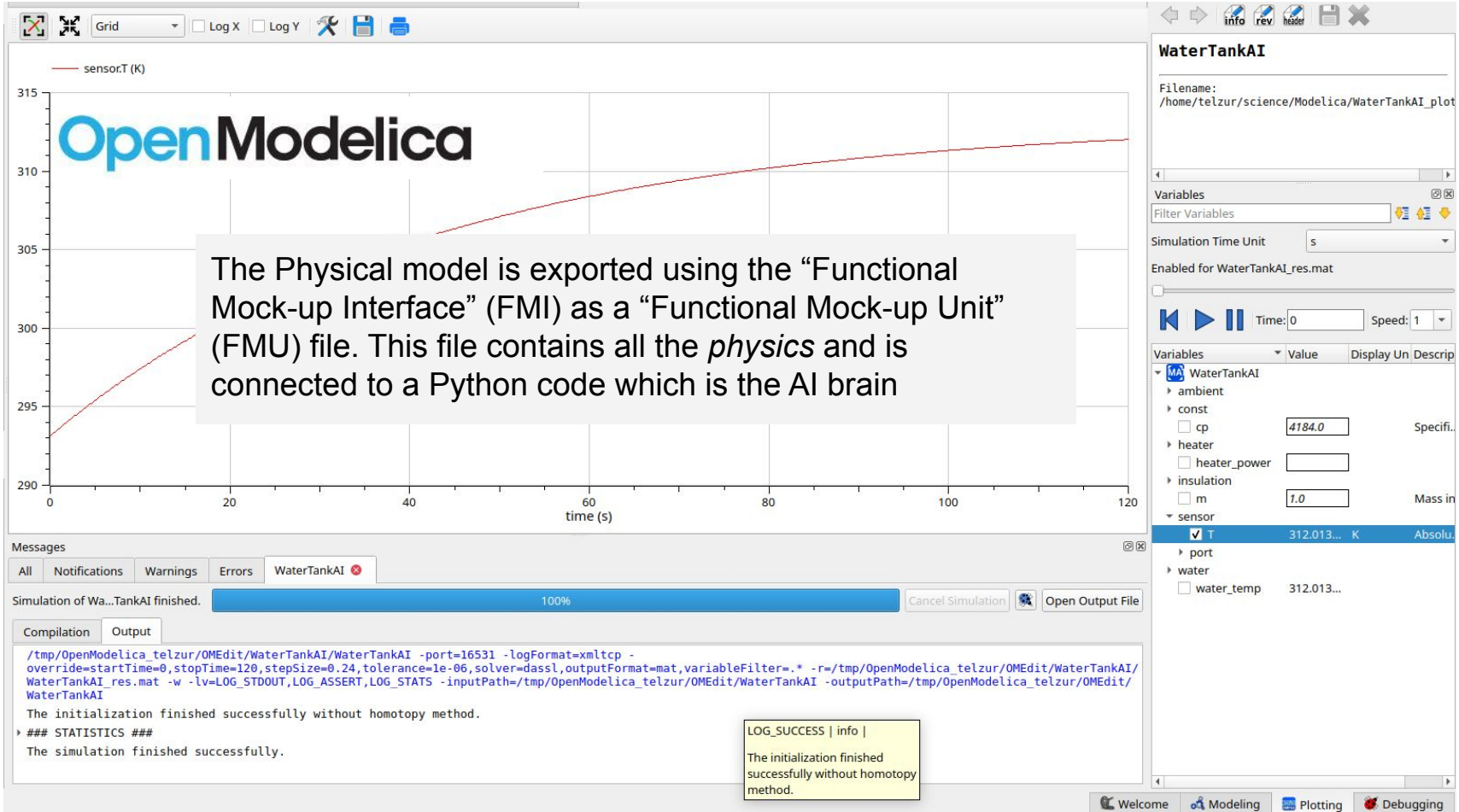
Water-Heating Control with an AI-Assisted Controller and an FMU-Based Digital Twin

- A kettle with water
- Heater: up to 2500W
- Heating profile: Starting from room temperature (20°C), heat the water to 40°C within 120 seconds, then keep the temperature steady at 40°C for an additional 60 seconds and then let it naturally cool down.
- Water mass (1 Kg. ↣ 1 Liter)
- Heat loss via the kettle's sides to the outside environment
- Assumptions: A simple lumped-parameter model with a uniform water/pot temperature model (i.e. no boiling, no bubbles, no turbulence, etc').

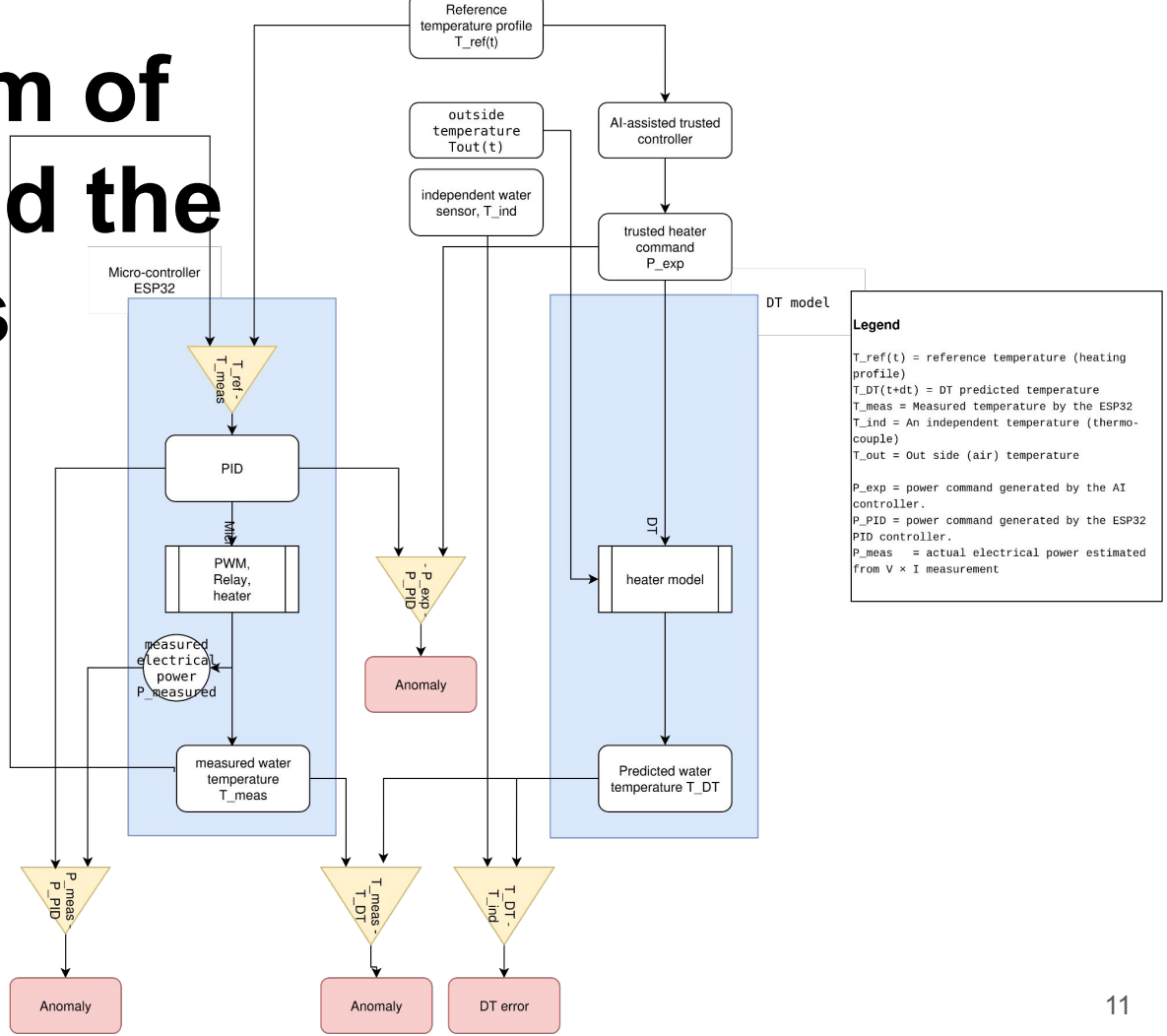
The physical model in Modelica



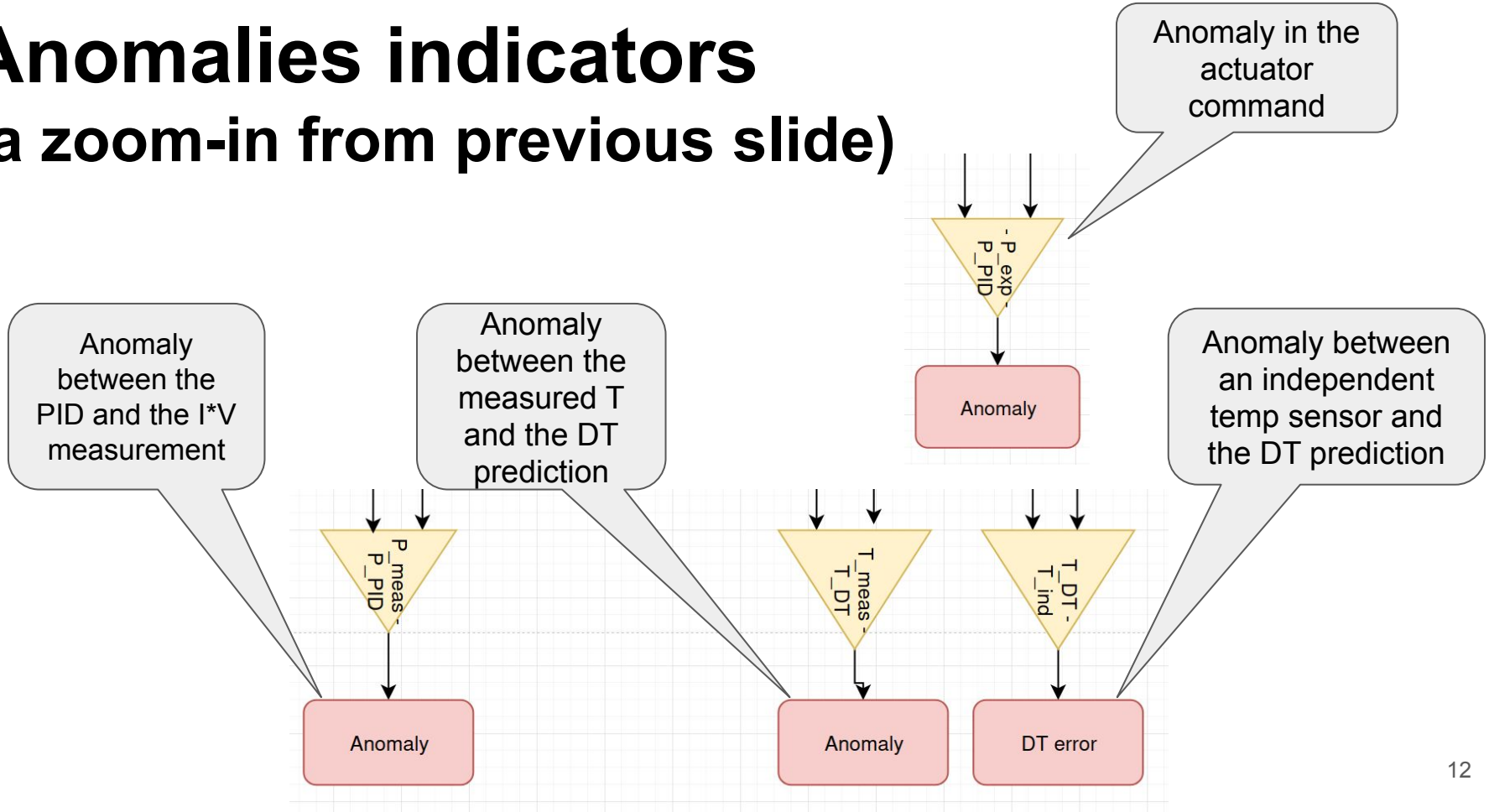
Model simulation in OpenModelica



A Block diagram of the Physical and the Virtual systems

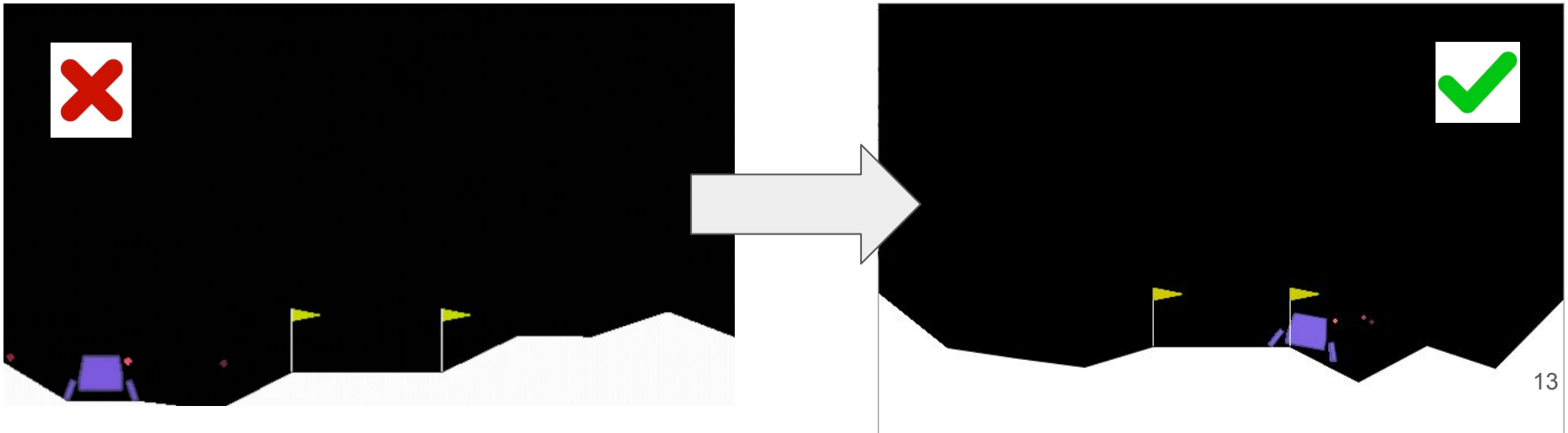


Anomalies indicators
(a zoom-in from previous slide)



Selecting an AI Method for the Trusted Controller

- MLP
- LSTM
- Other models...?
- Reinforcement Learning (RL) models:
  - **Proximal Policy Optimization (PPO)** - failed
  - **Advantage Actor-Critic (A2C)** - not evaluated; supports both discrete and continuous policies
  - **Soft Actor Critic (SAC)** - success. May be an overkill for this specific project (continuous)



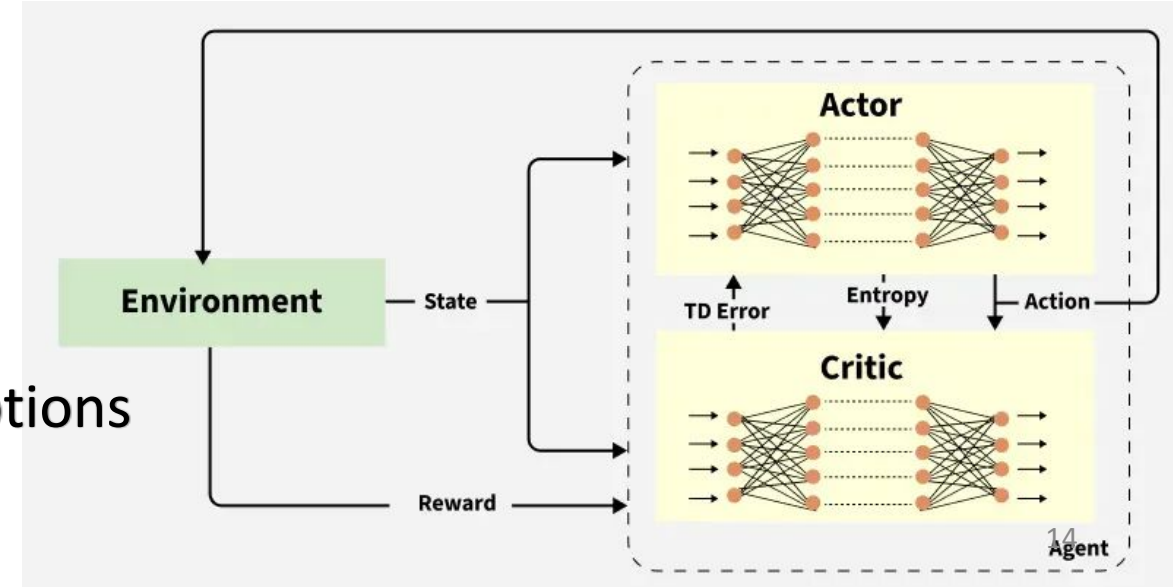
Soft Actor-Critic (SAC): Core Idea

- Actor: chooses actions probabilistically
- Critic: evaluates actions

Key idea: Maximize Reward + α × Randomness (Entropy)

Why?

- Encourages exploration
- Avoids getting stuck
- Keeps multiple good options

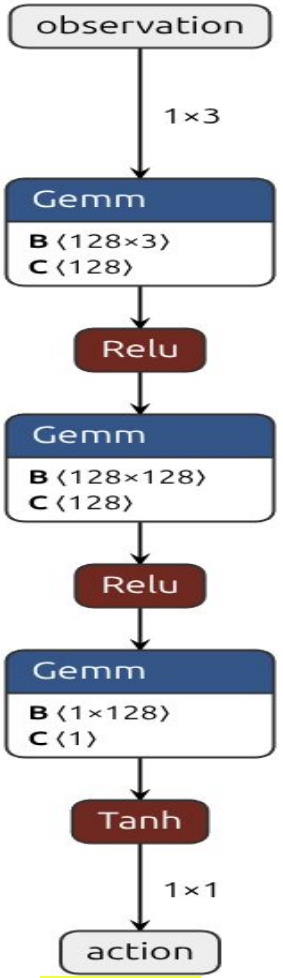


The NN structure

#Parameters = Inputs X Outputs + Biases

1. **Layer 1 (Gemm 1):** Inputs: 3 ⟶ Outputs: 128
   - Weights: 3 × 128 = 384
   - Biases: 128
   - **Subtotal: 512**
2. **Layer 2 (Gemm 2):** Inputs: 128 ⟶Outputs: 128
   - Weights: 128 × 128 = 16,384
   - Biases: 128
   - **Subtotal: 16,512**
3. **Layer 3 (Gemm 3):** Inputs: 128 ⟶ Outputs: 1
   - Weights: 1 × 128 = 128
   - Biases: 1
   - **Subtotal: 129**

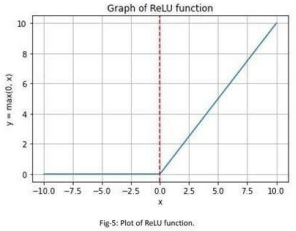
Total number of trainable parameters: 512 + 16,512 + 129 =17,153



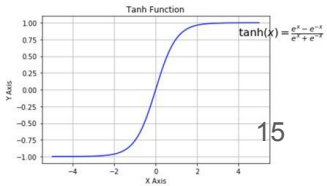
GEMM = General Matrix Multiply

Relu = Rectified linear unit

$$f(x) = max(0, x) = \begin{cases} x_i, & if\ x_i \geq 0 \\ 0, & if\ x_i < 0 \end{cases}$$

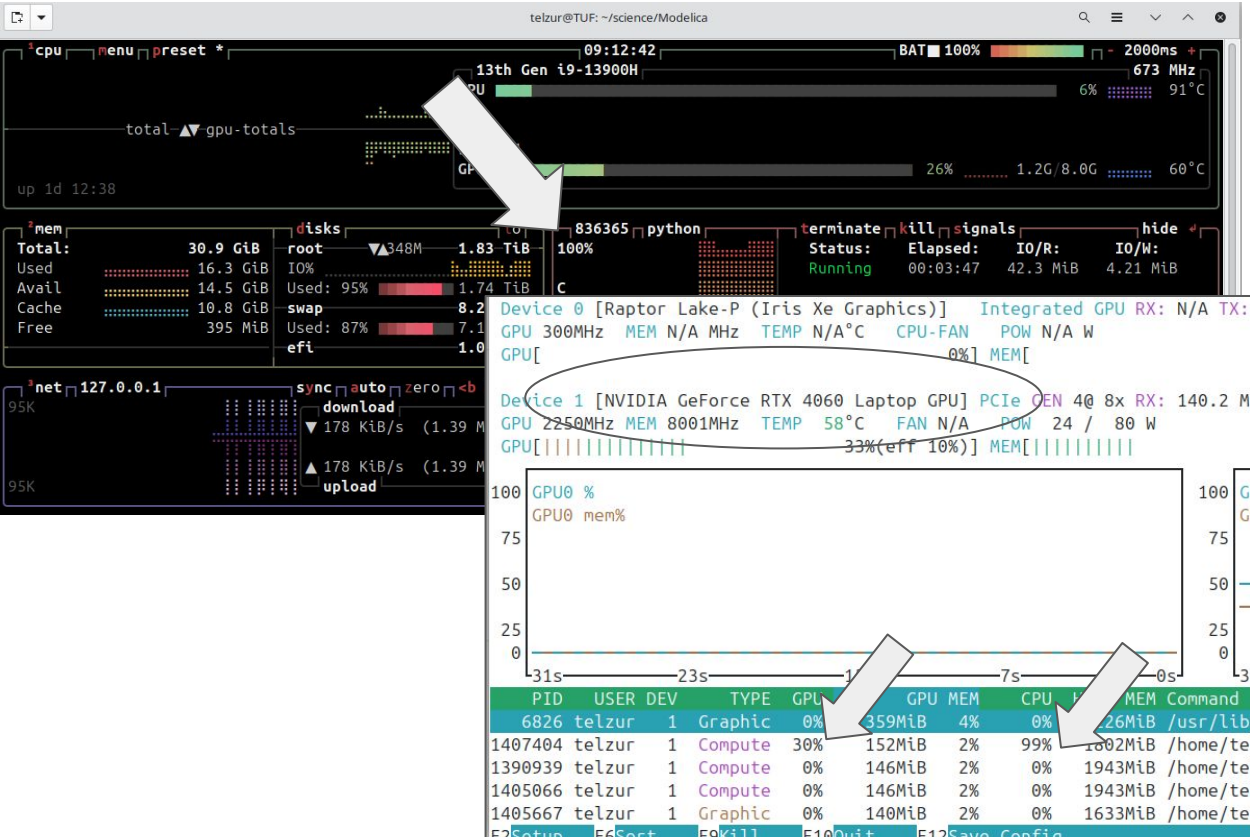


Tanh



SAC training both on the CPU and the GPU

```
real    9m45.703s
user    9m22.963s
sys     0m9.045s
(mypython) telzur@TUF:~/science/Modelica$
```



RL SAC Coding in Python

Gymnasium PyTorch

Gymnasium is a standard interface for reinforcement learning (RL) environments.

- A unified API for RL environments (like a "playground" where agents learn)
- A `reset()` → `step()` loop: observation, action, reward, done
- Popular environments (CartPole, MountainCar, Atari, etc.)
- Originally forked from OpenAI's Gym, now maintained separately

Stable Baselines 3 (SB3) is a collection of pre-implemented RL algorithms.

- Ready-to-use algorithms: PPO, A2C, DQN, SAC, TD3, etc.
- Clean, well-documented implementations
- Works on top of Gymnasium environments
- Handles the training loop, logging, and evaluation

Soft Actor-Critic:
Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

Tuomas Haarnoja [1] Aurick Zhou [1] Pieter Abbeel [1] Sergey Levine [1]

Abstract

Model-free deep reinforcement learning (RL) algorithms have been demonstrated on a range of challenging decision making and control tasks. However, these methods typically suffer from two major challenges: very high sample complexity and brittle convergence properties, which necessitate meticulous hyperparameter tuning. Both of these challenges severely limit the applicability of such methods to complex, real-world domains. In this paper, we propose soft actor-critic, an off-policy actor-critic deep RL algorithm based on the maximum entropy reinforcement learning framework. In this framework, the actor aims to maximize expected reward while also maximizing entropy. That is, to succeed at the task while acting as randomly as possible. Prior deep RL methods based on this framework have been formulated as Q-learning methods. By combining off-policy updates with a stable stochastic actor-critic formulation, our method achieves state-of-the-art performance on a range of continuous control benchmark tasks, outperforming prior on-policy and off-policy methods. Furthermore, we demonstrate that, in contrast to other off-policy algorithms, our approach is very stable, achieving very similar performance across different random seeds.

1. Introduction

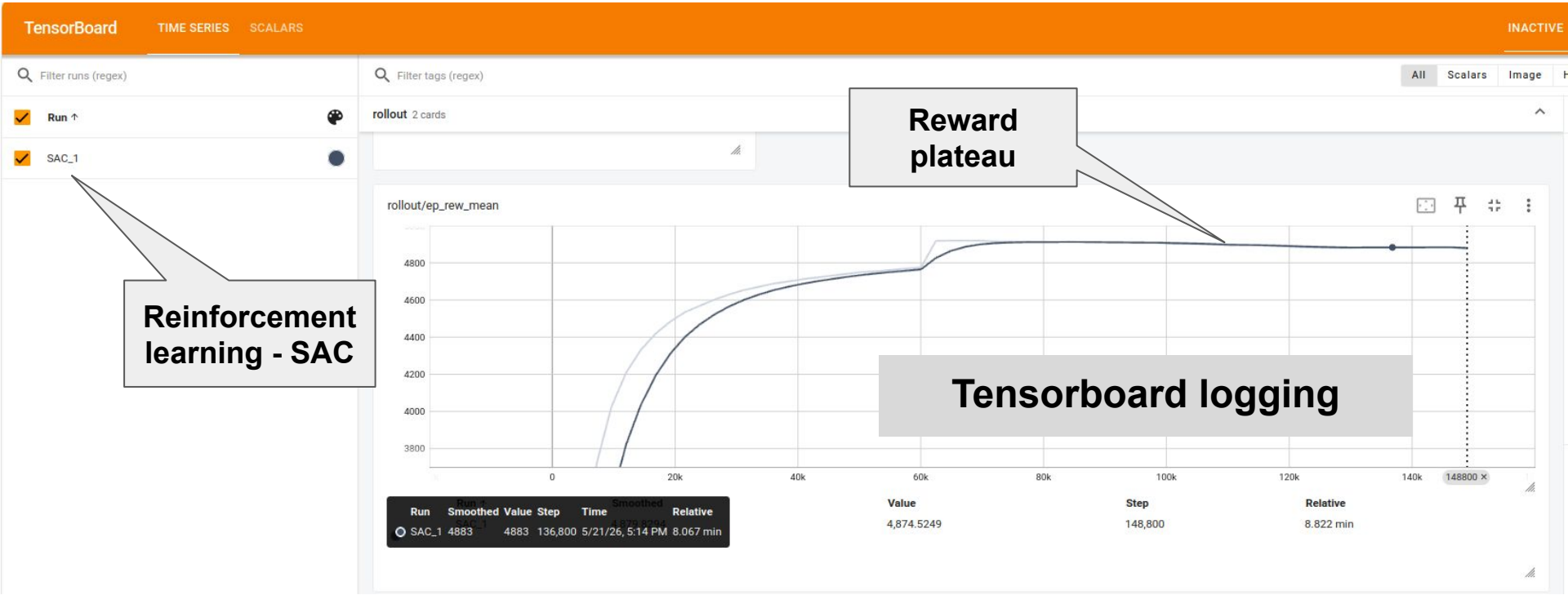
Model-free deep reinforcement learning (RL) algorithms have been applied in a range of challenging domains, from games (Mnih et al., 2013; Silver et al., 2016) to robotic control (Schulman et al., 2015). The combination of RL and high-capacity function approximators such as neural networks holds the promise of automating a wide range of decision making and control tasks, but widespread adoption of these methods in real-world domains has been hampered by two major challenges. First, model-free deep RL methods are notoriously expensive in terms of their sample complexity. Even relatively simple tasks can require millions of steps of data collection, and complex behaviors with high-dimensional observations might need substantially more. Second, these methods are often brittle with respect to their hyperparameters: learning rates, exploration constants, and other settings must be set carefully for different problem settings to achieve good results. Both of these challenges severely limit the applicability of model-free deep RL to real-world tasks.

One cause for the poor sample efficiency of deep RL methods is on-policy learning: some of the most commonly used deep RL algorithms, such as TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017b) or A3C (Mnih et al., 2016), require new samples to be collected for each gradient step. This quickly becomes extravagantly expensive, as the number of gradient steps and samples per step needed to learn an effective policy increases with task complexity. Off-policy algorithms aim to reuse past experience. This is not directly feasible with conventional policy gradient formulations, but is relatively straightforward for Q-learning based methods (Mnih et al., 2015). Unfortunately, the combination of off-policy learning and high-dimensional, nonlinear function approximation with neural networks presents a major challenge for stability and convergence (Bhatnagar et al., 2009). This challenge is further exacerbated in continuous state and action spaces, where a separate actor network is often used to perform the maximization in Q-learning. A commonly used algorithm in such settings, deep deterministic policy gradient (DDPG) (Lillicrap et al., 2015), provides for sample-efficient learning but is notoriously challenging to use due to its extreme brittleness and hyperparameter sensitivity (Duan et al., 2016; Henderson et al., 2017).

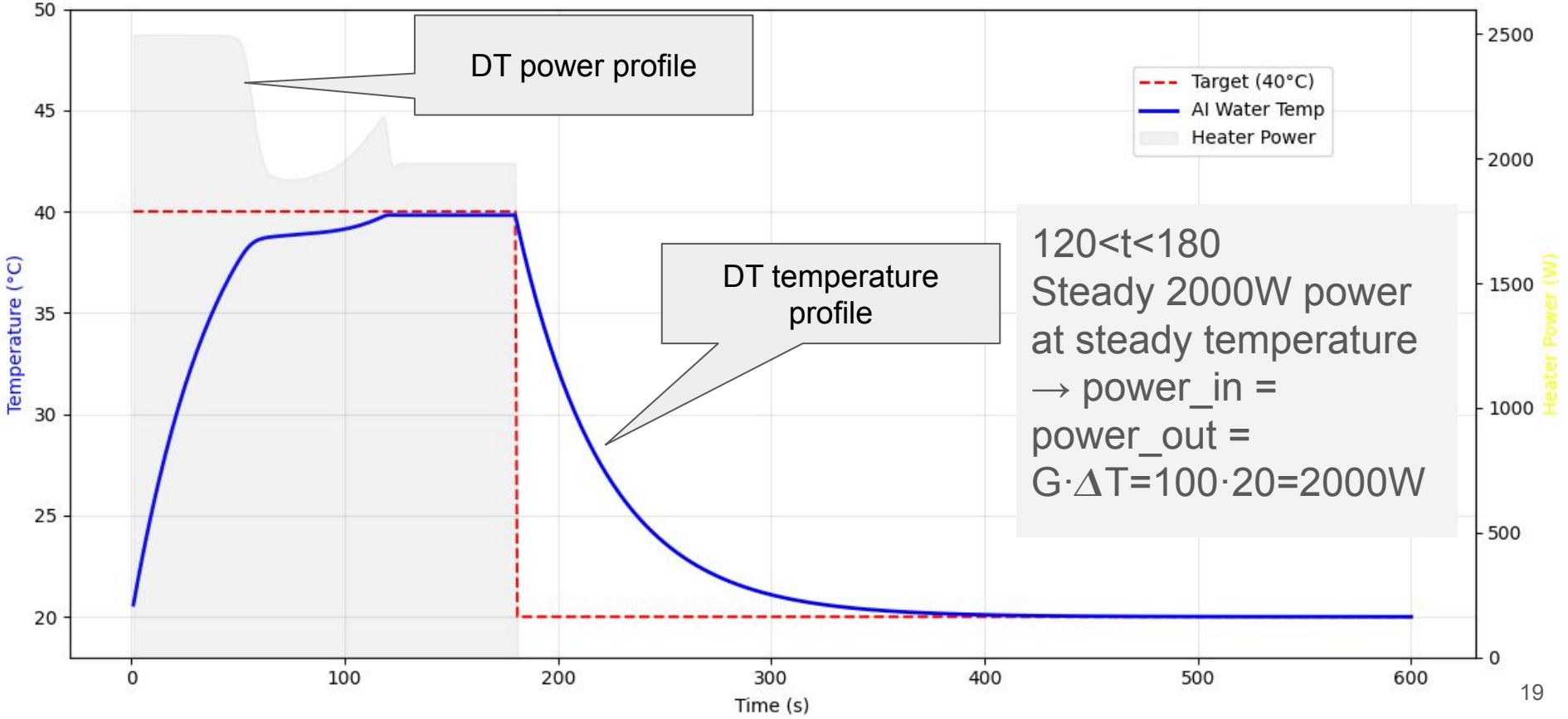
We explore how to design an efficient and stable model-free deep RL algorithm for continuous state and action spaces. To that end, we draw on the maximum entropy framework, which augments the standard maximum reward reinforcement learning objective with an entropy maximization term (Ziebart et al., 2008; Toussaint, 2009; Rawlik et al.,

[1]Berkeley Artificial Intelligence Research, University of California, Berkeley, USA. Correspondence to: Tuomas Haarnoja <haarnoja@berkeley.edu>.

Training the SAC-Based Trusted Controller



SAC Controller and Digital-Twin Response

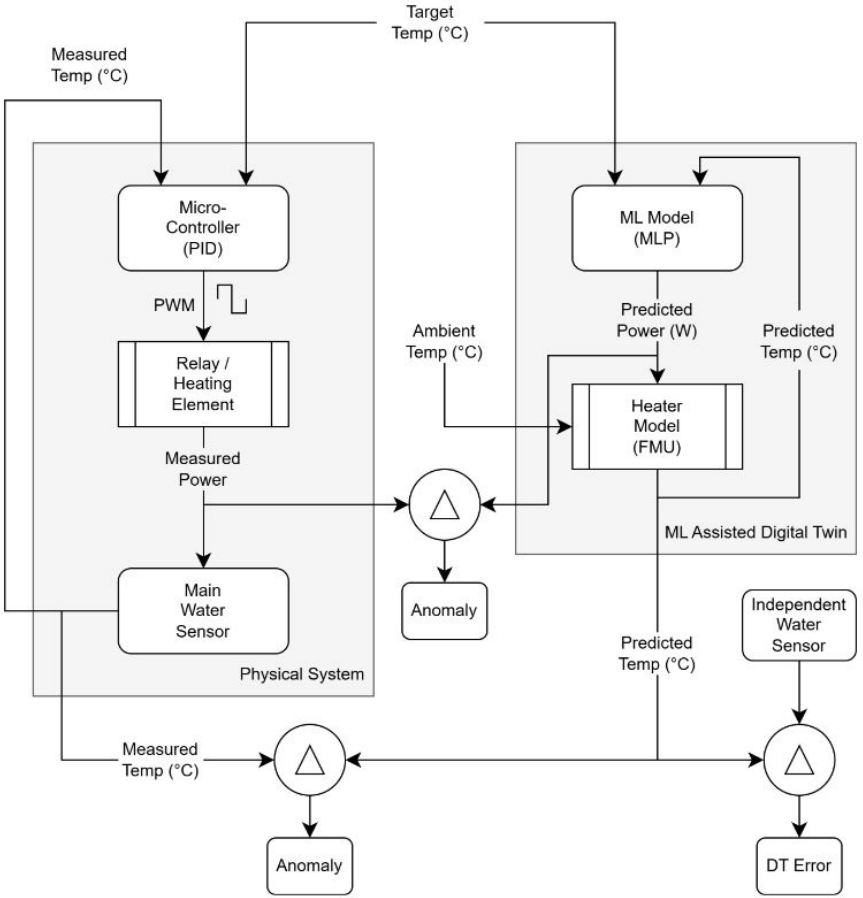


Final project. BGU 4th year engineering students project

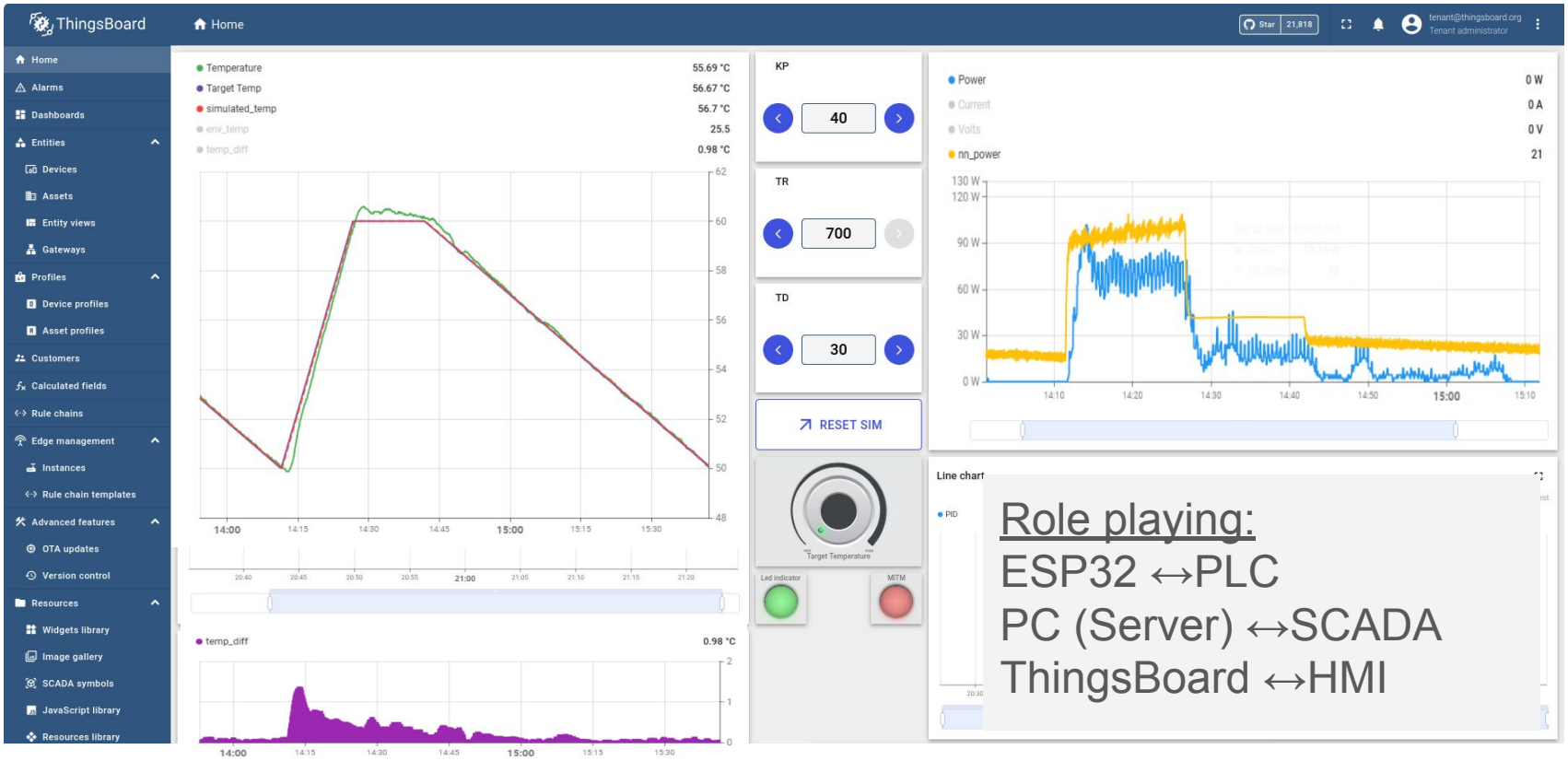
- Ori Dagan
- Elad Elmakias

Key features

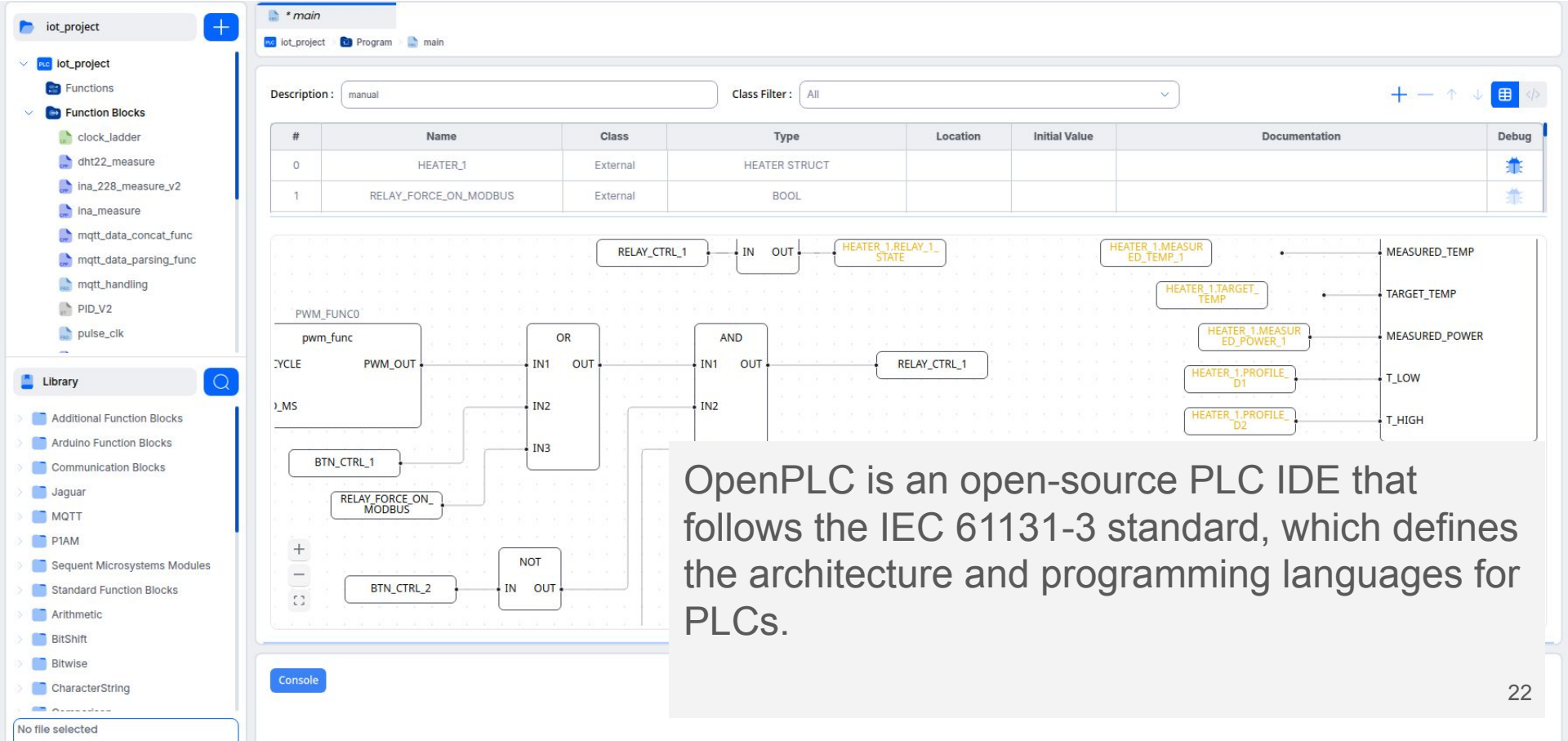
- Open Source tools
- OpenPLC
- ThingsBoard
- Real physical model. ESP32
- MLP
- Data logging



The HMI


ThingsBoard
Home
Star 21,818
tenant@thingsboard.org
Tenant administrator
Home
Alarms
Dashboards
Entities
Devices
Assets
Entity views
Gateways
Profiles
Device profiles
Asset profiles
Customers
Calculated fields
Rule chains
Edge management
Instances
Rule chain templates
Advanced features
OTA updates
Version control
Resources
Widgets library
Image gallery
SCADA symbols
JavaScript library
Resources library
Temperature 55.69 °C
Target Temp 56.67 °C
simulated_temp 56.7 °C
env_temp 25.5
temp_diff 0.98 °C
14:00
14:15
14:30
14:45
15:00
15:15
15:30
KP
40
TR
700
TD
30
RESET SIM
Target Temperature
Led indicator
MITM
Power 0 W
Current 0 A
Volts 0 V
nn_power 21
130 W
120 W
90 W
60 W
30 W
0 W
14:10
14:20
14:30
14:40
14:50
15:00
15:10
Line chart
PID
temp_diff 0.98 °C
Role playing:
ESP32 ↔PLC
PC (Server) ↔SCADA
ThingsBoard ↔HMI

OpenPLC



OpenPLC is an open-source PLC IDE that follows the IEC 61131-3 standard, which defines the architecture and programming languages for PLCs.

Future plans

Continuation projects at BGU and at SCE

Suggestions:2027

Advisers:

Dr. Guy Tel-zur

Project name (english)

Protecting IoT, IIoT, and ICS networks using a Digital Twin

Abstract(english)

This project focuses on the development of a Digital Twin (DT)-based monitoring system that adds an independent security layer to a conventional industrial control system. The DT models the expected physical behavior of the controlled process and continuously compares it with measurements from the real system. Significant discrepancies between the real system and the DT may indicate equipment malfunction, process anomalies, or cyberattacks. In such cases, the system will trigger an alert and display it at the Human–Machine Interface (HMI) station.

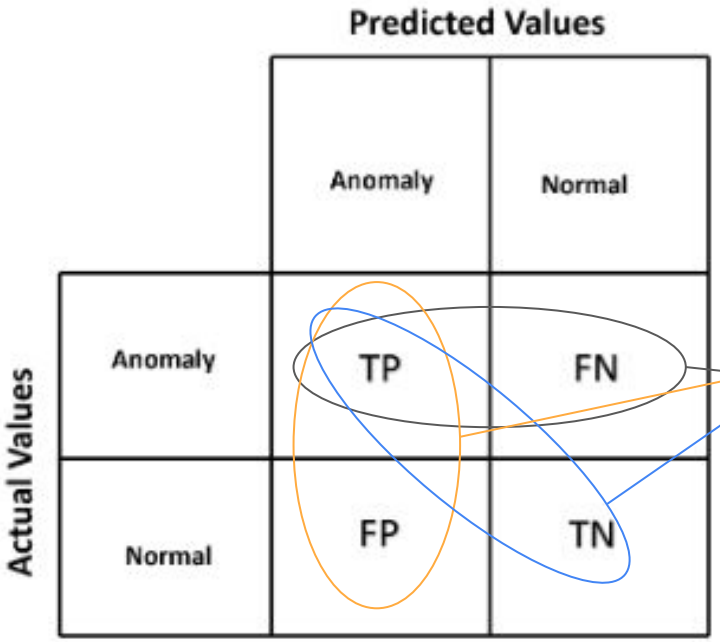
Project name (hebrew)

הגנת מערכות בקרה תעשיתית באמצעות תאום דיגיטלי

Abstract(hebrew)

מדמה את ההתנהגות הפיזית הצפויה של התהליך המבוקר ומשווה אותה באופן DT-המוסיפה שכבת אבטחה עצמאית למערכת בקרה תעשייתית קונבנציונלית. ה (DT) פרויקט זה מתמקד בפיתוח מערכת ניטור מבוססת תאום דיגיטלי עשויים להצביע על תקלה בציוד, אנומליות בתהליך או מתקפות סייבר. במקרים כאלה, המערכת תפעיל התראה ותציג אותה בתחנת ממשק DT-רציף למדידות מהמערכת האמיתית. פערים משמעותיים בין המערכת האמיתית ל אדם-מכונה (HMI).

Anomaly Detection metrics



Confusion matrix

$$Accuracy = \frac{TN + TP}{FN + FP + TN + TP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Summary

Innovation

- **100% open source** tools (ThingsBoard, OpenPLC, OpenModelica, Python, local AI model)
- **AI-assisted Digital Twin** - an improved robustness and independent security layer
- Reinforcement learning implementations

Next steps

- BGU 2027 4th year ECE continuation project
- SCE 2027 CS final projects (including physical HW to be tested in the new Cyber OT laboratory.

Questions?

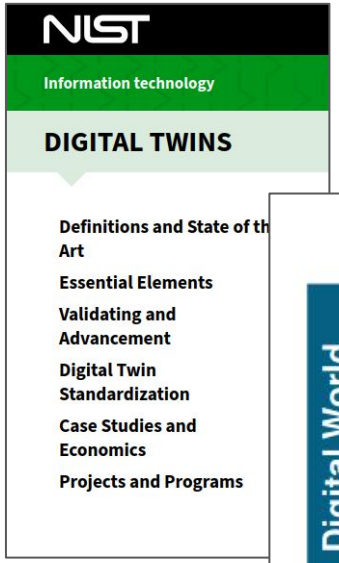
E-mail: gtelzur@bgu.ac.il

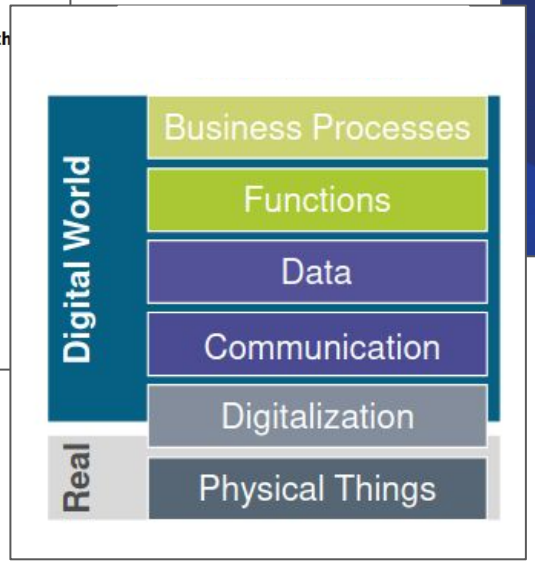
Linkedin: linkedin.com/in/telzur

That's it!

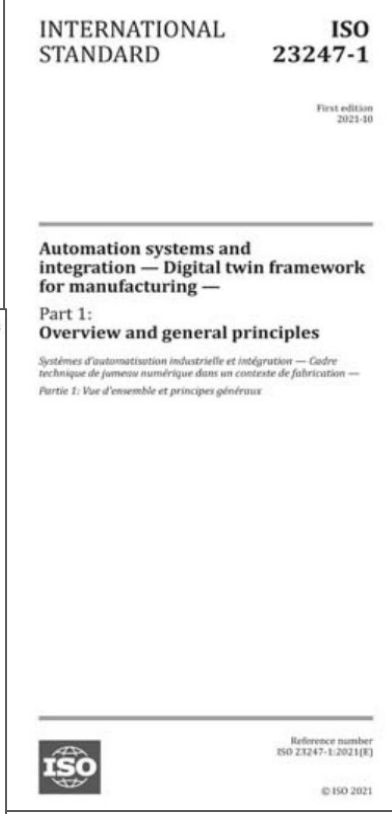
Extra Slides

Regulation





RAMI 4.0 - **R**eference **A**rchitectural **M**odel for **I**ndustrie 4.0



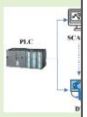
Literature

IEEE Sensors Council IEEE SENSORS JOURNAL, VOL. 25, NO. 10, 15 MAY 2025 17739

PLC-MDT: A Framework for Detecting Anomalies With Digital Twins of Industrial Control Systems

Zidong Xu, Zhenyong Zhang, and Hetu Zheqiu

***Abstract*—The attack events (such as the Stuxnet and BlackEnergy) that targeted the industrial control system (ICS) have validated its vulnerability to cyber intrusions. The prevention of ICS from cyberattacks is undoubtedly important and practical as it operates dangerous equipment and produces products that people need. Traditional anomaly detection methods, which rely heavily on SCADA data, are often unreliable due to the limitation of such data in accurately reflecting real-time system states under attacks. Therefore, we propose PLC-MDT, a novel anomaly detection method based on digital twins (DTs), exploiting the modern communication technology and computing resources of ICS. PLC-MDT does not depend solely on SCADA data; instead, it leverages the interaction between DT and real-world ICS, utilizing both DT a detection. Specifically, for PLC-MDT, the prediction model in DT is trained base values are used as input to the mechanism model in DT to generate simulated guarantees real-time data collected from field sites for DT. Therefore, PLC-M difference between the predicted and real-time data, revealing control channel c a threshold. The effectiveness of PLC-MDT is demonstrated through experimer potential for real-time detection of anomalies in ICS.**



***Index Terms*— Anomaly detection, digital twin (DT), industrial control systems**

I. Introduction

INDUSTRIAL control systems (ICSs), which are widely deployed in critical infrastructures such as metallurgy, electricity grids, nuclear plants, and gas distribution systems, have attracted attention from potential attackers [1], [2]. ICS controls physical processes by associating various control components. Unlike the IT system, ICS not only faces physical damage but also suffers from cyberattacks. Especially with the development of cyber-physical systems, ICSs are becoming increasingly connected to the open network. Although networked ICS can improve efficiency, it is not easy to ensure system security.

Received 13 March 2025; accepted 25 March 2025. Date of publication 10 April 2025; date of current version 15 May 2025. This work was supported in part by the Natural Science Foundation of China under Grant 62303126 and Grant 62362008 and in part by the Major Scientific and Technological Special Project of Guizhou Province under Grant [2024]014. The associate editor coordinating the review of this article and approving it for publication was Dr. Ke Feng. *(Corresponding author: Zhenyong Zhang.)*

The authors are with the State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China (e-mail: tzudsblue@gmail.com; zhangzy@gzu.edu.cn; auyuwhale@gmail.com).

Digital Object Identifier 10.1109/JSEN.2025.3555889

Several incidents
ICSs are vulnerable
attacker has the mo
the security defense
the underlying phys
infected Iran's nucl
sive damage to cer
company was attack
to widespread pow
In 2017, a petroleu
in the Middle Eas
compelling its shut
Steel Plant was tar
(APT) attack, where
plant's production
attack ultimately le
duction to halt in t
events demonstrate t
of developing appre
infrastructures.
The threat of cy
only in network i
machines but more

Scientific Reports
https://doi.org/10.1038/s41598-026-53863-z
Article in Press

A digital twin and deep-learning ensemble for cyber attack detection in industrial control systems at the IoT edge

d: 13 October 2025
d: 14 May 2026
d online: 22 May 2026
article as: Sayghe A.,
di M.D. & Gharawi A.A. A digital
d deep-learning ensemble for
tack detection in industrial
ystems at the IoT edge. *Sci*
26). https://doi.org/10.1038/
026-53863-z

Ali Sayghe, Moha

We are providing
findings. Before
note there may
apply.
If this paper is
Review reports v

ARTICLE

Causal Digital Twins for Cyber-Physical Security in Water Systems: A Framework for Robust Anomaly Detection

Mohammadhossein Homaei[a,*], Mehran Tarif[b], Pablo García Rodríguez[a], Andrés Caro[a] and Mar Ávila[a]

[a]Department of Computer Systems and Telematics Engineering, University of Extremadura, Av. Universidad s/n, Cáceres, 10003, Spain
[b]Department of Computer Science, University of Verona, Strada le Grazie 15, Verona, 37134, Italy

ARTICLE INFO

Keywords:
Causal Inference
Digital Twin
Industrial Control Systems
Cyber-Physical Security
Anomaly Detection
SWaT Testbed

ABSTRACT

Industrial Control Systems (ICS) in water distribution and treatment face cyber-physical attacks exploiting network and physical vulnerabilities. Current water system anomaly detection methods rely on correlations, yielding high false alarms and poor root cause analysis. We propose a Causal Digital Twin (CDT) framework for water infrastructures, combining causal inference with digital twin modeling. CDT supports association for pattern detection, intervention for system response, and counterfactual analysis for water attack prevention. Evaluated on water-related datasets SWaT, WADI, and HAI, CDT shows 90.8% compliance with physical constraints and structural Hamming distance 0.133 ± 0.02. F1-scores are 0.944 ± 0.014 (SWaT), 0.902 ± 0.021 (WADI), 0.923 ± 0.018 (HAI, $p < 0.0024$). CDT reduces false positives by 74%, achieves 78.4% root cause accuracy, and enables counterfactual defenses reducing attack success by 73.2%. Real-time performance at 3.2 ms latency ensures safe and interpretable operation for medium-scale water systems.

1. Introduction

The ICS have changed from isolated proprietary networks to interconnected cyber-physical systems that combine operational technology with information technology infrastructure. This change started in the 1990s with the use of standard networking protocols and has grown faster with Industry 4.0 initiatives. Today, ICS need Internet connectivity for remote monitoring, predictive maintenance, and distributed control across facilities that are far apart Mathur & Tippenhauer (2016); Xu et al. (2023); Abdoli & Ye (2023); Jiang et al. (2024). Critical infrastructure sectors such as water treatment, power generation, and manufacturing depend more on these connected systems, with the global ICS market expected to reach $124 billion by 2025. The Secure Water Treatment (SWaT) testbed, developed by iTrust at Singapore University of Technology and Design, is an important example for cyber-physical security research. It provides a realistic scaled-down industrial environment with 51 sensors and actuators across six operational stages Goh et al. (2017); Jeremiah et al. (2024); Lucchese et al. (2024).

However, more connectivity and complexity create new cybersecurity vulnerabilities that traditional IT security cannot fully solve. The 2010 Stuxnet attack on Iranian nuclear facilities showed that advanced attackers can exploit both cyber weaknesses and physical process knowledge to cause real damage Raman & Mathur (2022). Recent events, like the 2021 Colonial Pipeline ransomware attack and the 2021 Oldsmar water treatment plant breach, show the ongoing threat to critical infrastructure. Current anomaly detection methods in ICS mostly use statistical correlation and machine learning techniques, which cannot separate real causal relationships from false associations Kravchik & Shabtai (2018); Kim et al. (2023). These correlation-based methods often have high false positive rates, cannot do root cause analysis, and are vulnerable to adversarial attacks that take advantage of missing causal understanding.

The main limitation of current methods is their inability to reason about causal relationships in cyber-physical systems. Correlation can show statistical dependencies between sensor readings, but it cannot tell if anomalies come from attacks, cascading failures, or normal operational changes. This problem is serious when security analysts need actionable information for incident response, because correlation-based alerts do not provide enough guidance to

*Corresponding author.
mhomaein@alumnos.unex.es (M. Homaei); mehran.tarifhokmabadi@univr.it (M. Tarif); pablogr@unex.es (P.G. Rodríguez); andresc@unex.es (A. Caro); mmavila@unex.es (M. Ávila)
ORCID(s):

International Journal of Information Security (2025) 24:151
https://doi.org/10.1007/s10207-025-01043-x
REGULAR CONTRIBUTION

Leveraging digital twins for advanced threat modeling in cyber-physical systems cybersecurity

Gizem Erceylan[1] · Aida Akbarzadeh[1] · Vasileios Gkioulos[1]

Published online: 3 June 2025

Abstract
Threat modeling is a critical proactive security technique for identifying threats and determining mitigations. However, traditional approaches often fall short for Industrial Control Systems (ICS), which automate operations in domains like manufacturing and energy and are a subset of Cyber-Physical Systems (CPS). CPS integrates computation, networking, and physical processes, with ICS requiring specialized cybersecurity approaches due to its operational and safety-critical nature. This study explores the use of digital twin technology as a promising cybersecurity tool for ICS, enabling testing and analysis without disrupting operations. By examining the capabilities of digital twins in analysis, simulation, and replication, the research evaluates their potential to enhance threat modeling across the CPS life-cycle. Insights from the European Cyber Security Organisation (ECSO) Technical Paper on Cybersecurity Scenarios and Digital Twins guide the exploration of their role in threat modeling. The study addresses four research questions: 1. What purposes do digital twins serve in cybersecurity? 2. What benefits do digital twins offer in cybersecurity? 3. How can digital twin technology be leveraged for threat modeling? 4. What advantages can the use of digital twins bring to threat modeling? Our findings reveal that digital twins enhance ICS threat modeling by enabling continuous, dynamic, and autonomous assessment, offering valuable insights for advancing cybersecurity strategies in ICS, CPS, and related domains.

Keywords Digital twin · Threat modeling · Cyber physical systems · Industrial control systems · Cybersecurity

1 Introduction

With the advent of Industry 4.0, technologies such as cloud computing, digital twins, and the Internet of Things (IoT) have significantly expanded the attack surface for ICS. Initially designed to operate in isolation, ICS are now increasingly connected to networks (including the Internet) to enhance functionality. However, this connectivity introduces new cybersecurity challenges, particularly in the manufacturing sector, as attack methods grow more complex and evolve rapidly. Detecting relationships between interconnected devices and tracing the progression of attacks over time has become increasingly difficult. Unlike traditional Information Technology (IT) networks, cyberattacks targeting ICS often require executing multiple coordinated actions across devices in a specific sequence. Addressing these unique challenges has led to recommendations for using digital twins in cybersecurity.

The motivation for this study stems from the growing need for proactive and adaptive security solutions in ICS environments. As ICS plays a critical role in sectors like energy, transportation, and manufacturing, their compromise can have severe economic, environmental, and safety implications. Traditional threat modeling approaches often struggle to capture the dynamic and interconnected nature of ICS, leaving gaps in identifying vulnerabilities and predicting attack pathways. One significant reason for this challenge is the difficulty in capturing the operational context of such systems and reasoning based on their operational semantics. ICS environments involve complex interactions between physical processes, networked components, and control systems, making it challenging to fully represent these intricacies in

✉ Gizem Erceylan
gizem.erceylan@ntnu.no
Aida Akbarzadeh
aida.akbarzadeh@ntnu.no
Vasileios Gkioulos
vasileios.gkioulos@ntnu.no
[1] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Gjøvik, Norway

Springer

SAC: Pros & Cons

Advantages:

- Better exploration
- More stable learning
- Robust policies
- Balanced behavior

Disadvantages:

- More complex
- Higher computation cost
- Sensitive hyperparameters
- Harder to interpret

Big Data Infrastructure for further processing