Privacy preserving and verifiable outsourcing of AI processing for cyber-physical systems

Georgios Spathoulas[1] , Angeliki Katsika[2], and **Georgios Kavallieratos[1]**

[1] Dept. of Information Security and Communication Technology, NTNU, Norway
[2] Dept. of Computer Science and Biomedical Informatics, University of Thessaly, Greece

26th International Conference on Information and Communications Security (ICICS 2024), August 26-28, 2024, at the University of the Aegean, in Mytilene, Greece.

Funded by the European Union

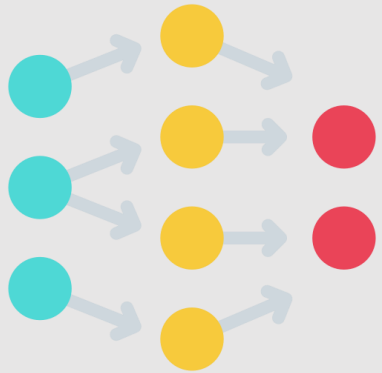
This work was supported by the European Commission [grant 101120657 "European Lighthouse to Manifest Trustworthy and Green AI" - ENFIELD]

Outline

- Introduction
- AI and CPSs
- Motivation
- Verifiable machine learning
- VML to CPS
- Results
- Discussion

AI in CPSs

- **Learn**, **adapt**, and **make informed decisions** in real-time.
- Analysis of **real-time data** from sensors.
- Predict the **states** of the CPSs.
- Enables proactive **maintenance**, reducing downtime, and **optimizing production efficiency**.
- Analysis of **vast amounts of data** generated by sensors and smart meters.

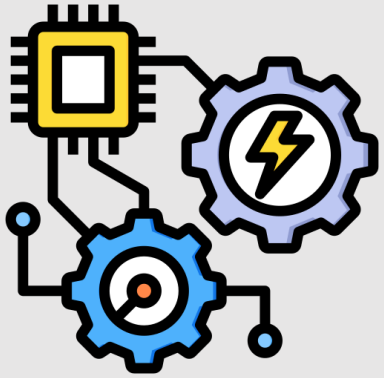


Algorithms

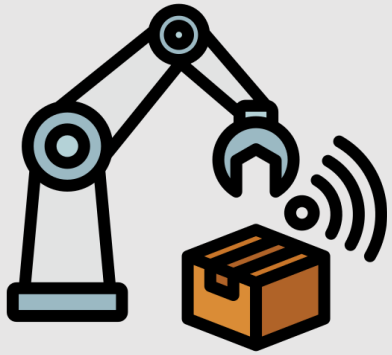
CPSs

Motivation

Control



Optimized and **flexible** control



Complex situations in the physical world

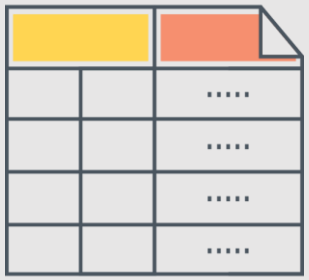
Analyze data

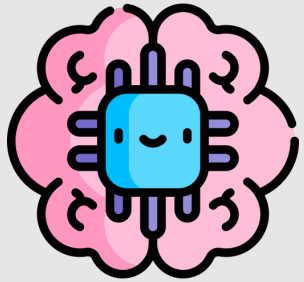
It **is not feasible for all CPSs** operators to collect a qualitative dataset and train an efficient AI model on their side.

Research Contributions

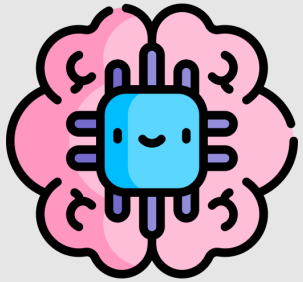
- Identification of how zero knowledge proofs can be employed for providing verifiable AI services to CPS.
- Validation of the hypothesis that such an approach would actually preserve data and model privacy and integrity of the inference operations.
- Assessment of computational complexity overhead of such approaches and discussion on their applicability in real-world CPS.

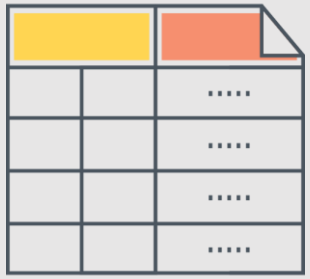
Traditional AI workflow





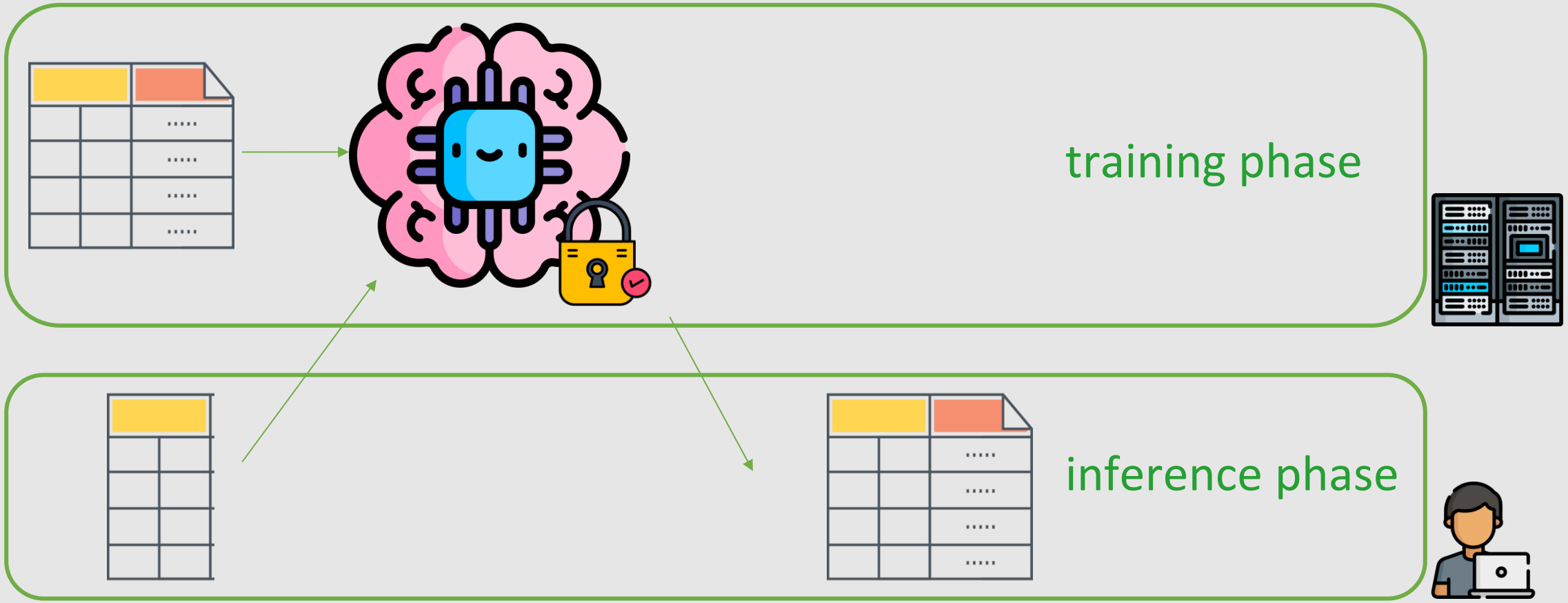
training phase





inference phase

Modern AI workflow



Verifiable machine learning..

AI service

AI model

data input

outcome

user

provider

user requires to be able to verify that outcome has been produced by feeding data input to AI model

Zero Knowledge

- A zero-knowledge (ZK) proof is a cryptographic protocol.
- One party, the prover, can prove to another party, the verifier, that a given statement is true, without revealing any additional information beyond the fact that the statement is true.
- ZK is used to create proofs of computational integrity for a set of given computations where:
  - the proof is significantly easier to verify than it is to perform the computation itself (succinctness)
  - hiding parts of said computation whilst preserving computational correctness is feasible(zero-knowledge)

..Verifiable machine learning

AI service

commitment

AI model

data input

outcome proof

user

provider

data input

outcome proof

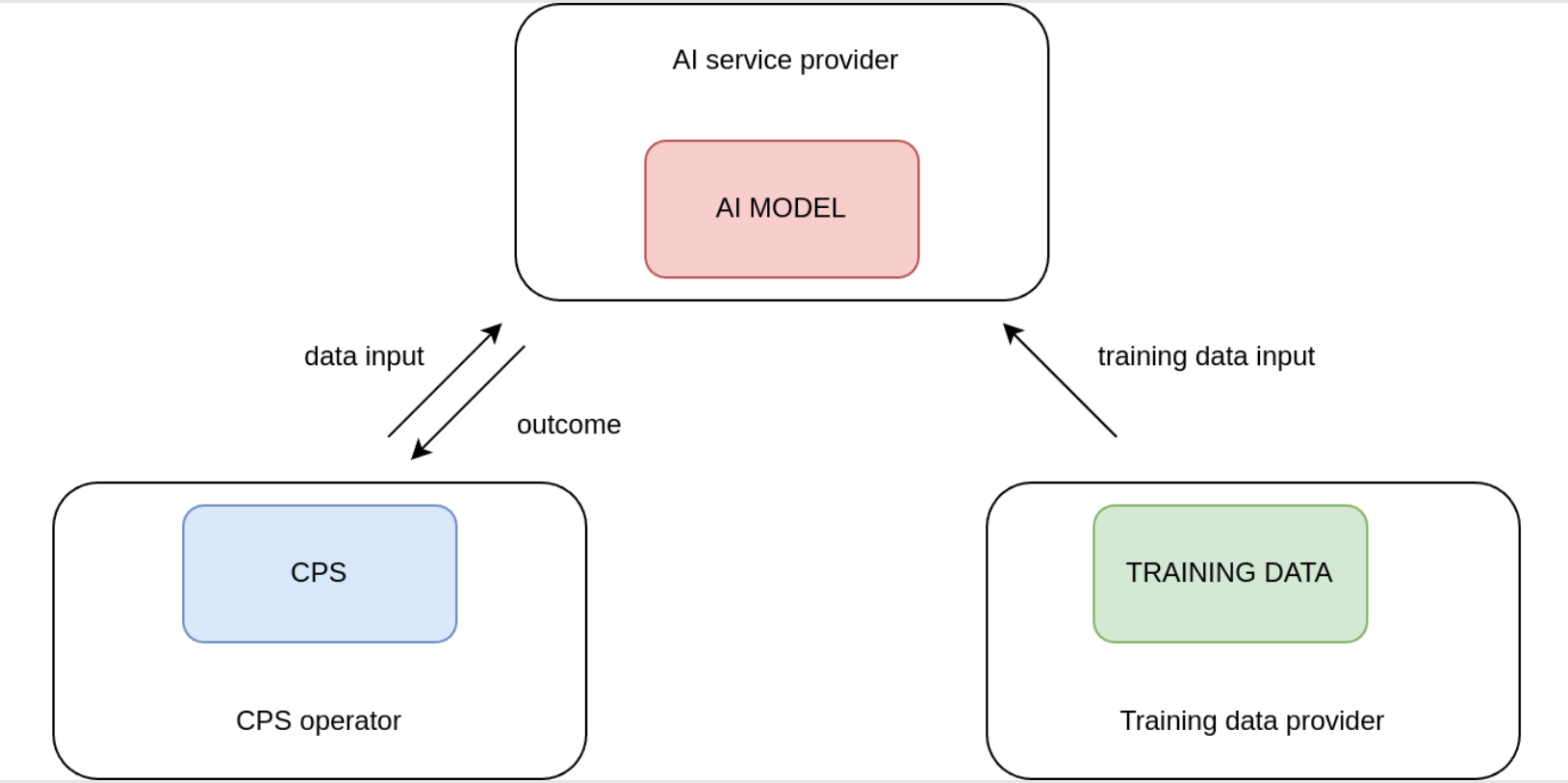
commitment

verification

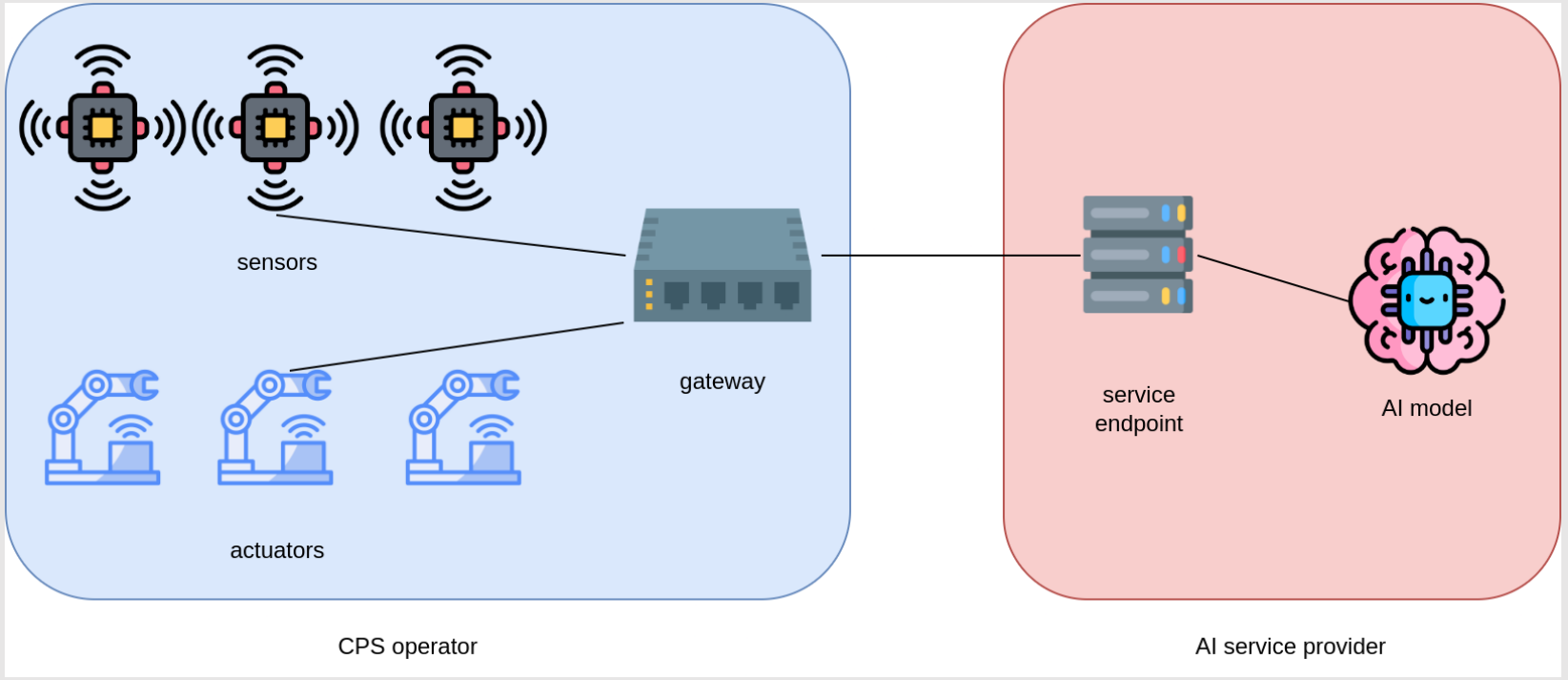
true or false

Applying VML to CPS

VML to CPS: Reference Setup



VML to CPS: Components



VML to CPS : Operation – Commitment Generation

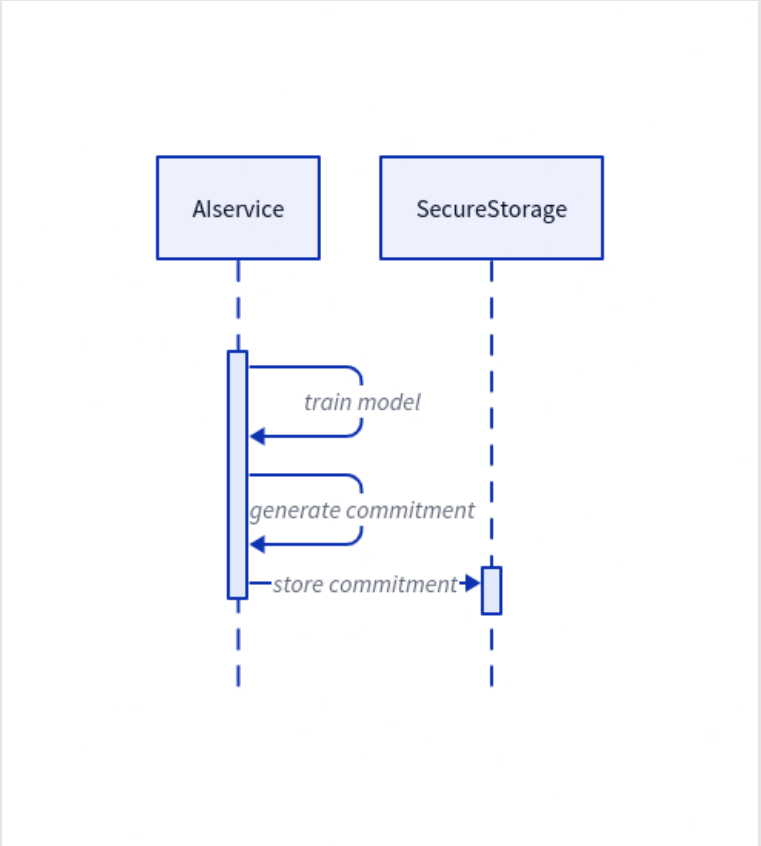
- An initial model m_0 is trained and transformed to a new instance m_1

  m1 = training(m0, data)

- The commitment generation process takes as input the model and produces a cryptographic commitment

  c_1 = commit(m_1)

- Commitment generation must be repeated in any subsequent update of the model

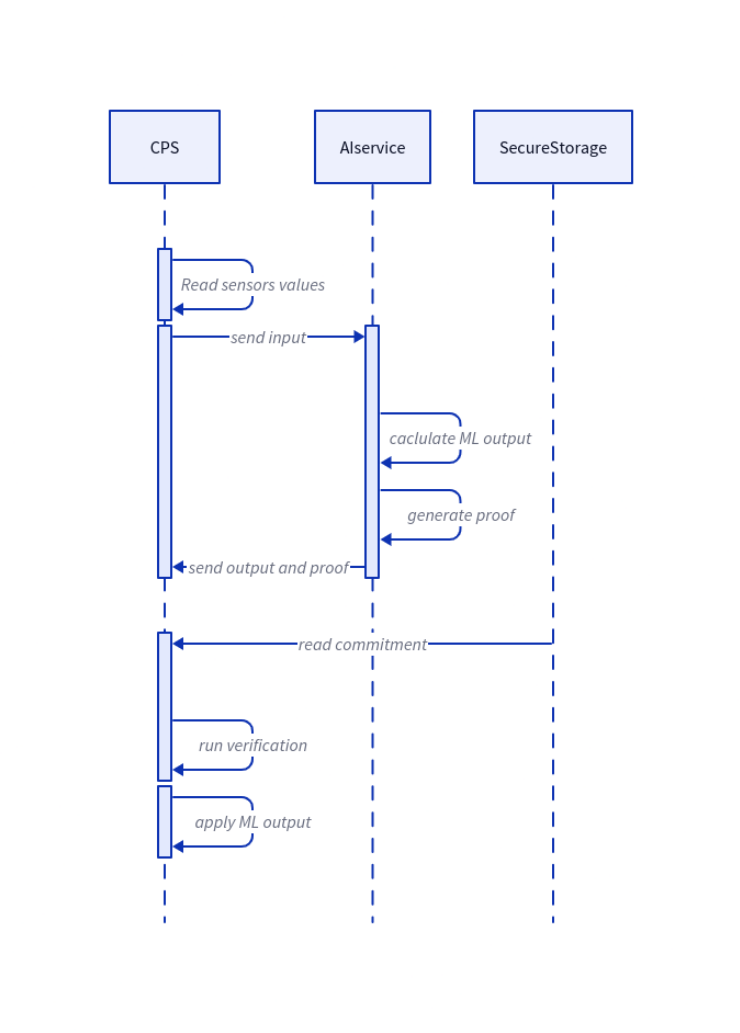


VML to CPS: Operation - Inference

- The AI model maintained by the AI service provider can decide the **optimal set of actions** for the CPS's actuators at a specific time point t: **$actions_t$**
- The input is the r most recent values monitored by all n CPS sensors and is denoted as **$values_{t,r}$**
- The CPS system sends to the AI service provider **$values_{t,r}$**
- The AI service provider feeds the data to the AI model m and calculates the best actions for the CPS actuators **$actions_t$**
- The AI service provider generates a proof related to the previous-step operations

$$proof_t = proof\_gen(m, values_{t,r}, actions_t)$$

- The AI service operator returns to the CPS the **$actions_t$** along with **$proof_t$**

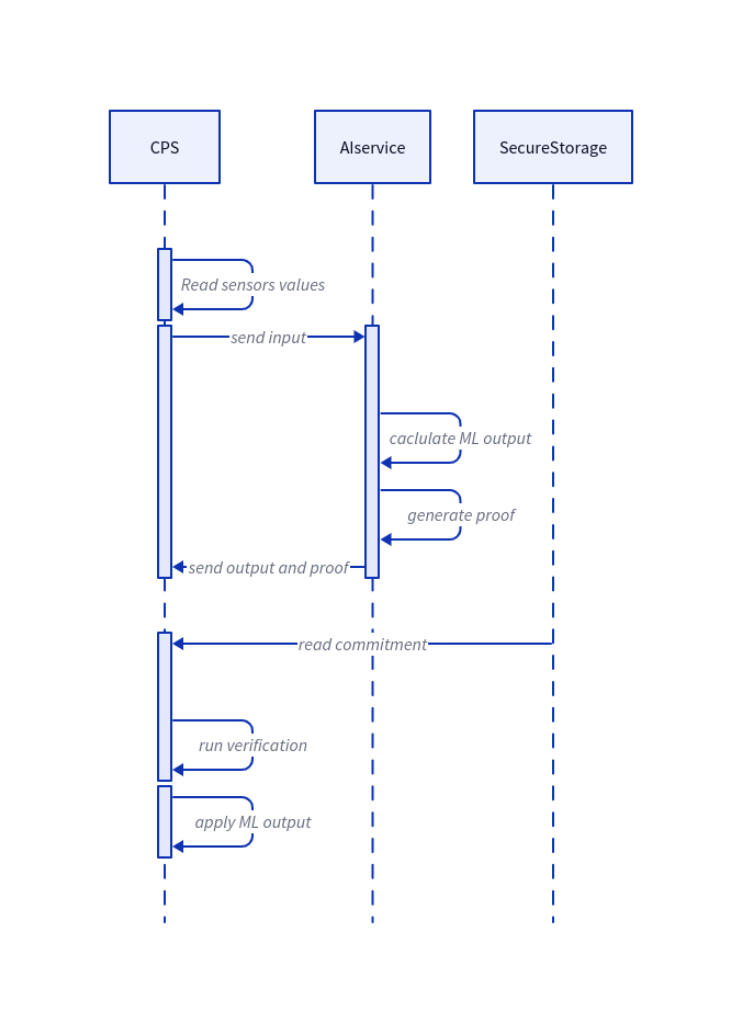


VML to CPS: Operation - Inference

- The CPS operator verifies the validity of the operations of the AI service provider.
- It uses as input
  - received data $actions_t$
  - $proof_t$
  - sent data $values_{t,r}$
  - the commitment c

result = proof_val(c, $proof_t$, $values_{t,r}$, $actions_t$)

- If the result of the verification in the previous step is positive, then the CPS can apply the received $actions_t$ to the actuators.



Results

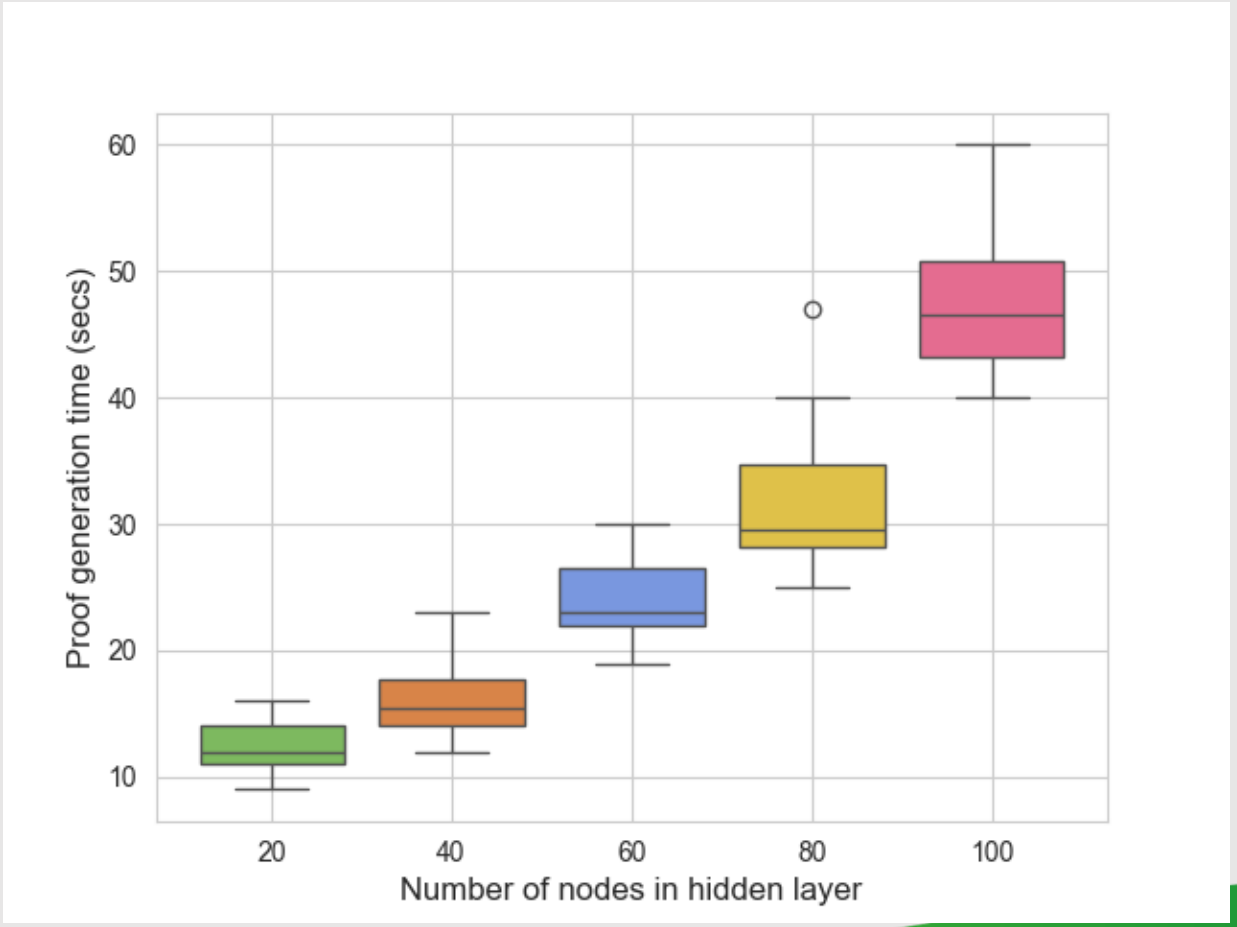
Experiments setup..

- HAI (HIL-based Augmented ICS) Security Dataset.
- ICS testbed augmented with a Hardware-In-the-Loop (HIL) simulator that emulates steam-turbine power generation and pumped-storage hydropower generation. Three types of cyberattacks have been simulated.
- The AI model aims to process CPS monitoring data and identify potential cyberattack.

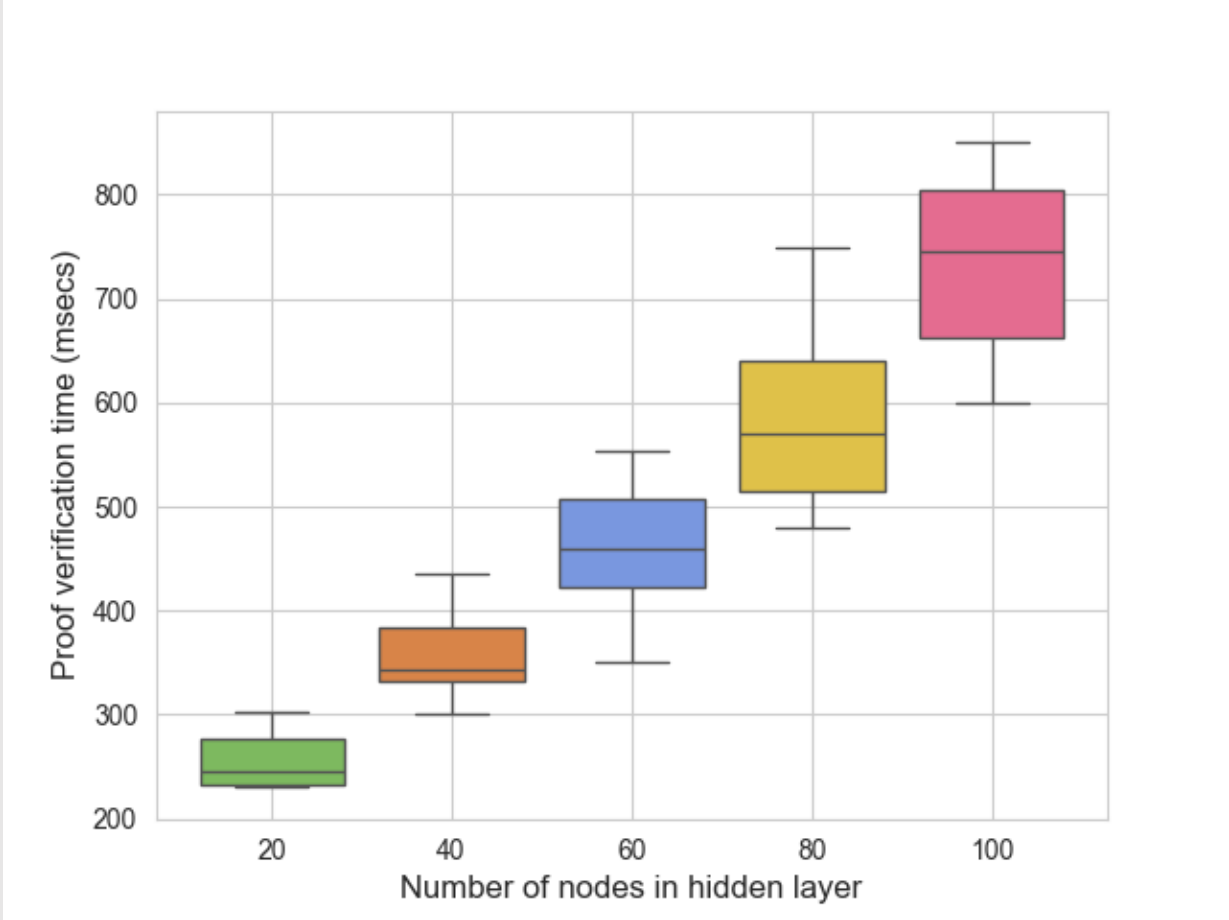
..Experiments setup

- We trained a neural network of a specific size, with the provided data, and tested the validity proof generation and validation process along with the time required for proof generation.
- Each neural network has three layers:
  1. the input layers consisting of 79 sensors
  2. a hidden layer (20,40,80 or 100 nodes)
  3. the output layer which has only 1 node

Proof Generation Times



Proof Verification Times



Discussion

- **Conclusions**
  - We validated that it is feasible to apply the proposed methodology, and assessed the processing overhead that the proposed methodology brings in relation to the size of the AI model used.
  - Allows for new and update-able AI systems to be applied to a deployed CPS.
  - Protects privacy of AI models, and training data.
  - No additional hardware requirements for verification.
- The main **limitation** relates to the high proof generation time, which depends on the size of the AI model
- As plans for **future work**, we aim at exploring the use of GPU hardware

Dr. Georgios Kavallieratos
georgios.kavallieratos@ntnu.no

Funded by the European Union

This work was supported by the European Commission [grant 101120657 "European Lighthouse to Manifest Trustworthy and Green AI" - ENFIELD]