RS1: Securing the Autonomous System Usage Environment

Lancaster University

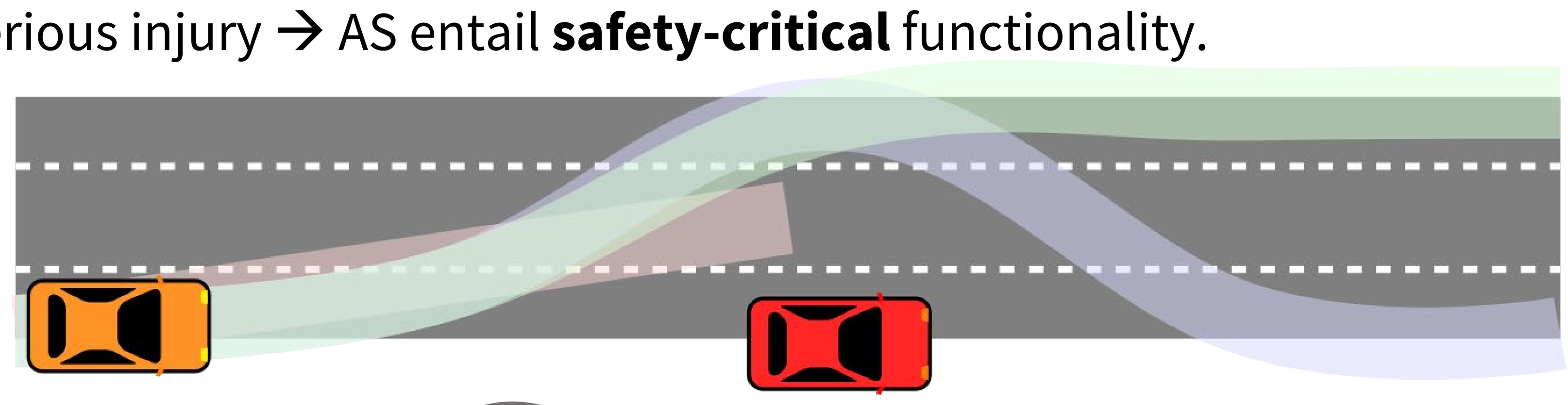
Researchers: Dr. Andrew Sogokon, Dr. Zhengxin Yu, Dr. Yi Li
Investigators: Prof. Neeraj Suri, Prof. Plamen Angelov

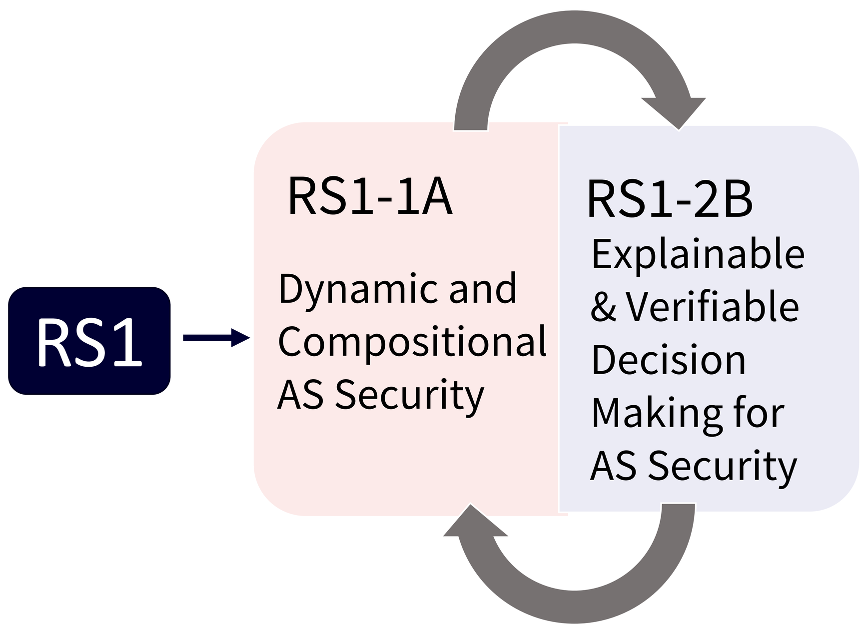
Engineering and Physical Sciences Research Council

TAS-S

RS1 - Secure Usage of Autonomous Systems

Autonomous Systems (AS) are typically *Cyber-Physical Systems* (CPS) where malfunctions can lead to catastrophic consequences, such as loss of life or serious injury → AS entail **safety-critical** functionality.



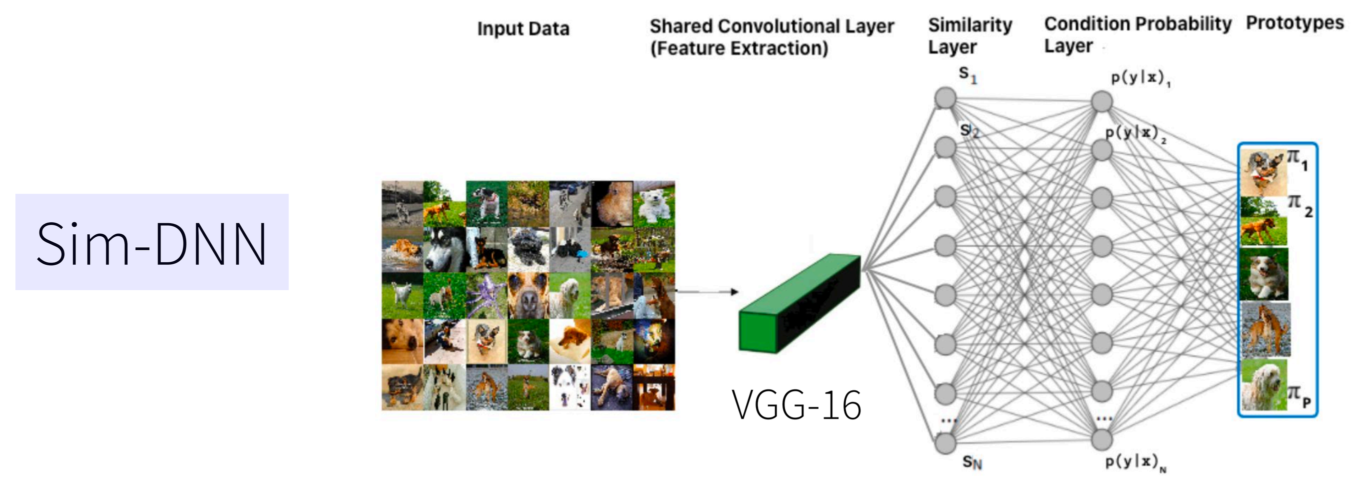


Challenges in RS1

- AS operates in dynamic, unpredictable environments.
- AS may be exposed to attacks; attacks are often complex – may be discrete, collusion and multi-layered.
- AS often process large amounts of data with complex data structures.
- AS needs to make "adaptive & run-time" decisions with *incomplete and uncertain* data streams & resources.
- AS nodes are *mobile*.

RS-1B (2): Detecting Imperceptible Attacks

- Similarity-based Deep Neural Networks (**Sim-DNN**) can be used to detect *imperceptible adversarial attacks* on the sensors (e.g. vision system) of AS.



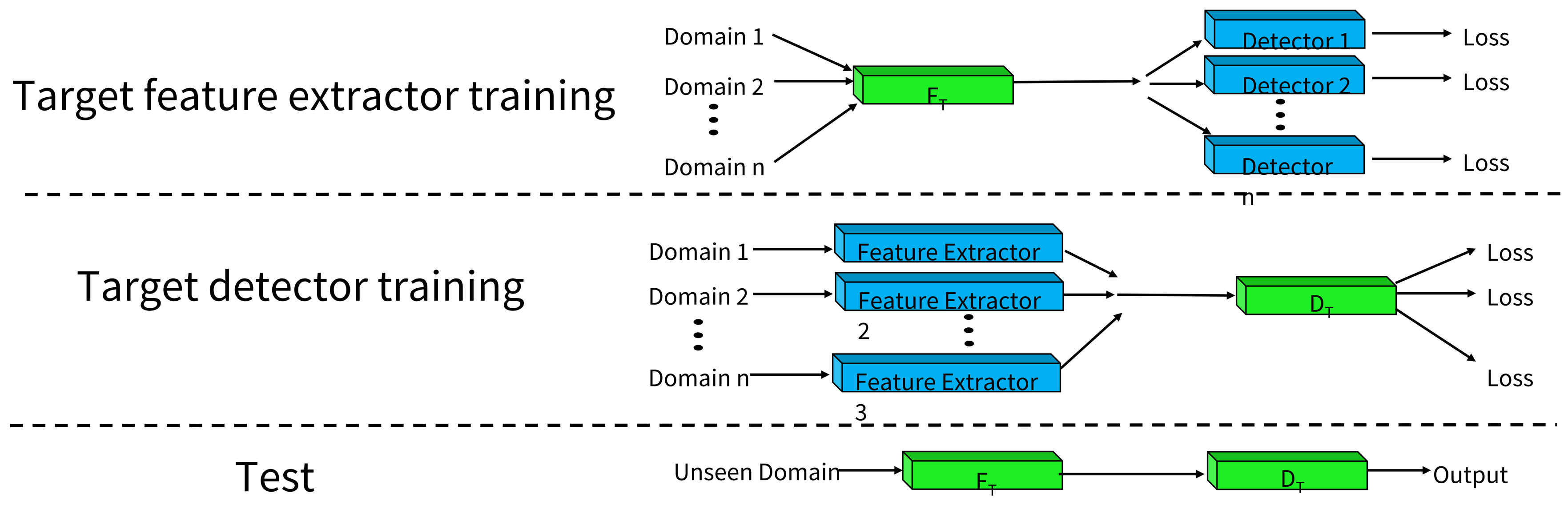
Pros:

- These frameworks provide excellent results for various attacks.
- These methods require few manual-engineering.

Cons:

- Weak adaptability and transferability to new domains, e.g., attacks or datasets.
- Slow training due to large model scales, particularly for the feature extractor (VGG-16).

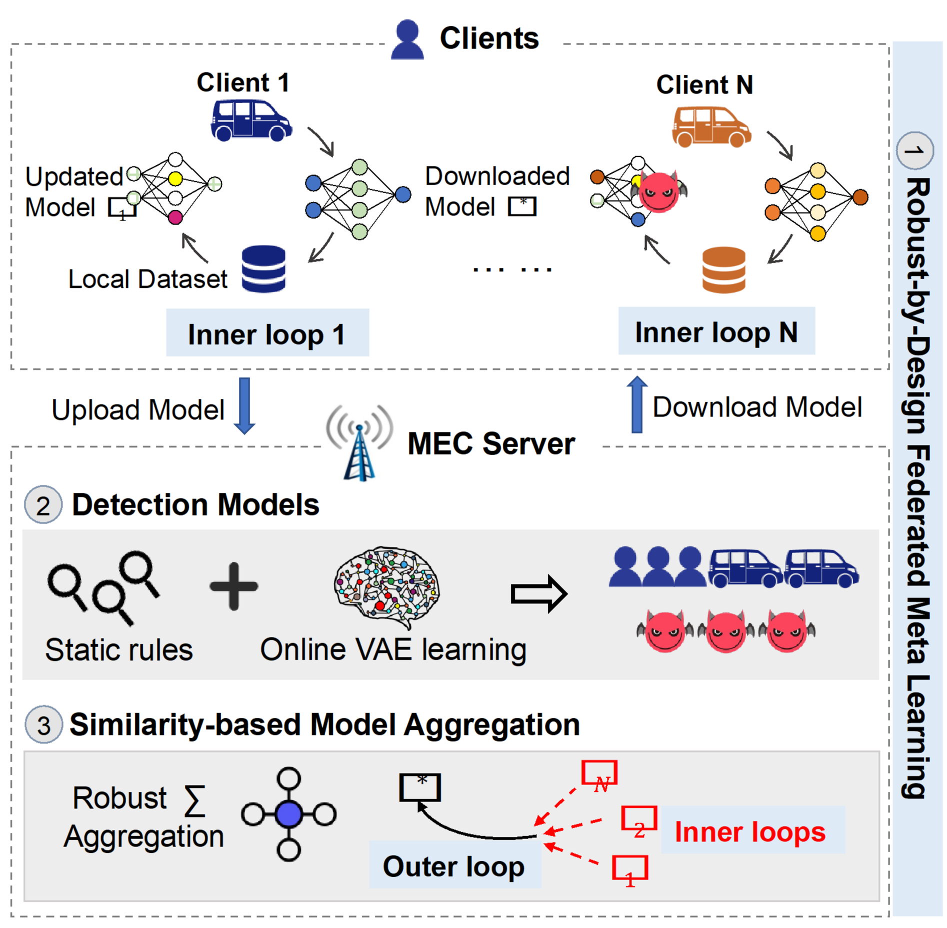
RS-1B (2): ML Domain Generalization Framework



- The feature extractor or detector is trained with a partner who is well tuned for different domains.
- In the test stage, the trained target feature extractor and detector are combined with the FFN to detect attacks in unseen domains.

RS-1A: RAFL- Dynamic & Compositional AS Security

- Develop a **robust and adaptive** federated meta-learning framework (**RAFL**) resilient against adversaries.



Goals:

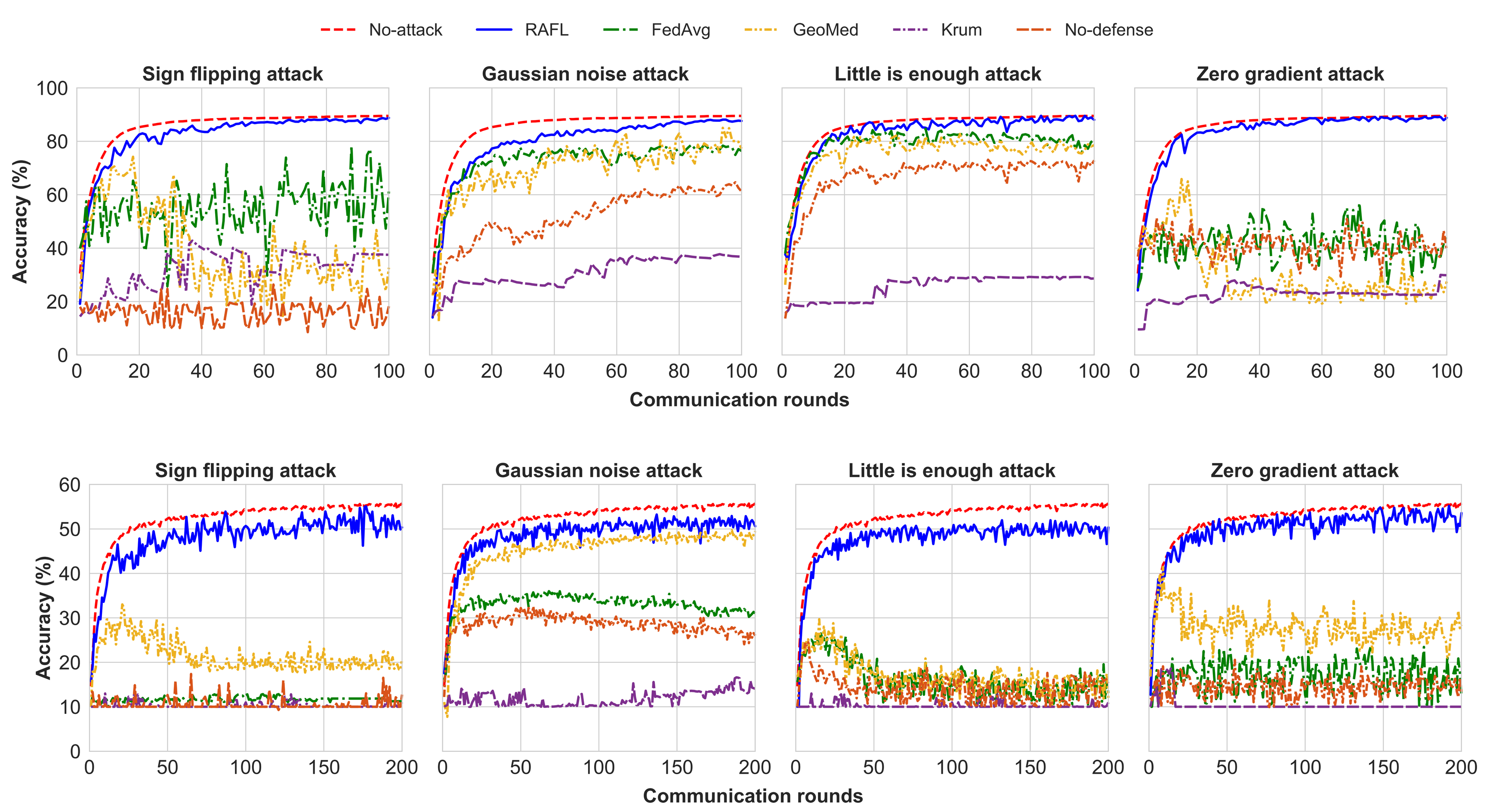
- Leverage distributed AS nodes to collaboratively train a global model to quickly adapt to new environments.
- Defend against adversarial attacks to reduce negative impact of attacks on ML models.

Key techniques:

- Federated meta-learning: Decentralized inner/out loops to train ML models.
- Rule-based and Variational Autoencoder (VAE) online learning-based detection model to detect adversarial attacks.
- A similarity-based model aggregation to conduct a global meta-model to further reduce the likelihood of uploading adversarial models from AS nodes.

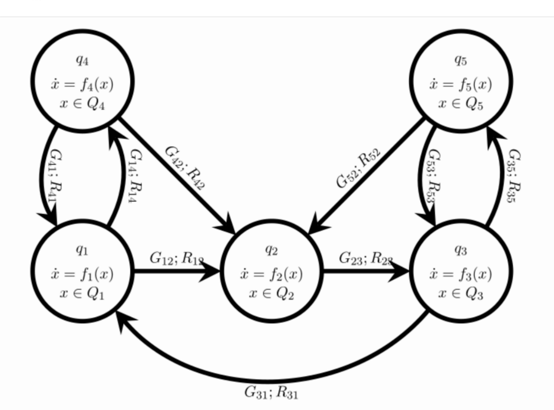
➲ RS-1A: RAFL- Experimental Results

RS-1A: The experimental results demonstrate that the proposed **RAFL** framework is robust by design and outperforms other baseline defensive methods against adversaries in terms of model accuracy and efficiency.



RS-1B (1): Safe Decision Making in AS

- Establishing **safe and secure** operation of an AS in uncertain and dynamic environments is part of the focus of our research in **RS-1B** (*Explainable and Verifiable Decision Making*). We have undertaken a survey of specifications of AS, focusing on *formal specification*.
- Formal modelling and verification of CPS is highly challenging, but can help in providing very strong guarantees about the behavior of AS.



- We are working towards adding support for reasoning about CPS in the formal verification framework of **TLA+** based on Lamport's Temporal Logic of Actions.
- Formal methods can provide *verifiable solutions* to trustworthy decision making in AS.

➲ RS-1B (1): Safe Decision Making in AS

RS-1B(1): We have implemented a *proof obligation generator* for checking continuous inductive invariants (the proof obligations are discharged using the SMT solver **Z3**) and are currently engaged in integrating it with the **TLA+ Toolbox**. Enables a convenient way of proving safety of continuous systems within the formal framework of the TLA+ Toolbox and will support formal verification of CPS.

RS-1A & 1B: Ongoing Work

- RS-1A: Develop a mobility-aware adaptive machine learning framework
- RS-1B (1): Formal specification of AS Safety and Security
- RS-1B (1): Case studies of safety verification of CPS in the TLA+ Toolbox. Integrate proof obligation generator into the Proof Manager in TLA+ Toolbox.
- RS-1B(2): Visualization results of the proposed algorithm will be completed.
- RS-1B(2): Adaptability and transferability will be evaluated in real-world photos.

Lancaster University · Cranfield University · UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]