OST

WO WISSEN WIRKT.

Multi-Agent Reinforcement Learning (MARL) in Cyber Security

Enhancing Cyber Attack Autonomy Through Self-Play

19.06.2025

Christoph Landolt

CYD Master Thesis Fellow now working at CISPA Helmholtz Center for Information Security

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus

Agenda

1. **Autonomous Intelligent Cyber Agents**
   - The Problem of Machine Learning in Intrusion Detection
   - Autonomous Intelligent Cyber Agent Reference Architecture
2. **Reinforcement Learning (RL) driven Attacker**
   - Introduction to RL
   - Single-agent RL for penetration tests
   - Experiments and results
3. **Active RL-Defender**
   - Multi-Agent Reinforcement Learning (MARL) in Cyber Security
   - Attacker-Defender Dynamics
   - MARL control loop and training setup
   - Observation and open challenges
4. **Q&A Session**
   - Open floor for questions, discussion, and feedback

OST

WO WISSEN WIRKT.

Introduction to Autonomous Intelligent Cyber Agents (AICA)

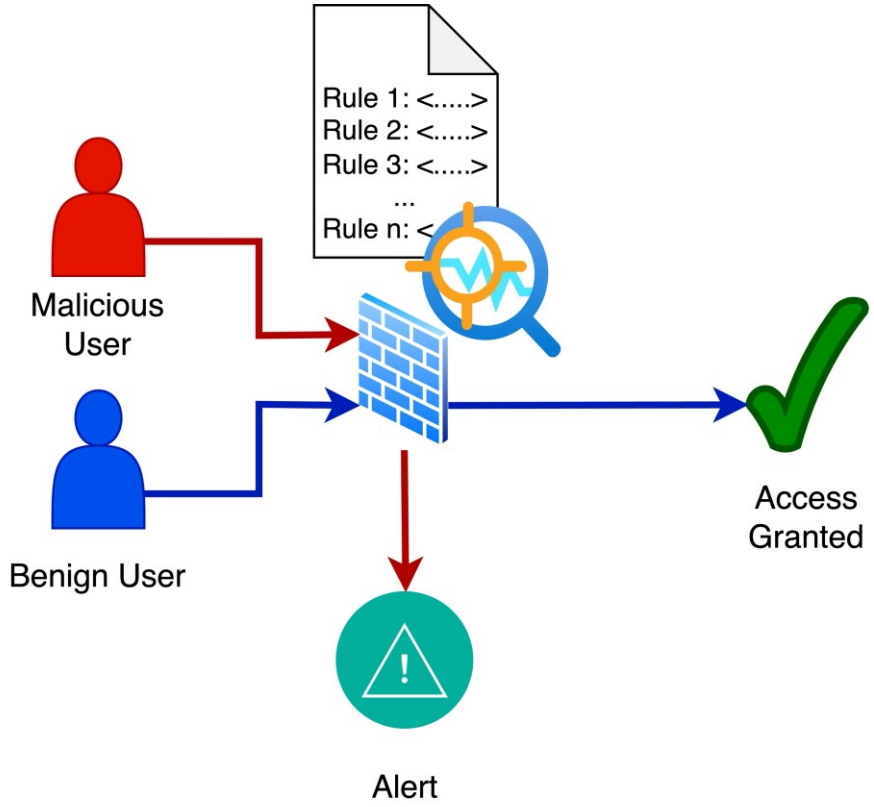
The Path to Agentic Cyber Defence

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

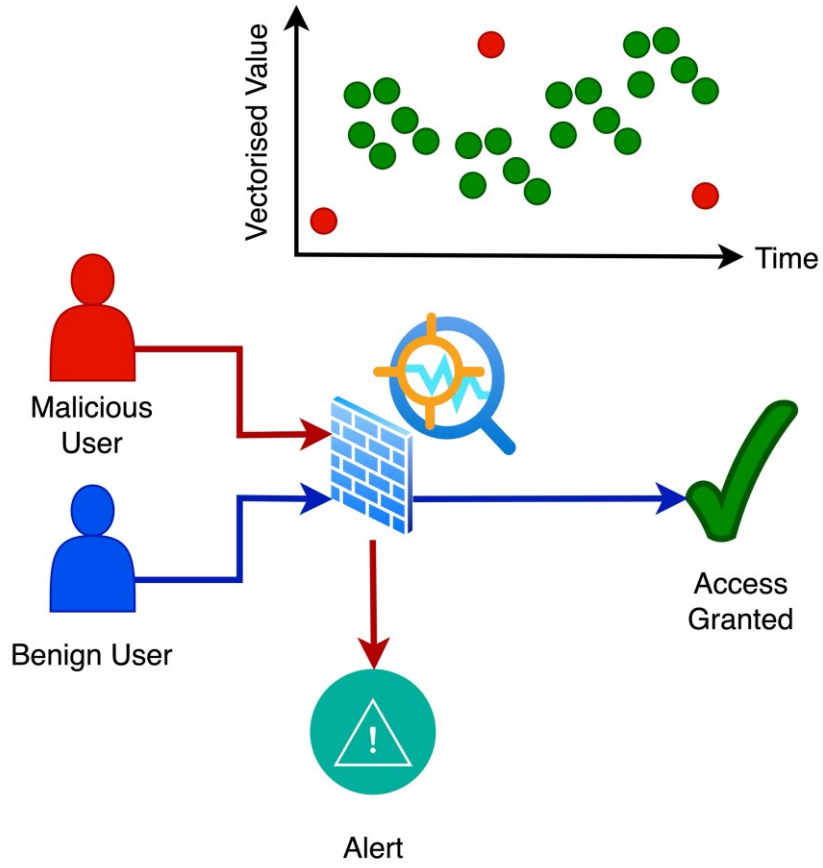
armasuisse Science and Technology
Cyber-Defence Campus

Intrusion Detection and Response

How to overcome static defence?



Rule Based Network Security Appliance

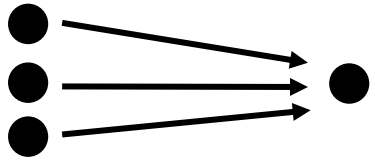


Machine Learning Based Network Security Appliance

Problem & Proposed Solution

Can Machine Learning Automate Coordinated Attacks?

Limited by Pre-existing Data
Traditional ML relies on ***existing datasets***, restricting its ability to discover novel strategies.



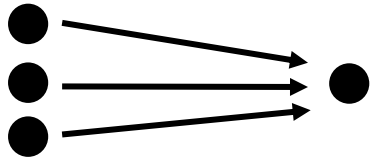
Distributed Attacks
Attacks are performed by ***multiple attackers***, complicating detection.

Non-Stationary Environments
Constant ***evolution of networks and attack strategies*** prevents stable training conditions.

Problem & Proposed Solution

Can Machine Learning Automate Coordinated Attacks?

Limited by Pre-existing Data
Traditional ML relies on ***existing datasets***, restricting its ability to discover novel strategies.

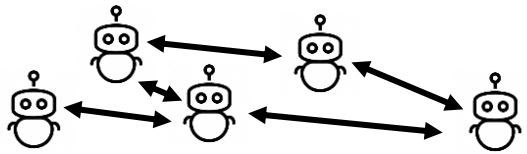


Distributed Attacks
Attacks are performed by ***multiple attackers***, complicating detection.

Non-Stationary Environments
Constant ***evolution of networks and attack strategies*** prevents stable training conditions.

Solution

Reinforcement Learning (RL)
RL explores actions through ***trial and error***, enabling it to find innovative and optimal strategies.

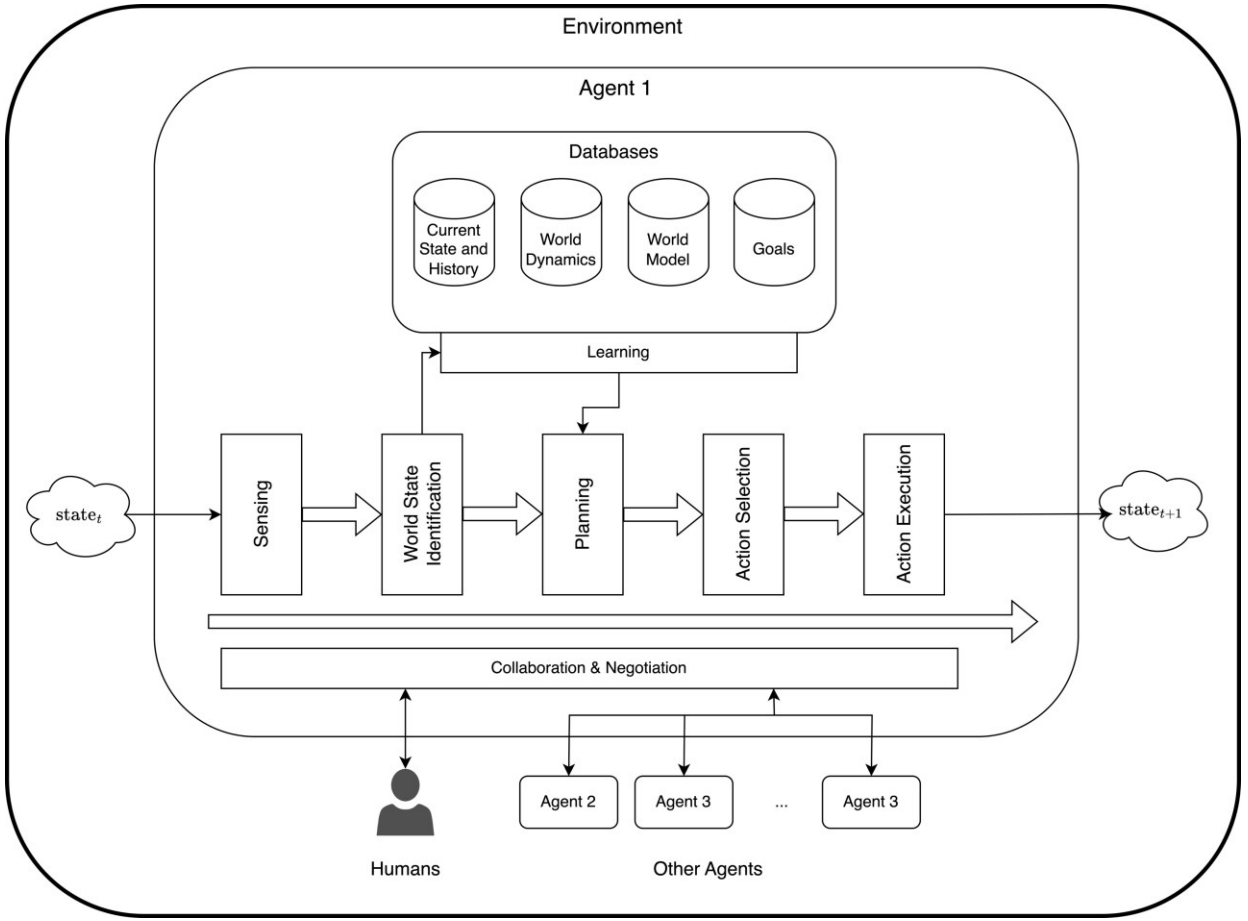


Coordinating Attack-Agents
Multi-agent systems can ***mimic distributed attacks*** for more realistic detection challenges.

Multi-Agent RL
Allows **adaption *in real-time***, addressing non-stationary and evolving attack strategies.

Autonomous Intelligent Cyber Agents

How to build an automated cyber defence?



NATO Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture

Autonomous Cyber Defense:

- **Sensing & World State**: Detect, gather/process data
- **Planning & Action**: Prioritize and select responses
- **Action Execution**: Implement and adapt actions
- **Collaboration**: Coordinate with agents or Humans
- **Learning**: Improve strategies via feedback

Architectures:

- **Centralized**: Master-agent control (e.g., SARL)
- **Distributed**: Self-organizing agents (e.g., MARL), more resilient but complex

OST

WO WISSEN WIRKT.

Reinforcement Learning (RL) driven Attacker

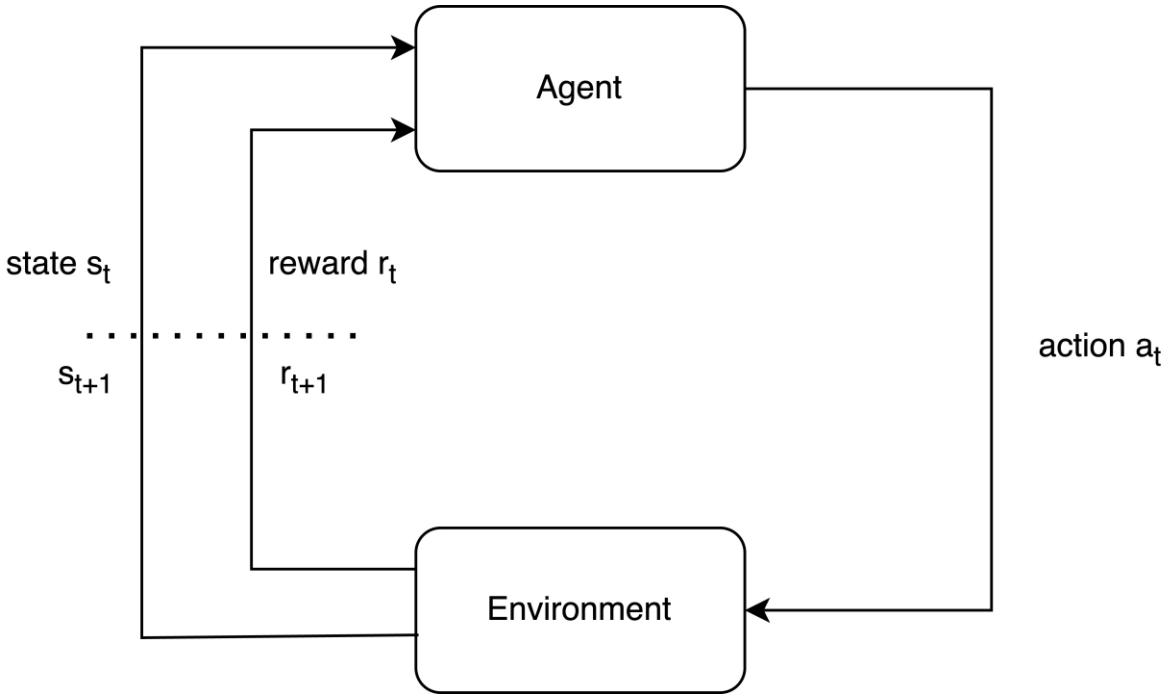
Penetration Testing as a Sequential Decision making Problem

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus

Introduction to Reinforcement Learning (RL)

How to learn through self play?



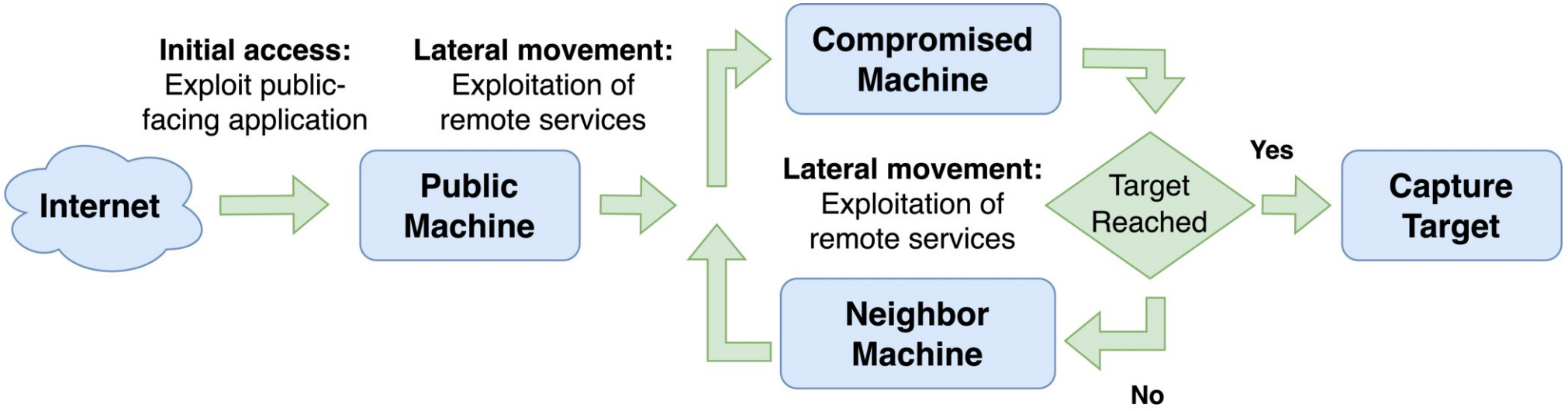
RL control loop

Use **Reinforcement Learning (RL)** when:

- **Sequential** decision-making is required
- **No labeled data**, but a reward signal is available
- Environment **dynamics are uncertain** or complex

Problem

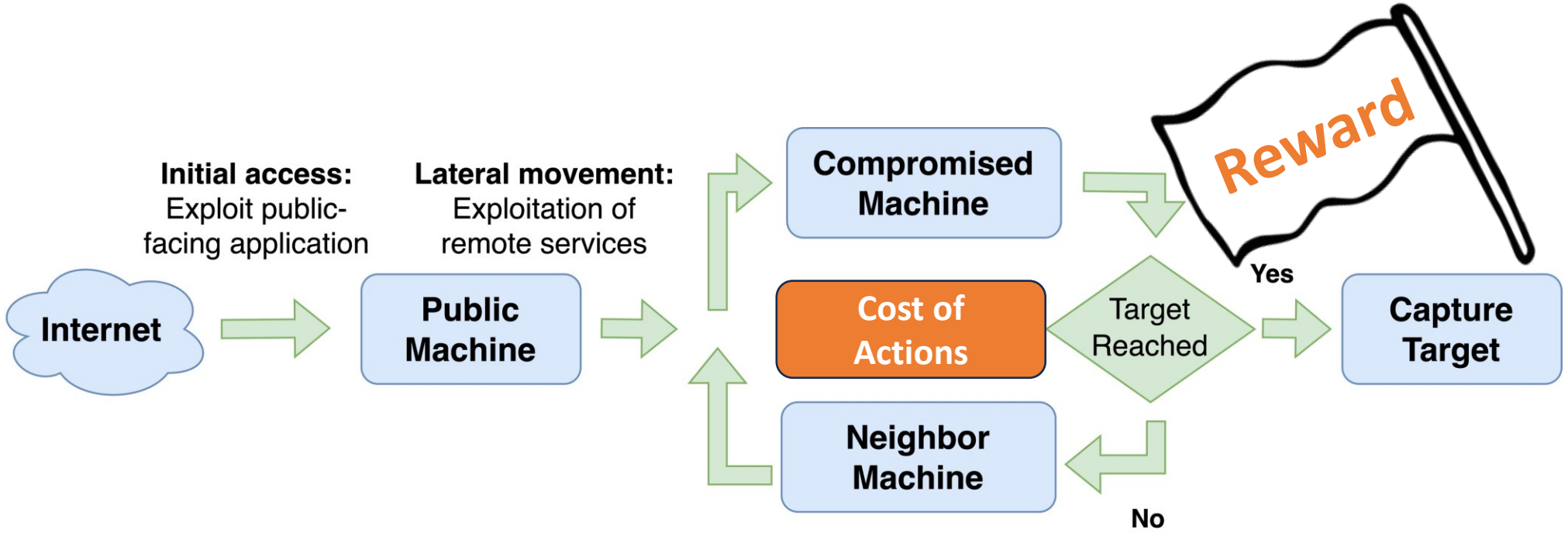
Can RL be used to automate Advanced Targeted Attacks (ATA)?



Attack graph starting from the Internet.

Problem

Can RL be used to automate Advanced Targeted Attacks (ATA)?

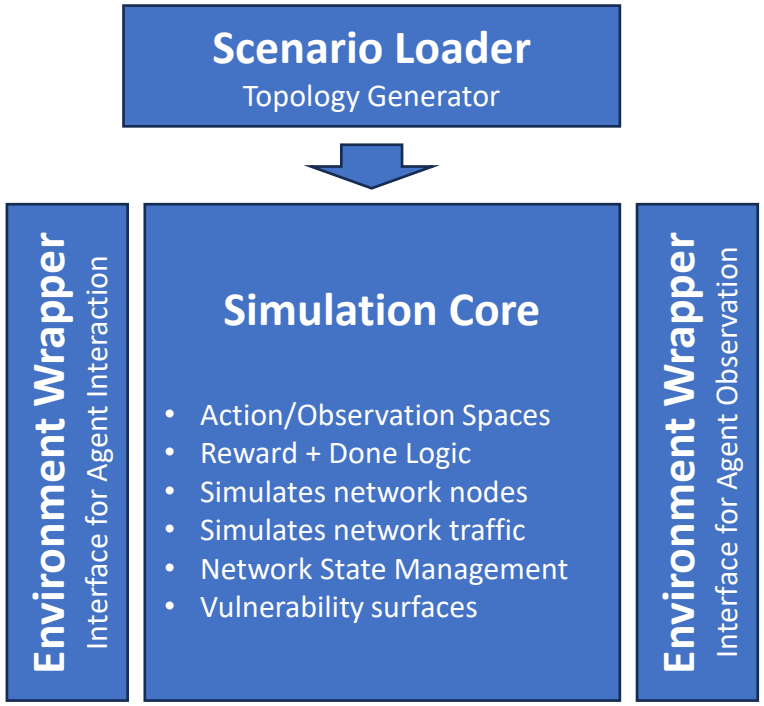


Attack graph starting from the Internet.

Sequential Actions / No training Data

RL Environment Architecture

How to simulate a computer network for RL?



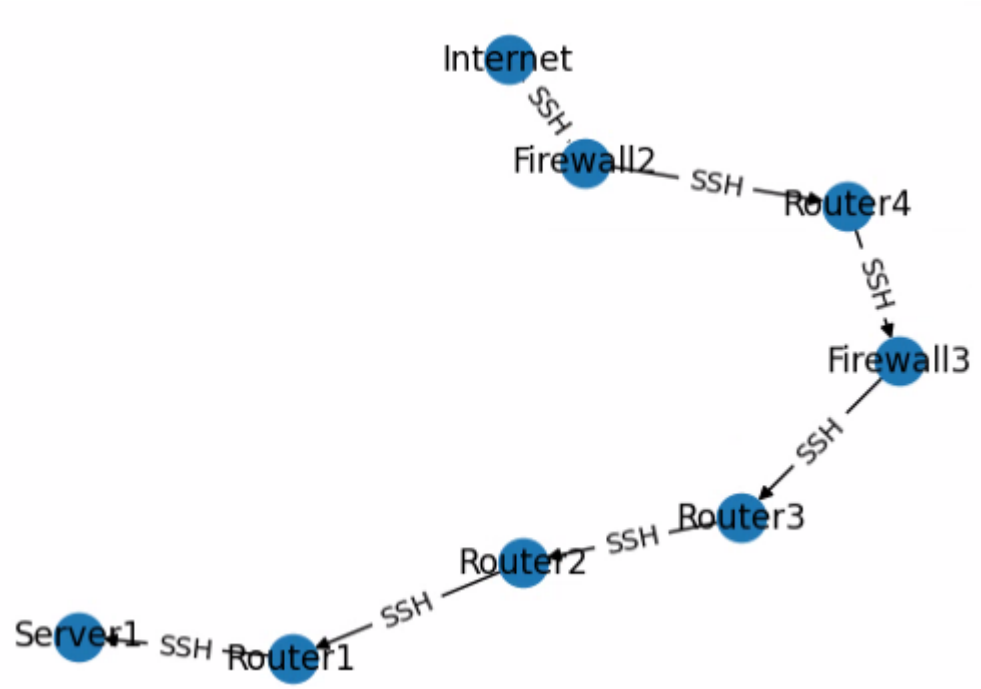
CyberGym Architecture

Environment	Red or Blue	Red and Blue	n blue or n Red
CyGIL (Li, Fayad, and Taylor 2021)	✓		
PrimAITE (Dstl 2023)	✓		
CSLE (Hammar and Stadler 2022)	✓		
Gym-IDS game (Hammar and Stadler 2020)	✓	✓	
CyberBattle Sim (Microsoft 2021)	✓		
MARLon (Kunz et al. 2022)	✓	✓	
Gym-Threat-defence (Miehling et al. 2015)	✓		
Gym-Optimal-Intrusion-Response (Hammar and Stadler 2021)	✓		
AtMOS (Akbari et al. 2020)	✓		
Yawning Titan (Collyer, Andrew, and Hodges 2022)	✓		
Farland (Molina-Markham et al. 2021)	✓		
CYST (Drašar et al. 2020)	✓		
CybORG (Standen et al. 2021)	✓		✓

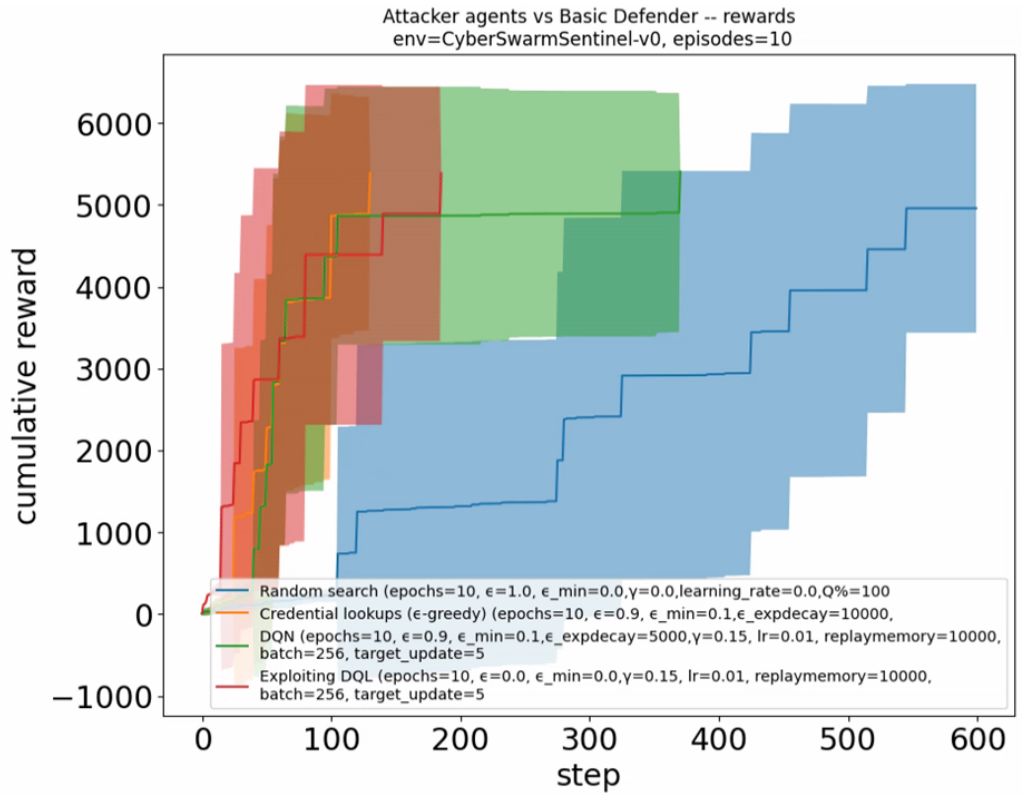
Source: Kiely et al. 2024, AAAI-25

Single Agent Training in CyberBattleSim Gym

Can an attacker move laterally?



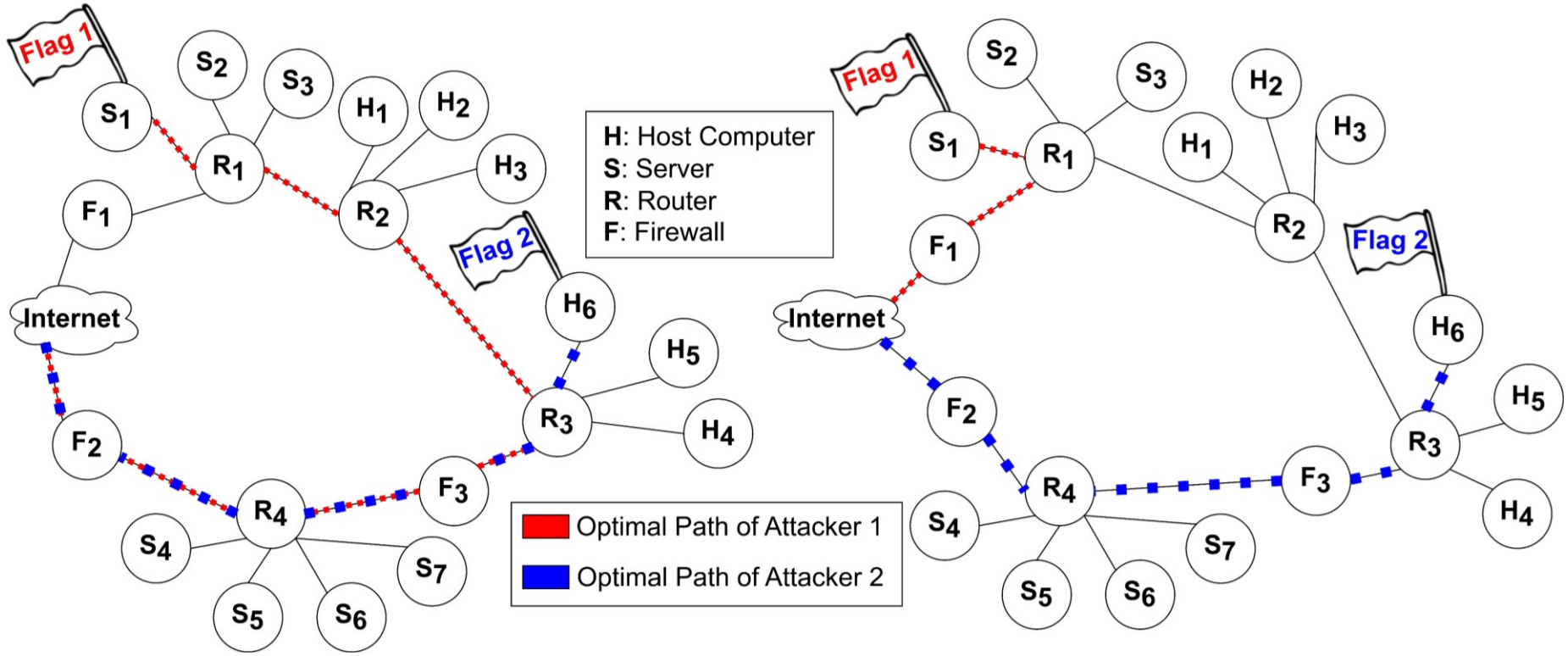
A chain of abstract network components, all of which have been intentionally made vulnerable.



Comparison of different RL algorithms

Single Agent Training in NASimEmu Gym

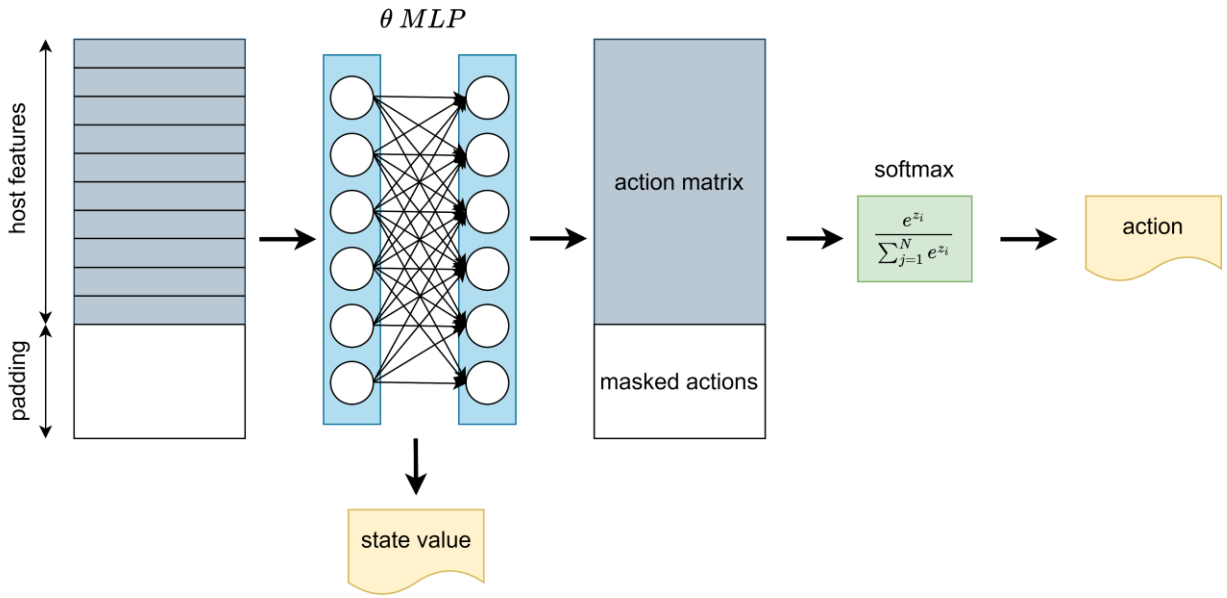
Can an attacker navigate complex networks?



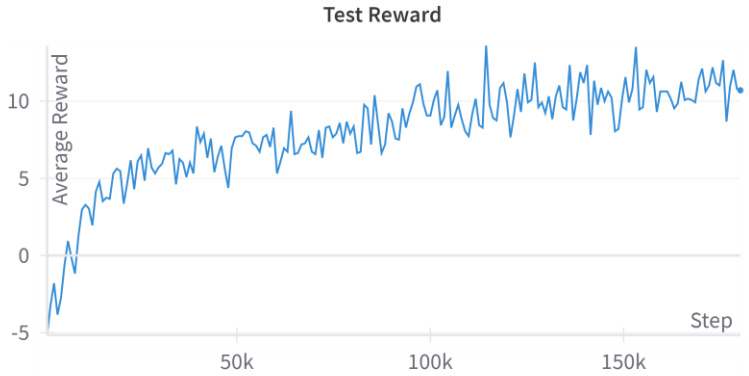
Visualization of the test network based on: A. Basak et. al. (2021), Scalable Algorithms for Identifying Stealthy Attackers in a Game-Theoretic Framework Using Deception

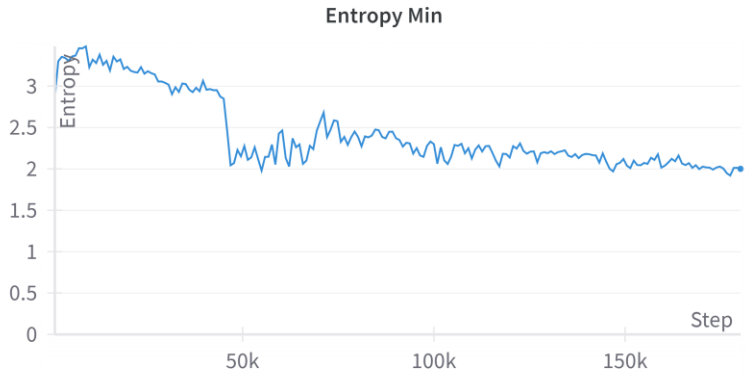
MLP Policy Network

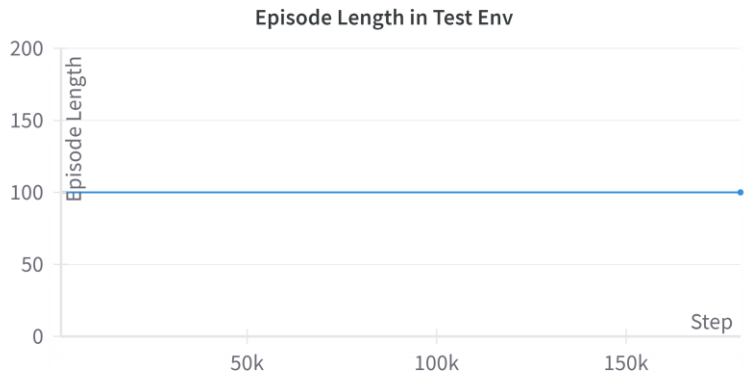
Can a Feed Forward Neural Network be used for Autonomous Cyber Agents?



Multilayer Perceptron (MLP) Policy and Value Network with shared Feature Extractor.

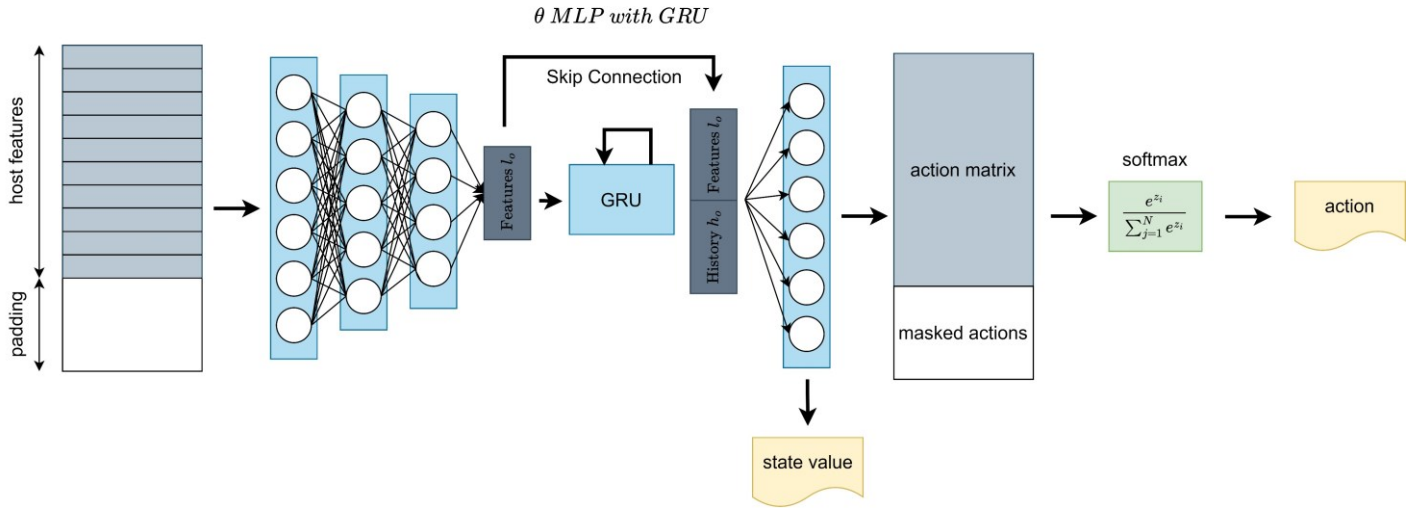




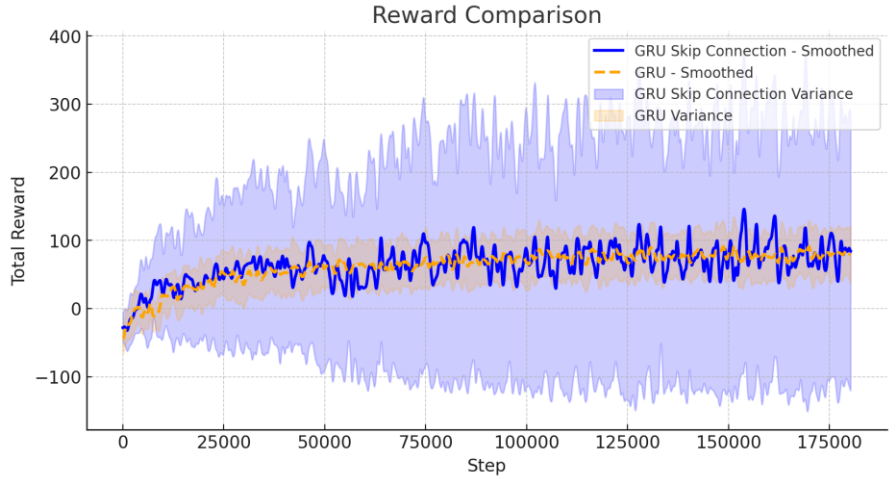


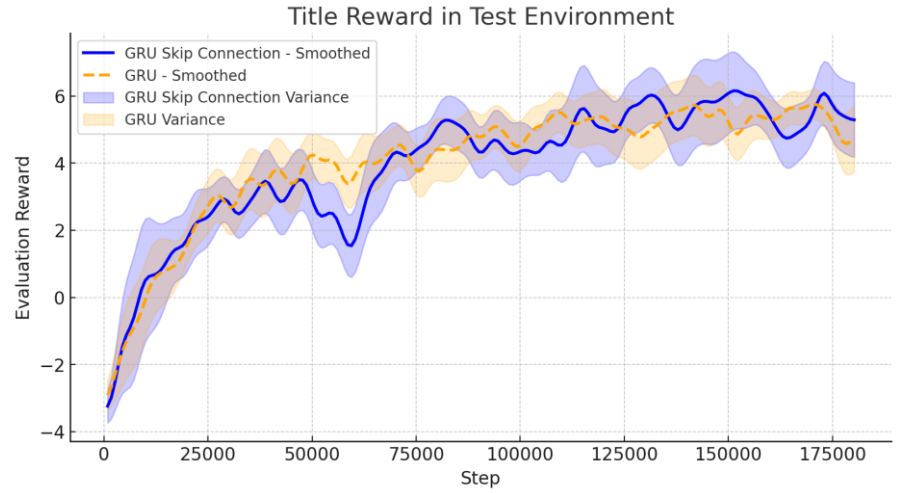
GRU Policy Network with Skip Connections

Can residuals help to prevent vanishing gradients?



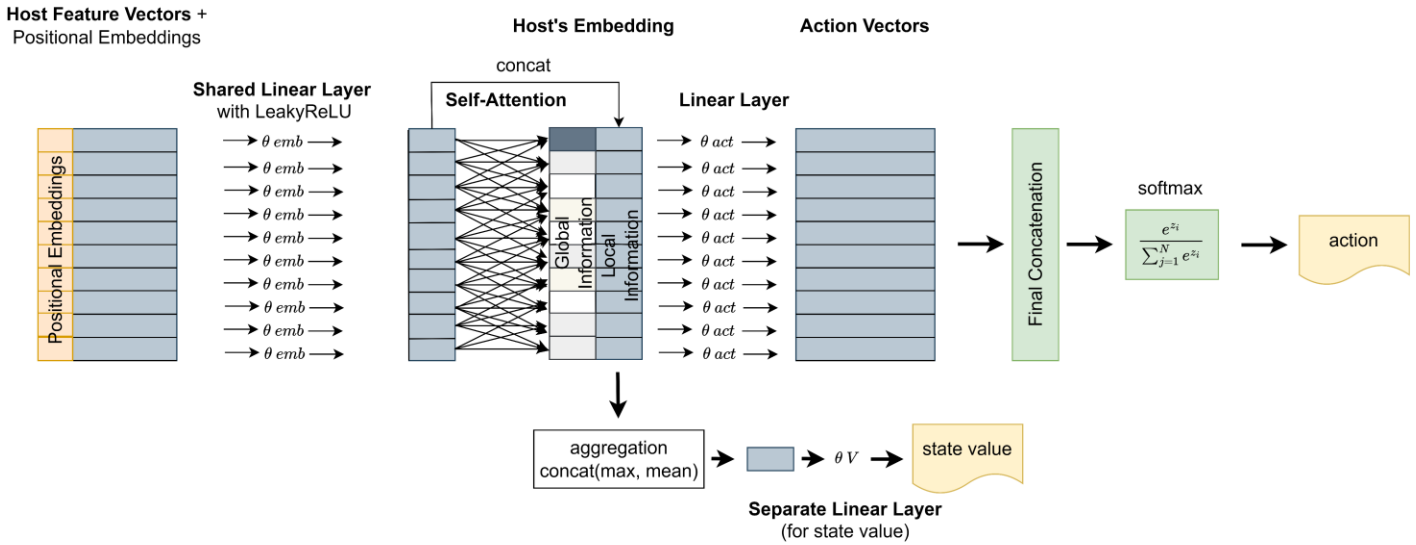
Implementation of a Gated Recurrent Unit (GRU) to realize the memory component in the policy-value network with Skip Connections to prevent vanishing gradients.



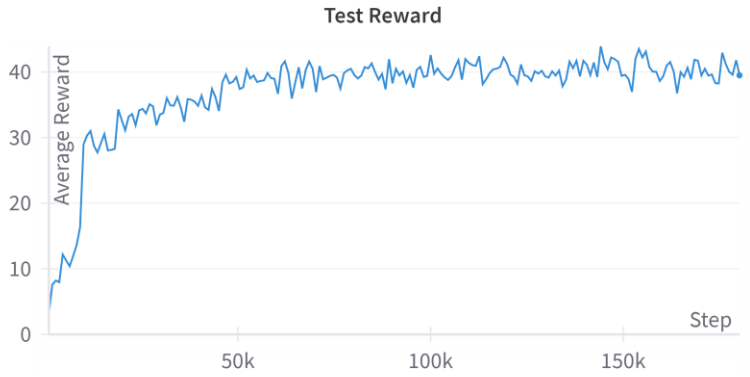


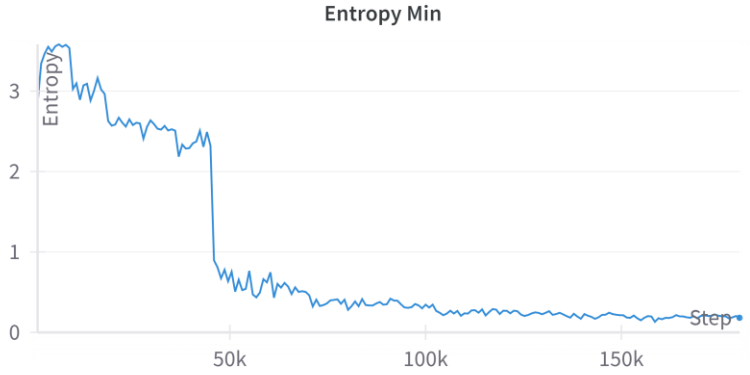
Self Attention Policy Network

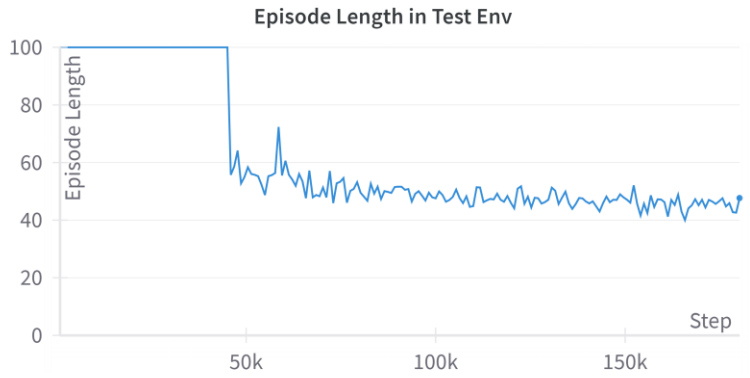
Can the weighting of local information improve decision-making?



Implementation of a Self-Attention Mechanism to realize the memory component to leverage local and global information.







OST

WO WISSEN WIRKT.

Active Defender

How can cyber security be learned as a game?

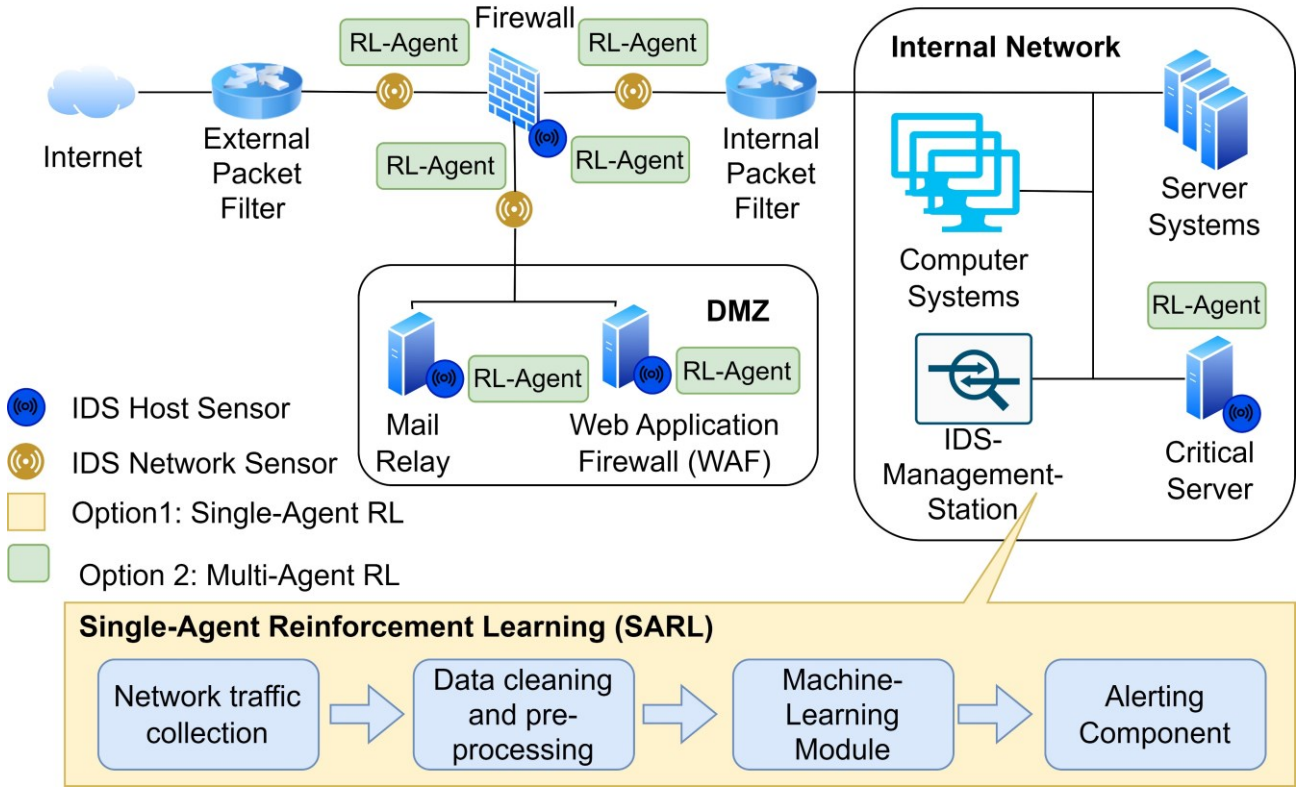
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus

The Defender

How can the defender protect his resources?

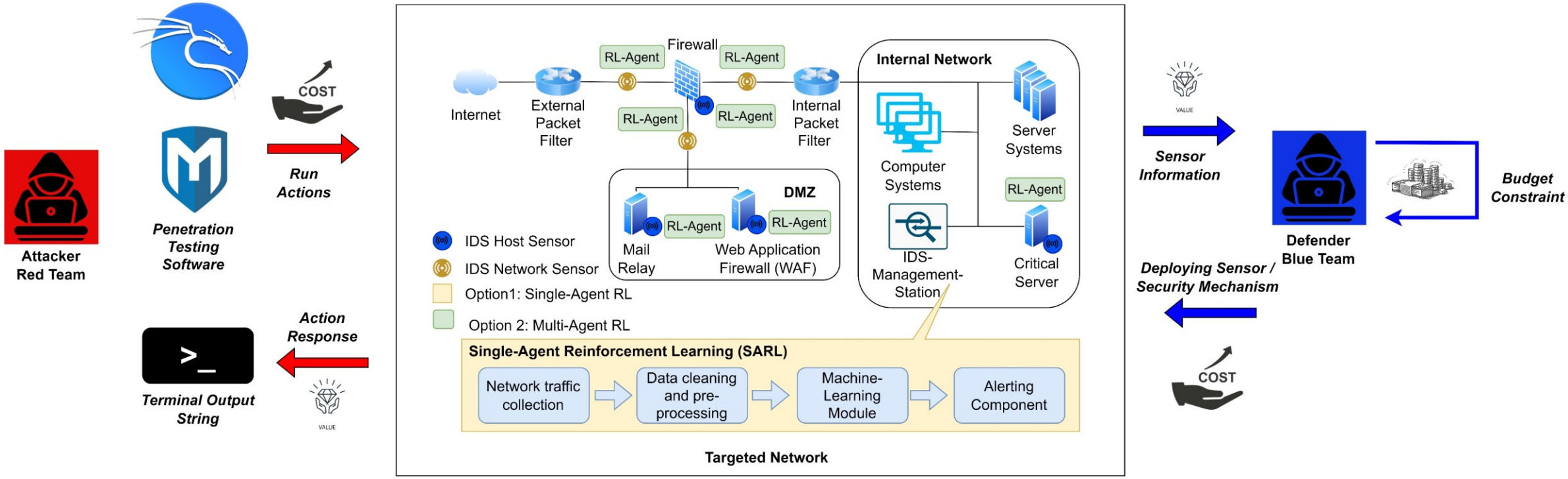
- Deploying Host sensors
- Deploying Network Sensors
- Deploying Security Mechanisms



Placement of network sensors for different architectures for intrusion detection systems (IDS)

Attacker-Defender Dynamics

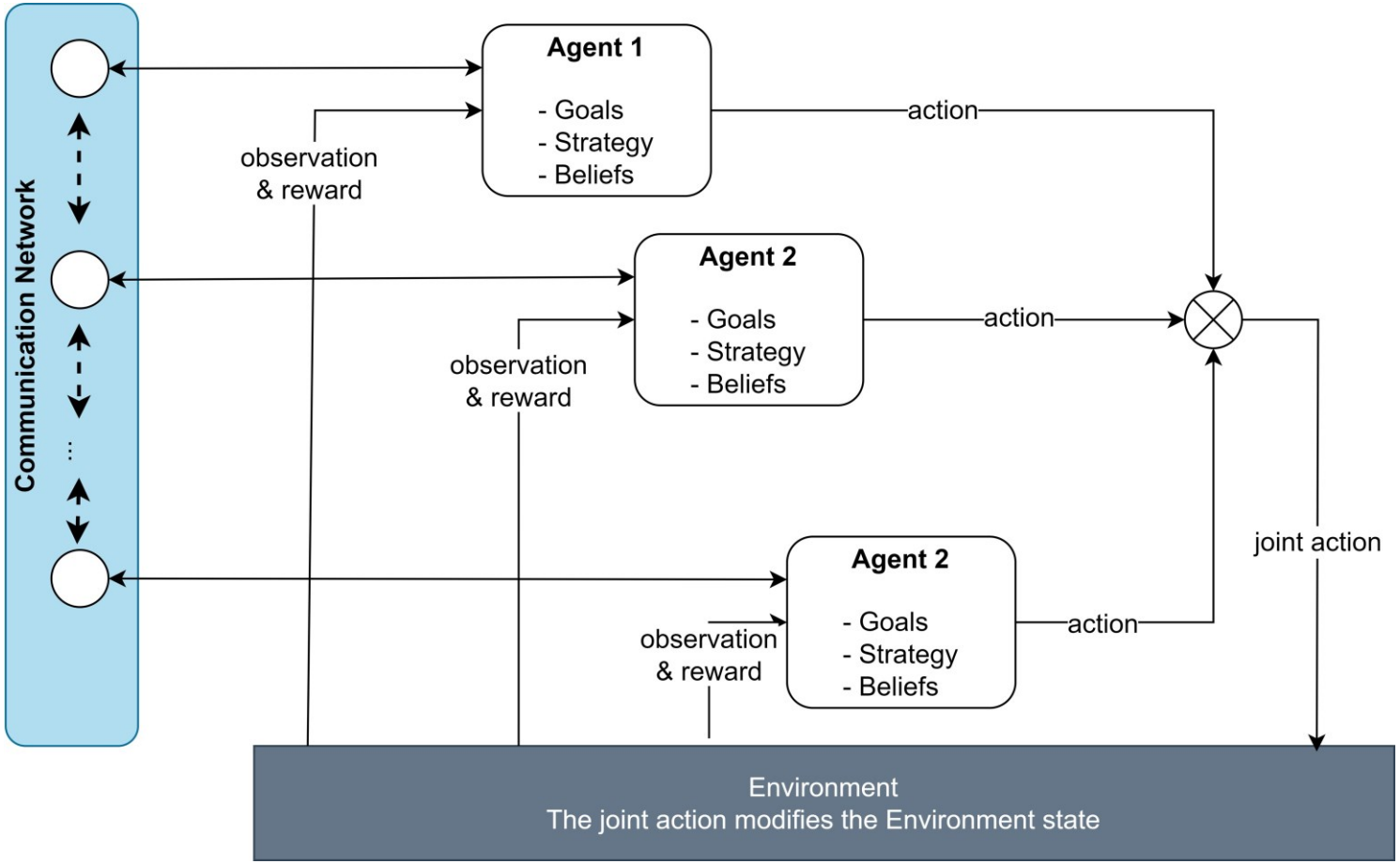
How Do Attackers and Defenders Compete in the Cybersecurity Game?



Gamification of attacker-defender dynamics with defender cost constraint.

Multi-Agent Reinforcement Learning Loop

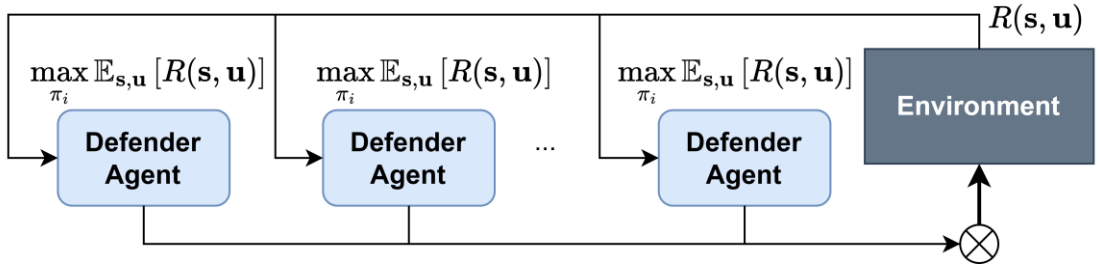
How to train multiple agents in a shared environment?



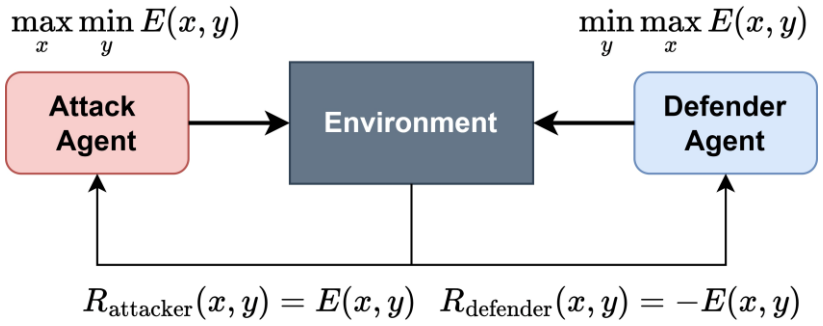
MARL control loop

MARL Training Setup

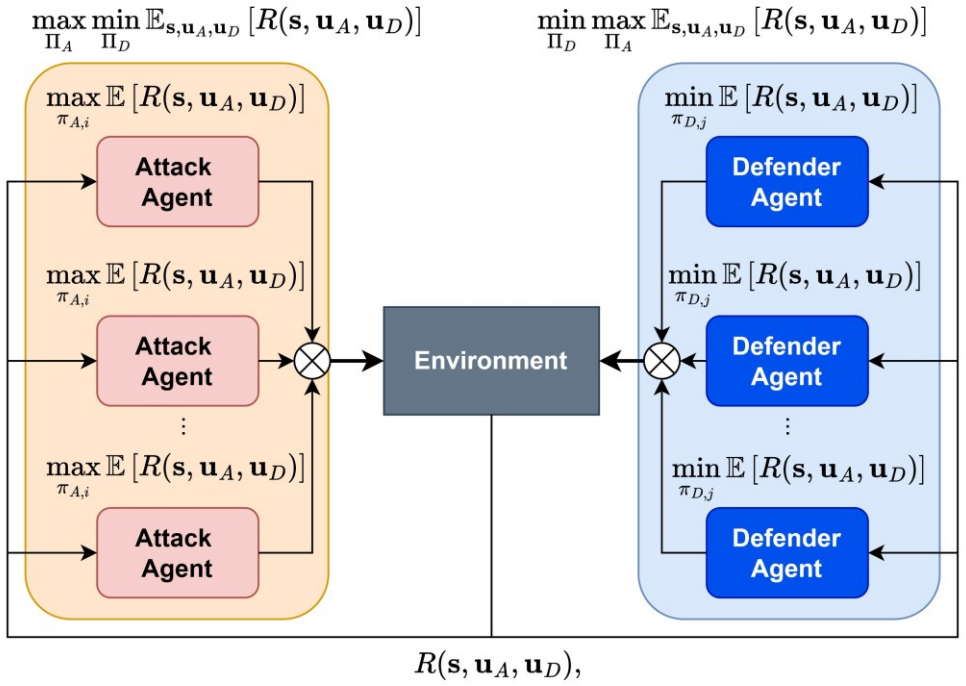
How can MARL be used to train AICA?



Training of a distributed defender



Training of an attacker vs a defender

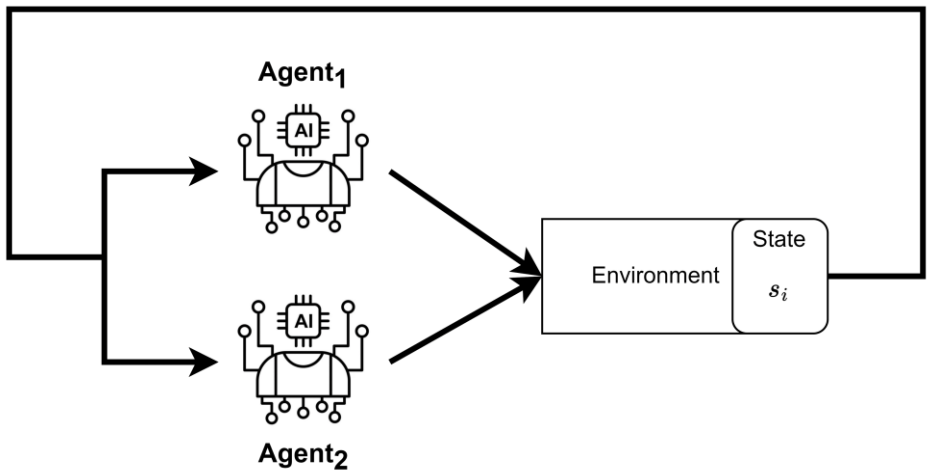


Team of Attacker vs a distributed defender

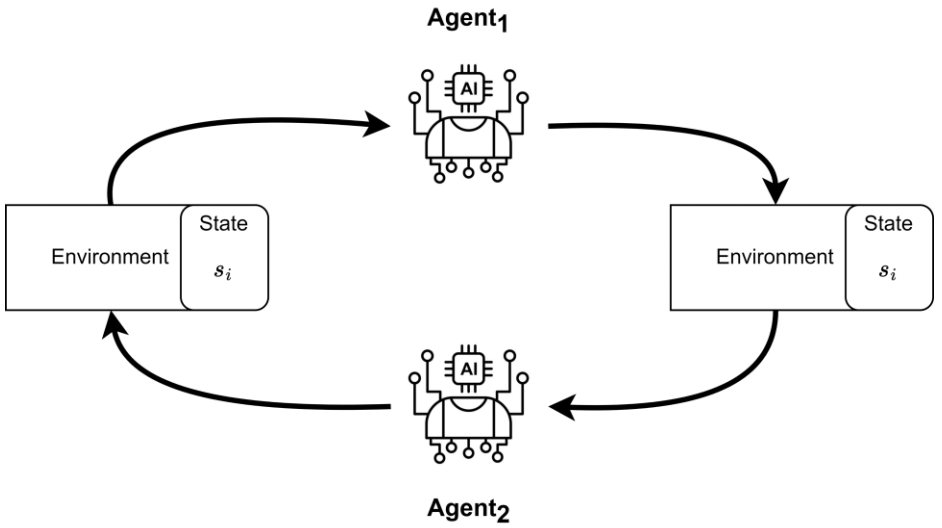
MARL – Game Design

How can game-theoretical dynamics be modeled in MARL?

Feature	Zero-Sum Game	Stackelberg Game
Interaction	Simultaneous, direct competition	Leader-follower (sequential)
Defender's Role	Reacts equally to attacks	Moves first, optimizes proactively
Attacker's Role	Always competes to maximize own gain	Observes and optimizes attack based on defense



Partially Observable Stochastic Games (POSGs) modelled as Parallel Game



Game model as Agent Environment Cycle (AEC)

Open Challenges

What should be done next?

- **Policy Generalization Failure:**
  - Abstract simulations lead to overfitting and poor transferability to real-world systems.
- **▪Large State/Action Spaces:**
  - Impede the convergence and efficiency of the training process.
- **▪Limited MARL Tooling:**
  - Existing frameworks lack robust support for multi-agent scenarios.

Check out our Paper

OST

WO WISSEN WIRKT.

Q&A Session

Open floor for questions, discussion, and feedback

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus