Enhancing Cyber Attack Autonomy Through Multi-Agent Reinforcement Learning

Christoph R. Landolt[(1)], Valentin Mulder[(2)], Julian Jang-Jaccard[(2)],
Christoph Würsch[(3)], Roland Meier[(2)], Mario Fritz[(1)]

28.05.2026

CYCON

(1) CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

(2) Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
armasuisse

(3) OST Ostschweizer Fachhochschule

The Rise of Autonomous Vulnerability Discovery

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Aug 2025

1. AUTOMATION
AI handles tasks

2. ORCHESTRATION
AI coordinates operations

3. EXPLOITATION AT SCALE
AI discovers and weaponizes

4. ECOSYSTEM DISRUPTION
Incentives and systems break

CYCON

The Rise of Autonomous Vulnerability Discovery

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Aug 2025

AI-Orchestrated Espionage
First AI orchestrated campaign disclosed by Anthropic.

Nov 2025

1. AUTOMATION
AI handles tasks

2. ORCHESTRATION
AI coordinates operations

3. EXPLOITATION AT SCALE
AI discovers and weaponizes

4. ECOSYSTEM DISRUPTION
Incentives and systems break

CYCON

The Rise of Autonomous Vulnerability Discovery

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Aug 2025

AI-Orchestrated Espionage
First AI orchestrated campaign disclosed by Anthropic.

Nov 2025

curl Bounty Program Ended
Maintainer burnout from a massive influx.

Jan 2026

CYCON

The Rise of Autonomous Vulnerability Discovery

Aug 2025

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Nov 2025

AI-Orchestrated Espionage
First AI orchestrated campaign disclosed by Anthropic.

Jan 2026

curl Bounty Program Ended
Maintainer burnout from a massive influx.

Feb 2026

Claude Opus 4.6
Model discovers 500+ high-severity vulnerabilities in 48h.

CYCON

The Rise of Autonomous Vulnerability Discovery

Aug 2025

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Nov 2025

AI-Orchestrated Espionage
First AI orchestrated campaign disclosed by Anthropic.

Jan 2026

curl Bounty Program Ended
Maintainer burnout from a massive influx.

Feb 2026

Claude Opus 4.6
Model discovers 500+ high-severity vulnerabilities in 48h.

Mar 2026

IBB Paused
Internet Bug Bounty overwhelmed by AI reports.

CYCON

The Rise of Autonomous Vulnerability Discovery

DARPA AI Cyber Challenge (AIxCC)
Prove that AI-driven vulnerability detection and patching work.

Aug 2025

AI-Orchestrated Espionage
First AI orchestrated campaign disclosed by Anthropic.

Nov 2025

curl Bounty Program Ended
Maintainer burnout from a massive influx.

Jan 2026

Claude Opus 4.6
Model discovers 500+ high-severity vulnerabilities in 48h.

Feb 2026

IBB Paused
Internet Bug Bounty overwhelmed by AI reports.

Mar 2026

Claude Mythos Preview
Anthropic cancels launch of Claude Mythos.

Apr 2026

1. AUTOMATION
AI handles tasks

2. ORCHESTRATION
AI coordinates operations

3. EXPLOITATION AT SCALE
AI discovers and weaponizes

4. ECOSYSTEM DISRUPTION
Incentives and systems break

CYCON

Is Autonomous Vulnerability Discovery Solved?

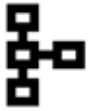
CYCON

Is Autonomous Vulnerability Discovery Solved?

Limited memory across long attacks
Agents forget earlier findings during long campaigns

CYCON

Is Autonomous Vulnerability Discovery Solved?



Limited memory across long attacks
Agents forget earlier findings during long campaigns

CTF Benchmark ≠ production
Winning a benchmark is not the same as breaching a real target

CYCON

Is Autonomous Vulnerability Discovery Solved?

Limited memory across long attacks
Agents forget earlier findings during long campaigns

CTF Benchmark ≠ production
Winning a benchmark is not the same as breaching a real target

Speed creates a detectable signature
Machine-speed actions look nothing like a human operator

CYCON

Is Autonomous Vulnerability Discovery Solved?

Limited memory across long attacks
Agents forget earlier findings during long campaigns

CTF Benchmark ≠ production
Winning a benchmark is not the same as breaching a real target

Speed creates a detectable signature
Machine-speed actions look nothing like a human operator

AI Agents act greedy
Agents chase the quickest win and get stuck in dead ends

CYCON

This presentation asks

What If Agents Escaped the Greediness by Working Together?

CYCON

From One Agent to a Team

Part 1

Single Agent

1

Advanced Targeted Attacks through Self-Play

Policy Network Selection

Part 2

Multi-Agent

2

Teach a team of agents

Results

Part 3

Conclusion

3

Findings

CYCON

From One Agent to a Team

Part 1

Single Agent

Part 2

Multi-Agent

Part 3

Conclusion

Advanced Targeted Attacks through Self-Play

Policy Network Selection

Teach a team of agents

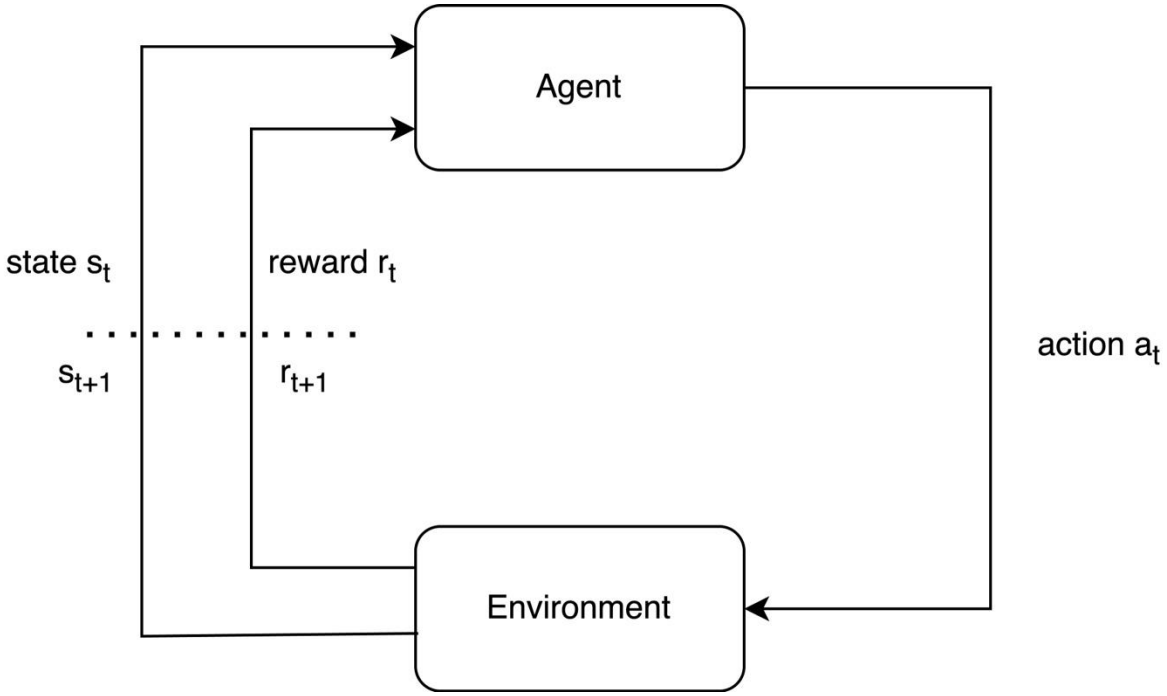
Results

Findings

CYCON

How to Learn Through Self-Play?

Introduction to Reinforcement Learning (RL)



RL control loop

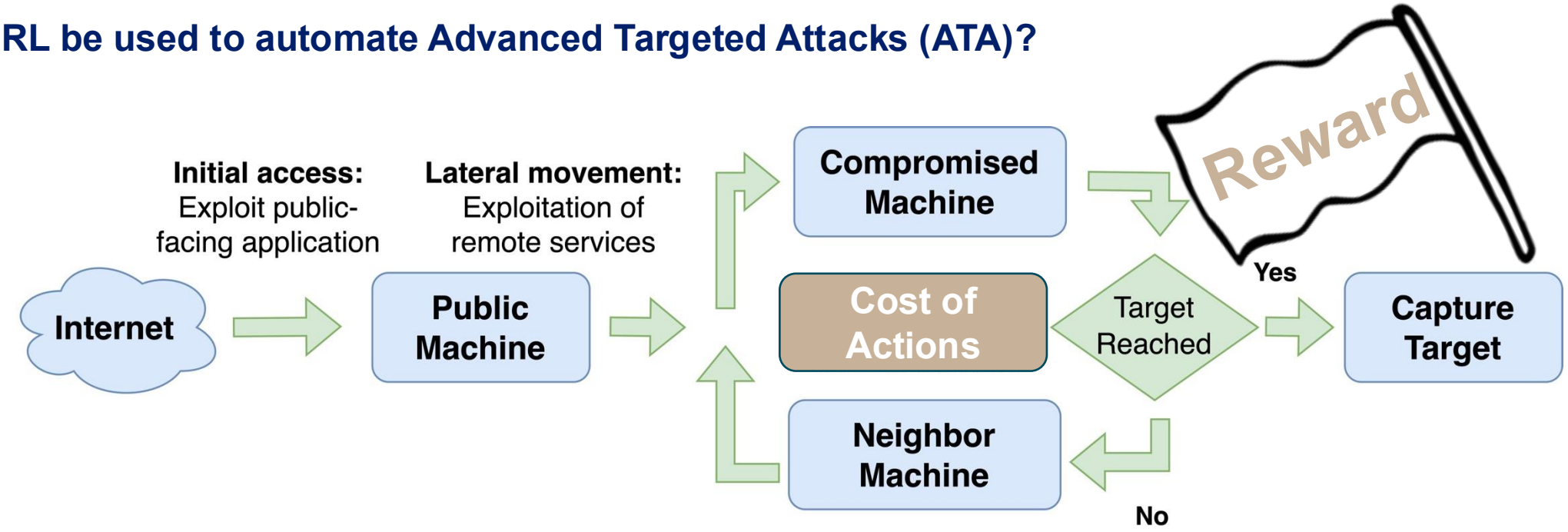
Use **Reinforcement Learning (RL)** when:

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

CYCON

Problem

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

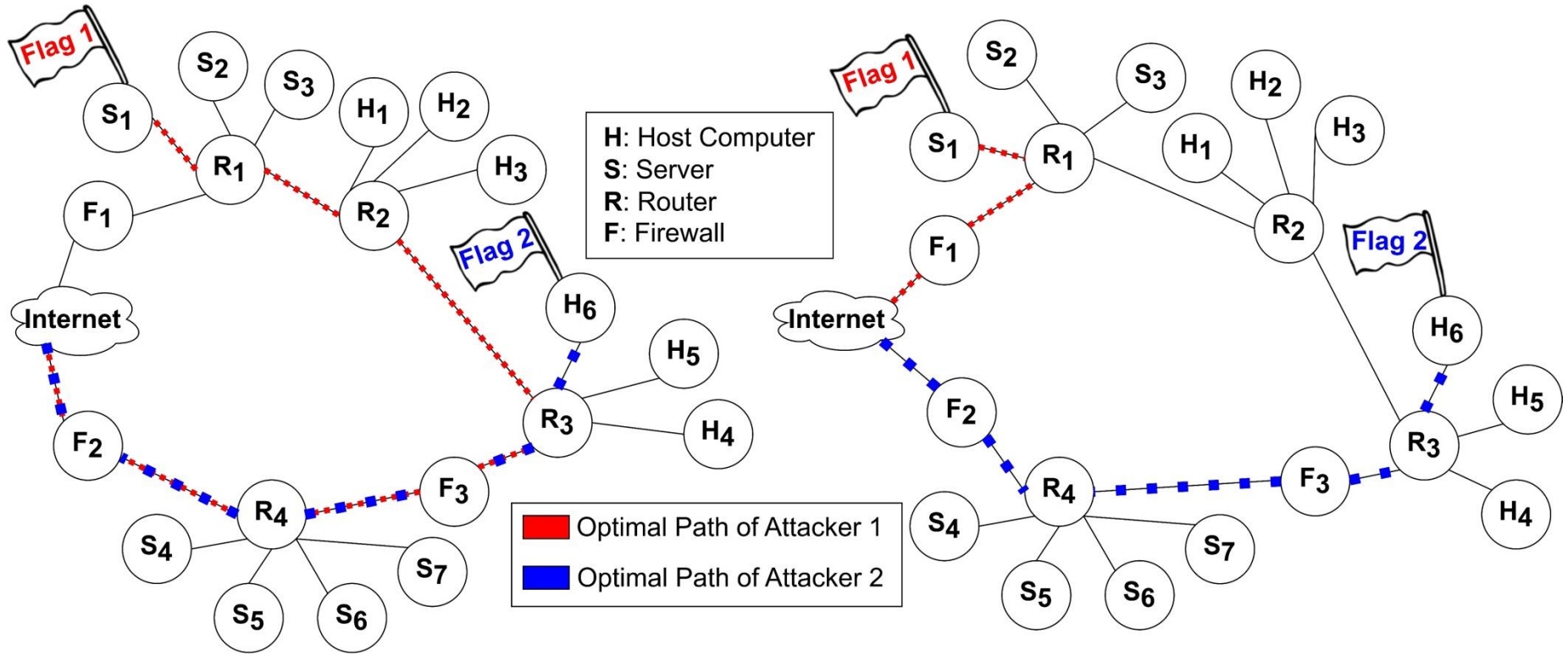


Attack graph starting from the Internet.

Sequential Actions / No training Data

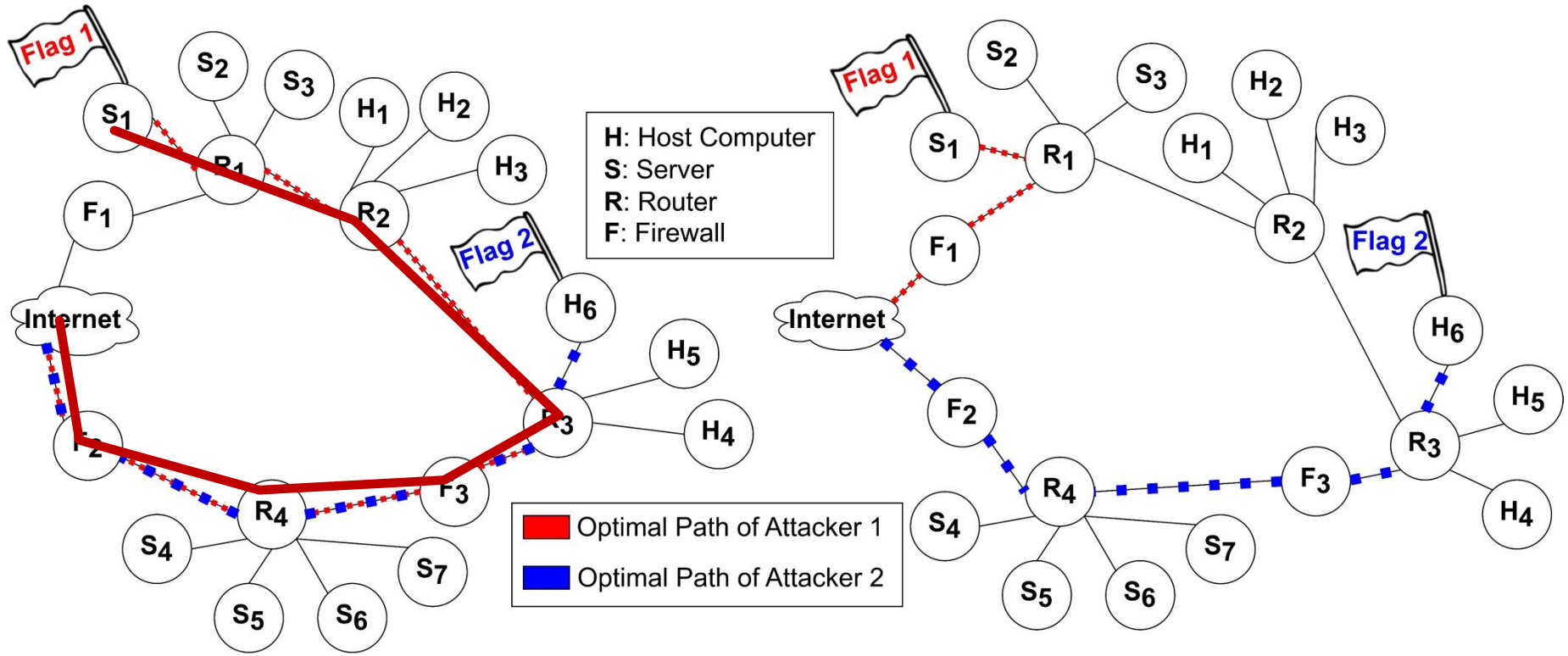
CYCON

Can an Attacker Navigate Complex Networks?



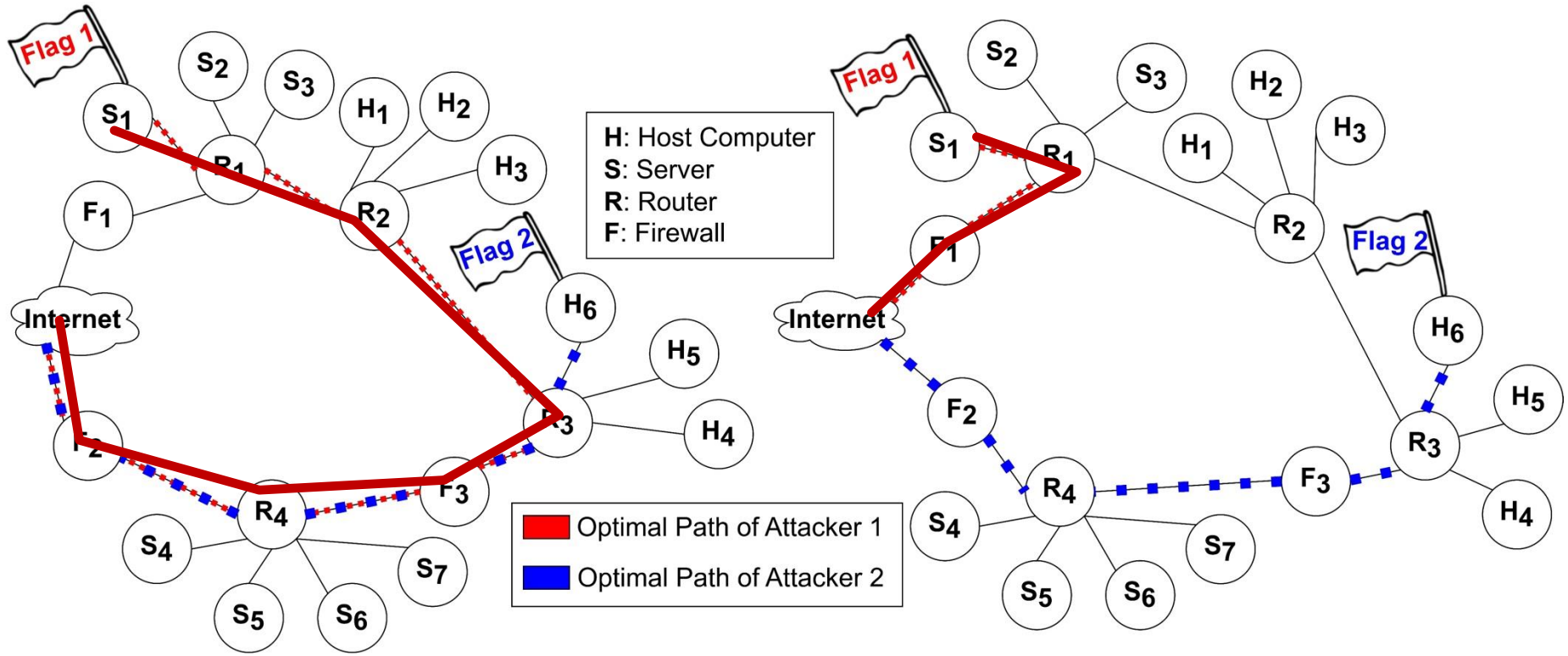
Visualization of the test network based on: A. Basak et. al. (2021), Scalable Algorithms for Identifying Stealthy Attackers in a Game-Theoretic Framework Using Deception

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

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

From One Agent to a Team

Part 1

Single Agent

Part 2

Multi-Agent

Part 3

Conclusion

Advanced Targeted Attacks through Self-Play

Policy Network Selection

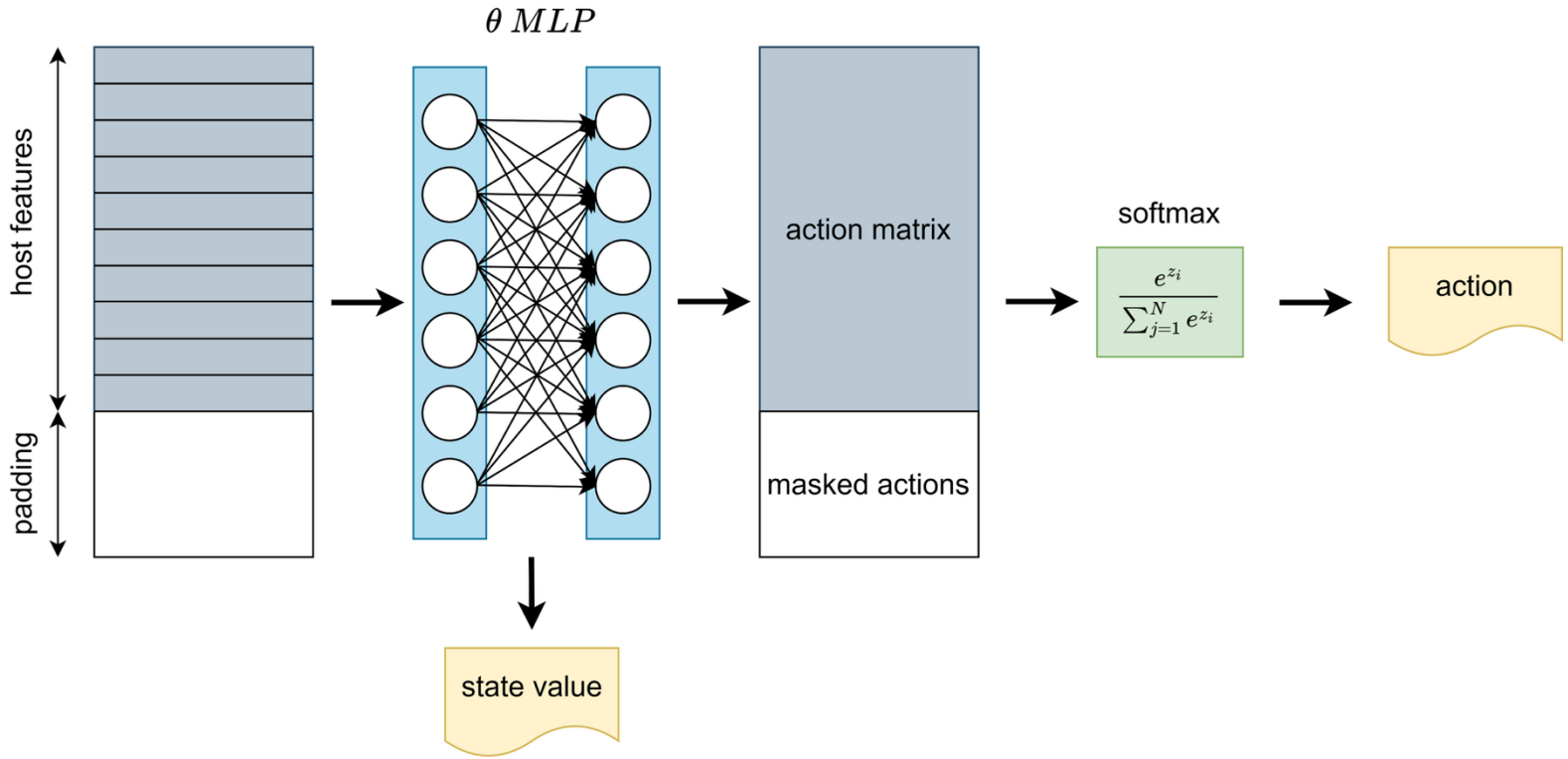
Teach a team of agents

Results

Findings

CYCON

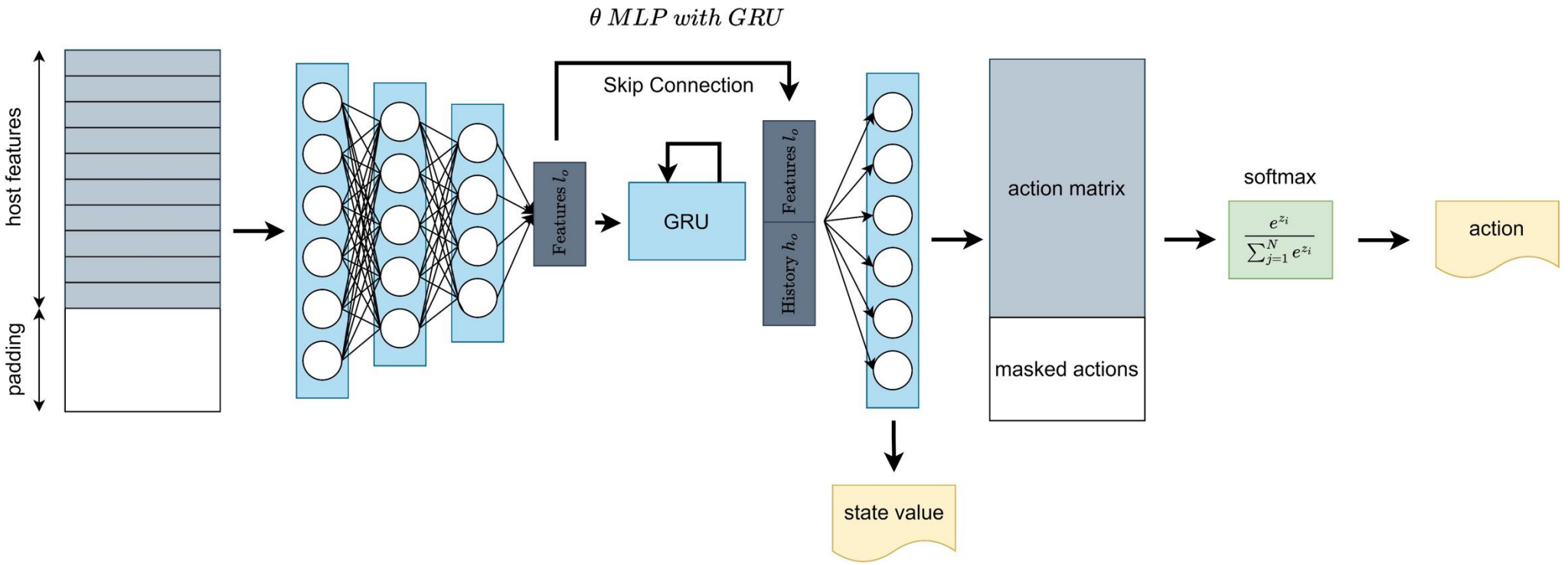
Can a Feed-Forward Neural Network Be Used for Autonomous Cyber Agents?



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

CYCON

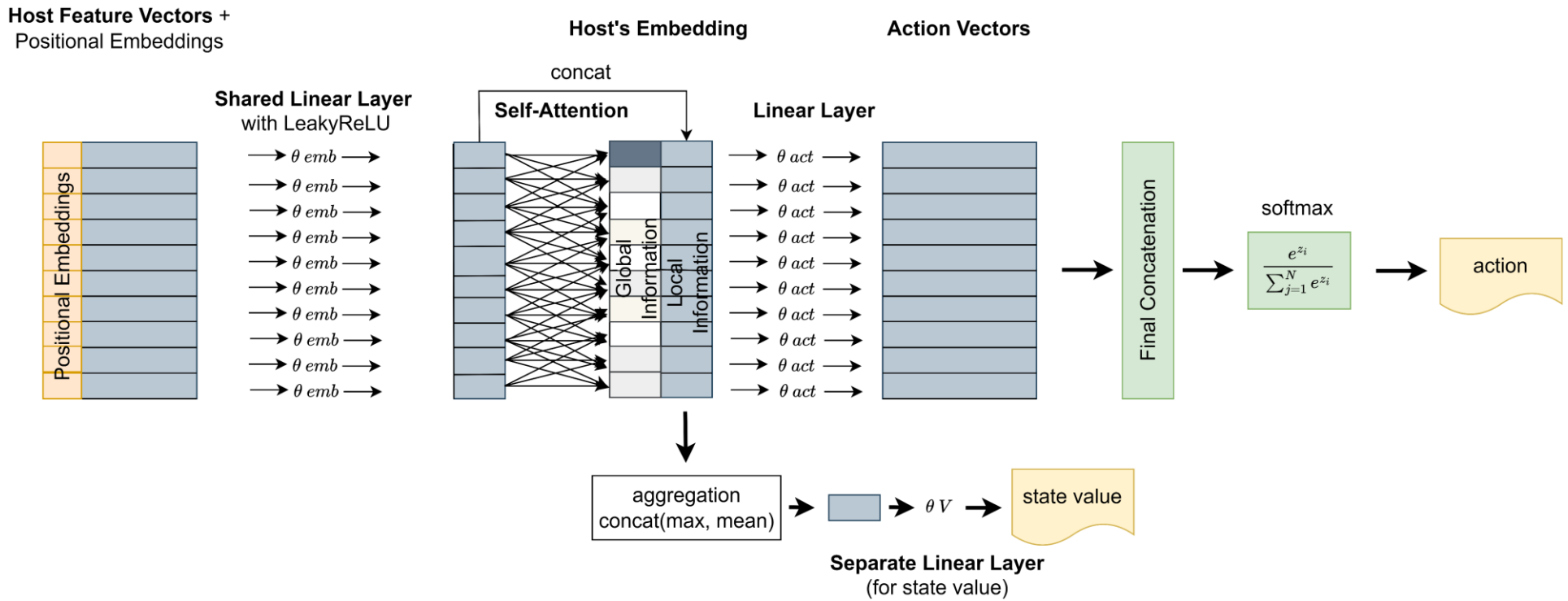
Can a Memory Module Help to Navigate Complex Scenarios?



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

CYCON

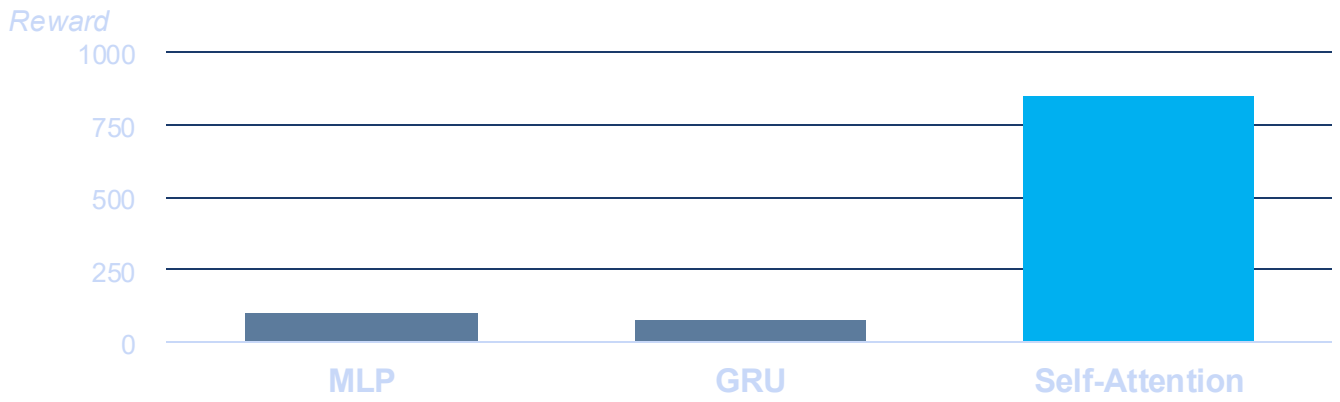
Can a Memory Module with Attention Help to Navigate Complex Scenarios?



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

CYCON

Results: Single Agent Training



Final rolling-mean reward at ~175k training steps (Fig. 6, paper).

Action distribution & convergence

	Entropy ↑	Kurtosis ↓	Status
MLP	2.33	8.54	Plateau
GRU	2.70	6.16	Plateau
Self-Attention	2.96	5.17	**Solved (grokked)**

CYCON

From One Agent to a Team

Part 1

Single Agent

Advanced Targeted Attacks through Self-Play

Policy Network Selection

Part 2

Multi-Agent

Teach a team of agents

Results

Part 3

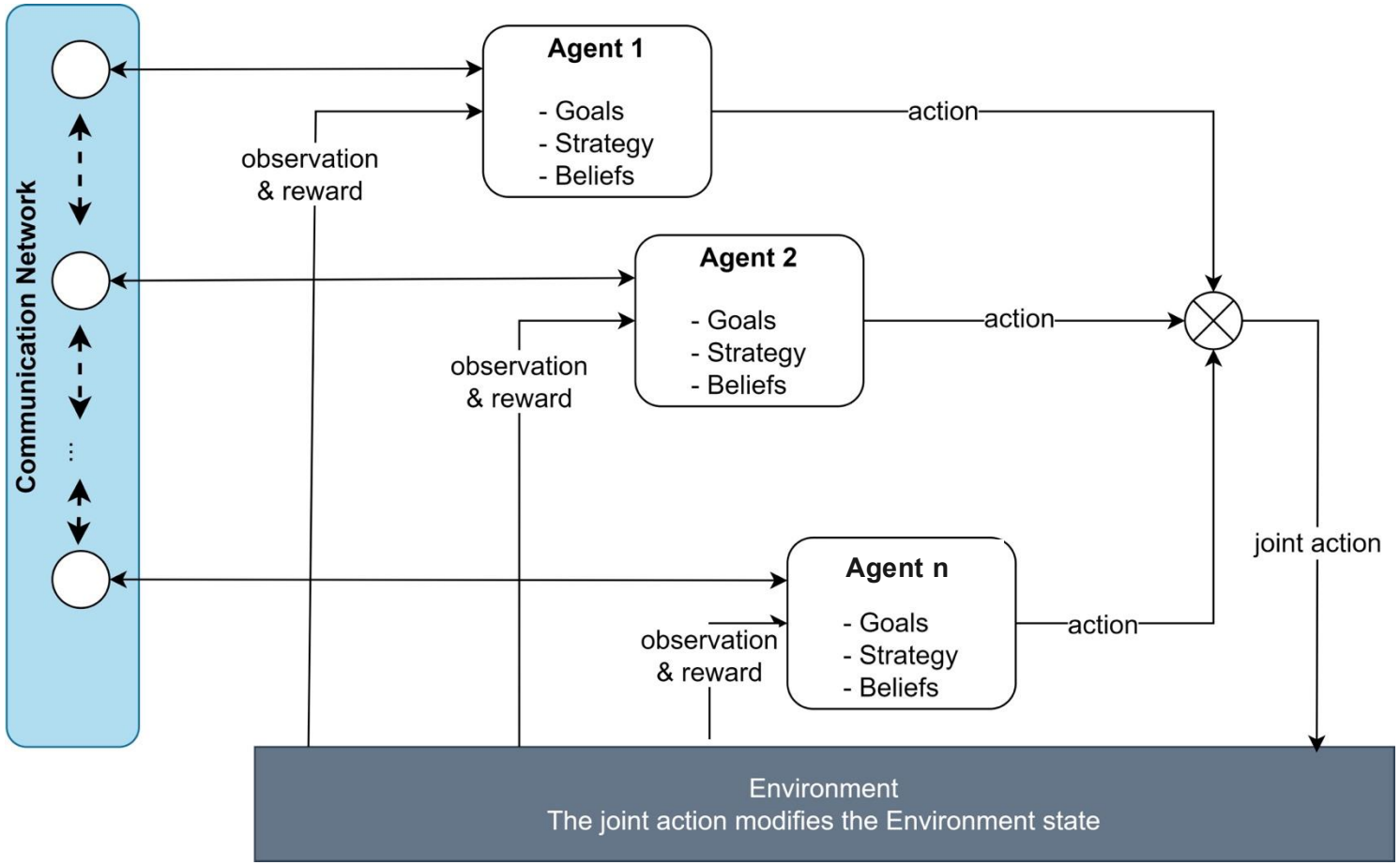
Conclusion

Findings

CYCON

Multi-Agent Reinforcement Learning Loop

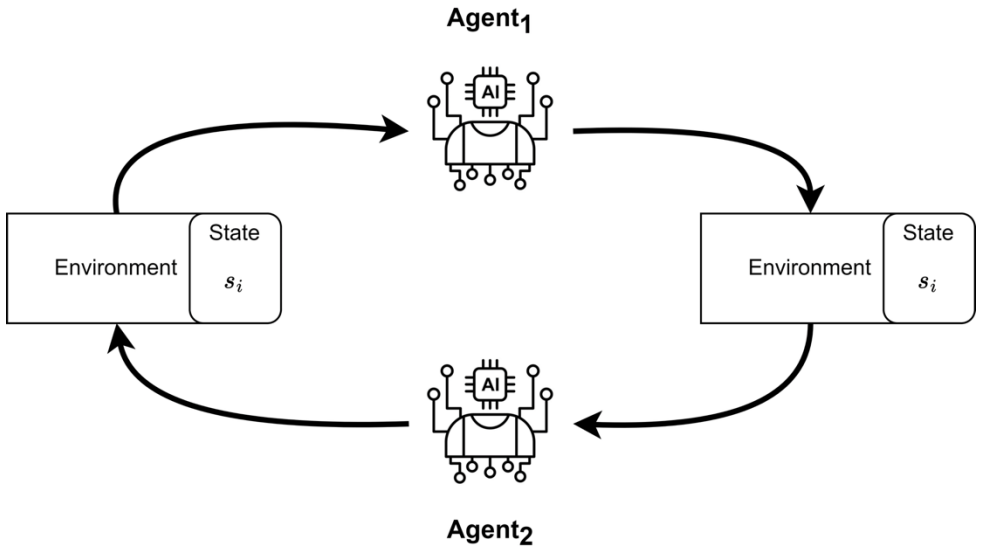
How to train multiple agents in a shared environment?



MARL control loop

Environment Non-Stationarity

How do agents in a self-play environment reach a cooperative equilibrium despite the observation corruption caused by simultaneous agent execution?

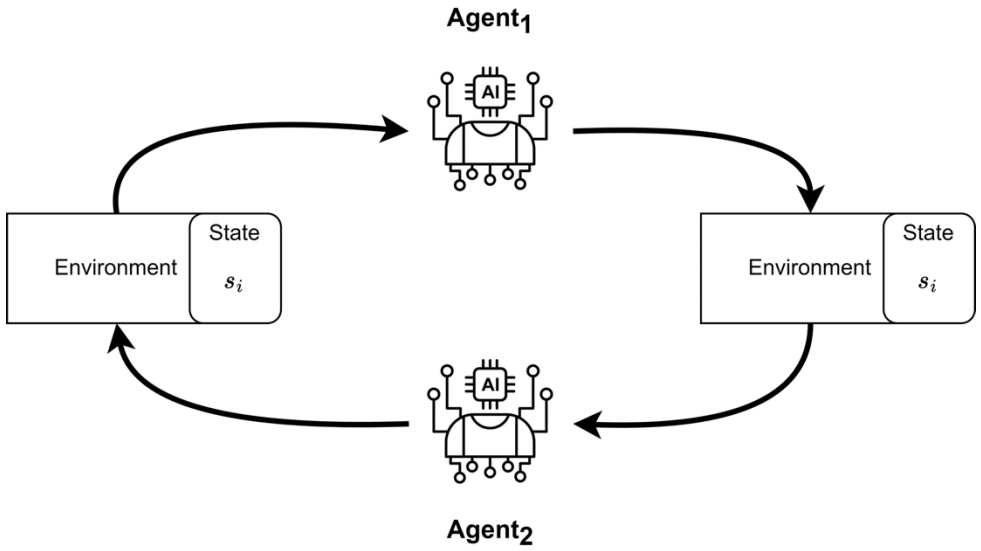


Game model as Agent Environment Cycle (AEC)

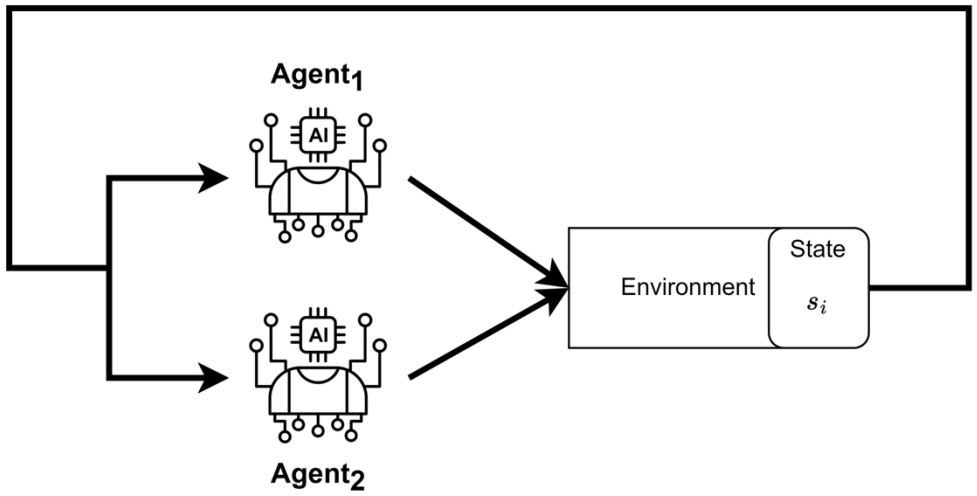
CYCON

Environment Non-Stationarity

How do agents in a self-play environment reach a cooperative equilibrium despite the observation corruption caused by simultaneous agent execution?



Game model as Agent Environment Cycle (AEC)



Partially Observable Stochastic Games (POSGs) modelled as Parallel Game

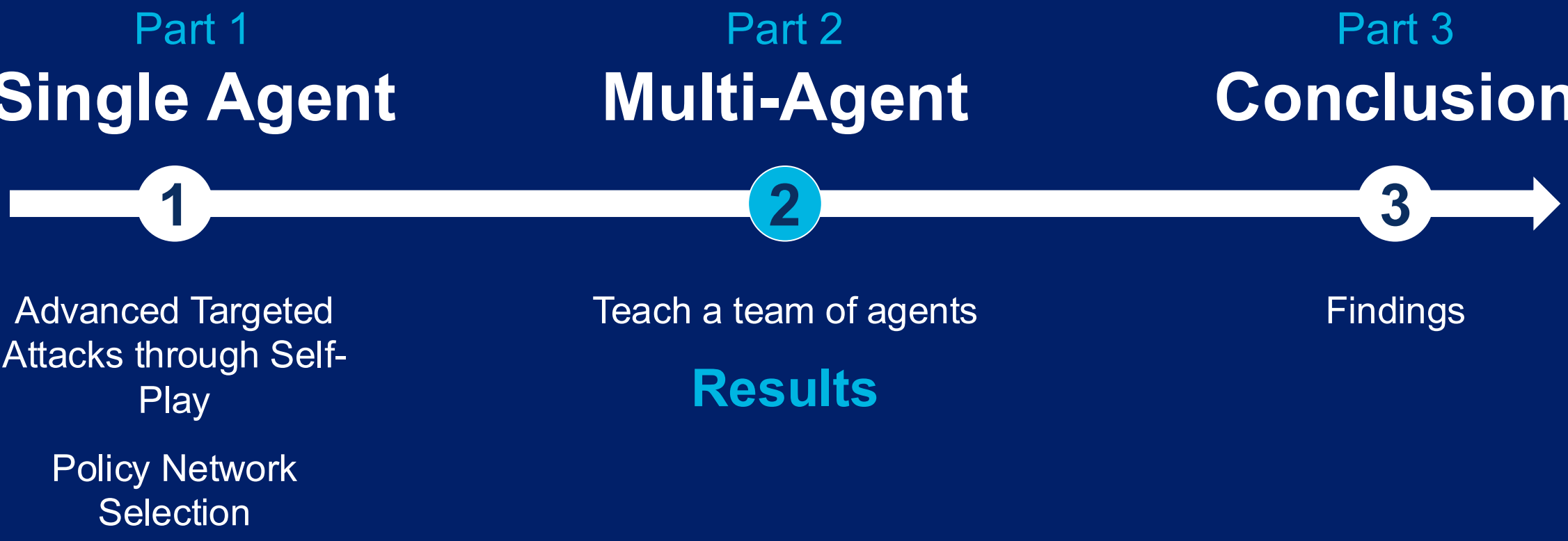
CYCON

The Effect of Non-Stationarity on the Memory Module

CYCON

Multi-agent (3 agents) MASAC policy evaluation for a tiny environment.

From One Agent to a Team



CYCON

The Scalability Bottleneck

Name	Subnets	Actions	States	SARL PPO	MASAC (3 Agents)	Step Limit
Tiny	4	18	576	**193.23 ± 0.68 (Reward)** 6.78 ± 0.36 steps	**58.51 ± 0.05** 8.82 ± 0.38 steps	1,000
Small	5	72	4,576	**184.83 ± 1.41 (Reward)** 8.48 ± 0.19 steps	**Did not Converge**	1,000
Medium	6	192	3.9×10^5	**181.87 ± 2.64 (Reward)** 9.15 ± 0.47 steps	**Did not Converge**	2,000
Large	8	322	4.5×10^6	**−89.93 ± 1.25 (Reward)** 100 ± 0 steps	**Did not Converge**	5,000

NASim Benchmark scenarios — specs and results

CYCON

From One Agent to a Team

Part 1

Single Agent

1

Advanced Targeted Attacks through Self-Play

Policy Network Selection

Part 2

Multi-Agent

2

Teach a team of agents

Results

Part 3

Conclusion

3

Findings

CYCON

Conclusion: What Can We Learn?

SARL Works

Needs entropy reg. + latent tuning

Memory Matters

Encoder + Memory + Controller

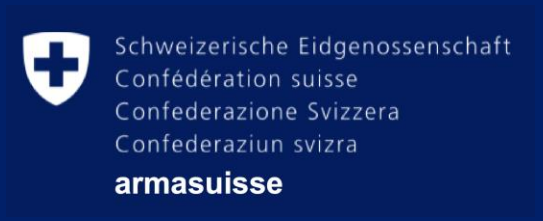
MARL: Promise vs. Scale

Non-stationarity kills scaling

💡 **Bottom Line:** MARL enables coordinated attacks, but scalability beyond tiny networks remains the critical open problem.

CYCON

Thank You



CISPA Helmholtz Center for Information Security

Christoph R. Landolt

christoph.landolt@cispa.de

CYCON