Building Better Environments for Autonomous Cyber Defence

Chris Hicks[1], Elizabeth Bates[1], Shae McFadden[1,2,3], Isaac Symes Thompson[1], Myles Foley[1], Ed Chapman[1], Nickolas Espinosa Dice[4], Ankita Samaddar[5], Joshua Sylvester[1], Himanshu Neema[5], Nicholas Butts[6], Nate Foster[7], Ahmad Ridley[8], Zoe M[1], Paul Jones[1]

[1]The Alan Turing Institute · [2]University College London · [3]King's College London · [4]Cornell University · [5]Vanderbilt University
[6]Microsoft · [7]EPFL · [8]NSA

The Alan Turing Institute

National Cyber Security Centre

Motivation

Agent-based **Autonomous Cyber Defence (ACD)** systems that monitor, adapt, and respond at machine speed are key to countering cyber threats to critical infrastructure. **Reinforcement Learning (RL)** is uniquely suited: it learns from interaction, without depending on prior human assumptions adversaries routinely exploit. We hosted a workshop to discuss what makes a good RL cyber environment, resulting in the following contributions:

1. **A framework** decomposing the components and modelling choices involved in mapping between ACD environments and real systems.
2. **Best-practice guidelines** for each framework component, a reference for anyone building cyber environments.

Workshop

25 DOMAIN EXPERTS

3 POINTS OF VIEW
- Network Security
- Reinforcement Learning
- Environment Development

A structured online workshop convening 25 domain experts from academia, industry and government with experience in RL-based cyber defence.

The Framework

- The **sim-to-real gap** hinders transfer of agent performance from training to deployment.
- High-fidelity environments help but are **not sufficient**. The RL task formulation itself constrains what and how an agent learns.
- Insufficiency in either **virtualisation** or **modelling** limits real-world performance.

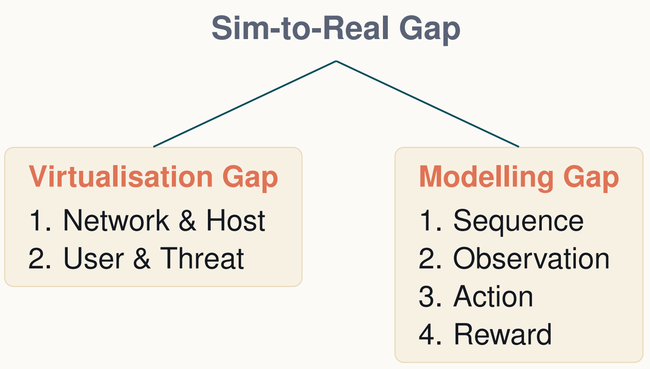

Figure 1: Breakdown of sim-to-real gap components.

Virtualisation Gap

Network & Host Simulation

High-fidelity is preferable. This minimises the virtualisation gap.

Low-fidelity creates simulator artefacts: sim performance ≠ real-world effectiveness.

User & Threat Simulation

Red agents shape network threats - simplistic attackers yield brittle defenders.

Green agents model benign users - omitting them risks policies that may disrupt normal traffic.

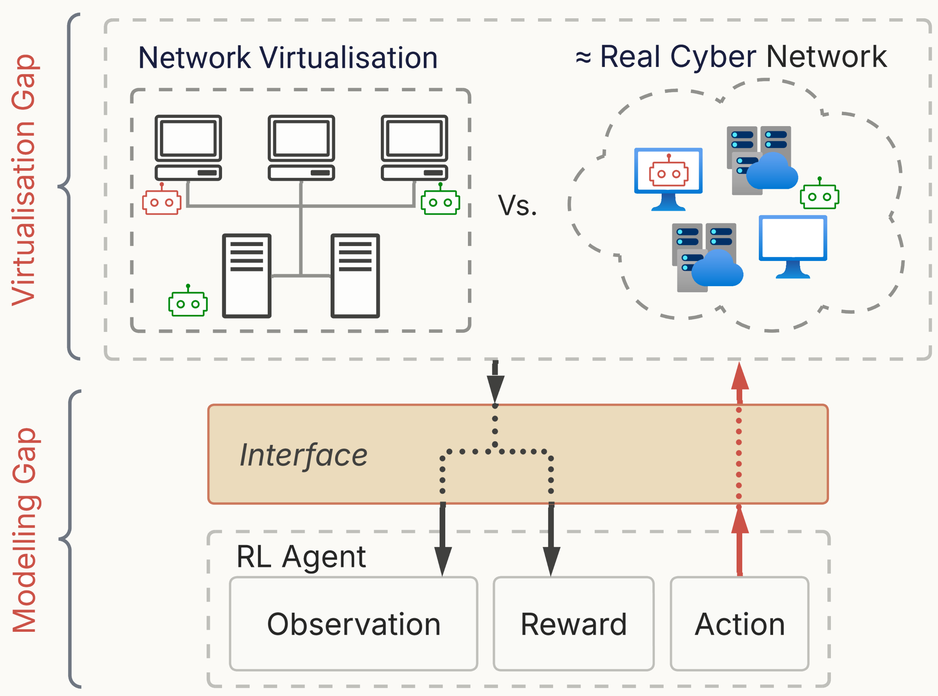

Figure 2: Virtualisation vs. modelling gap.

Modelling Gap

The modelling gap comprises the interface between the network defence problem and the RL task (i.e., a POMDP).

Sequence Modelling

Maps asynchronous network activity to discrete RL trajectories and drives choice of γ and episode horizon.

Observation Modelling

Aggregates network state into a vector from practically available data.

Action Modelling

Maps realistic defender operations to a discrete or continuous action set.

Reward Modelling

Provides scalar incentives whose maximisation corresponds to achieving the defensive goal.

Best-Practice Guidelines

Characterising the problem

- **Specify the problem, constraints & success criteria**
- **Identify areas of uncertainty**
- **Detail a minimum viable problem**
- **Explicitly document assumptions**

Virtualisation

- **Scope existing environments**
- **Validate virtualisation**
- **Move beyond fixed attackers**
- **Representative user activity**
- **Consider generative agents**

Sequence & Observation

- **Justify episodic structure & horizon**
- **Represent action duration & concurrency**
- **Reconsider turn-based modelling**
- **Define a realistic observation pipeline**
- **Consider factorised observations**

Action & Reward

- **Consider the granularity-capability trade-off**
- **Mask invalid actions**
- **Ensure effects are reflected in observations**
- **Motivate the reward function**
- **Align with the ACD problem goal**
- **Prefer sparse, goal-aligned rewards**

Evaluation

- **Evaluate policy, not just reward.** Average episodic reward is insufficient - track system-level behaviour, i.e., which hosts are attacked and action distributions.
- **Ensure statistical validity.** Multiple independent seeds with reported confidence intervals.
- **Test out-of-distribution robustness.** Evaluate against unseen attacker strategies and network configurations to assess whether the policy generalises beyond training conditions.

Acknowledgements

We would like to additionally thank Dr. Andres Molina-Markham for their contributions to this work both in the initial workshop and during the write-up. This research was funded and supported by the UK National Cyber Security Centre (NCSC).

Contact us: aicd@turing.ac.uk

READ THE PAPER
arxiv.org/abs/2604.08805