USENIX SECURITY '26 SYSTEMATIZATION OF KNOWLEDGE

The Pitfalls of Deep Reinforcement Learning for Cybersecurity

PAPER

Shae McFadden[1,2,3], Myles Foley[2,6], Elizabeth Bates[2], Ilias Tsingenopoulos[4], Sanyam Vyas[5,2], Vasilios Mavroudis[1], Chris Hicks[2], Fabio Pierazzi[3]

[1]King's College London, [2]The Alan Turing Institute, [3]University College London, [4]KU Leuven, [5]Cardiff University, [6]Devotion AI Labs

DRL's application to cybersecurity is well motivated by its success in other domains.

However, the adversarial, non-stationary, partially-observable nature of cybersecurity complicates this application.

DRL agents optimize reward over time through successive interactions with their environment. This makes it a natural fit for cybersecurity, where adaptive sequential decision making is essential. However, as a result of this transition we have identified methodological pitfalls across the four stages of the DRL4SEC development lifecycle.

66 papers reviewed

11 pitfalls identified

average pitfalls per paper

≥ pitfalls in every paper

11 Pitfalls Identified across Four Stages of Development

Environment Modeling

M-MS MDP Specification — **59%**

The MDP is ill-defined, leading to state, action, reward, or transition dynamics to remain unclear.

→ Fully specify all four MDP components.

M-MC Modeling Correctness — **35%**

The environment violates the Markov assumption or misrepresents the true MDP of the security task.

→ Ensure the task is Markovian; align the MDP to it.

M-PO Partial Observability — **61%**

An inherently partially-observable task is treated as if fully observable.

→ Model as a POMDP and use recurrence or beliefs.

Agent Training

T-HR Hyperparameter Reporting — **68%**

All hyperparameters are not reported.

→ Report all hyperparameters & architecture.

T-VA Variance Analysis — **67%**

The inherent variance of DRL training is never analysed across policies.

→ Report over multiple seeds with variance.

T-PC Policy Convergence — **71%**

Convergence is not demonstrated before training is terminated.

→ Show convergence prior to termination.

Performance Evaluation

E-AM Application Motivation — **44%**

Using DRL over traditional methods is not justified empirically or theoretically.

→ Justify DRL and compare to standard baselines.

E-GA Gain Attribution — **50%**

Gains stem from extra information in the MDP rather than the learned policy.

→ Delineate gains between the MDP and policy.

E-EC Environment Complexity — **41%**

Training and evaluation occur in contrived or oversimplified environments.

→ Evaluate in realistic environments.

System Deployment

D-UA Underlying Assumptions — **29%**

Training or runtime requirements cannot be met in real-world deployment.

→ Verify assumptions hold at deployment.

D-NS Non-Stationarity — **55%**

A static-dynamics assumption is broken by evolving adversaries and drift.

→ Account for non-stationarity and drift.

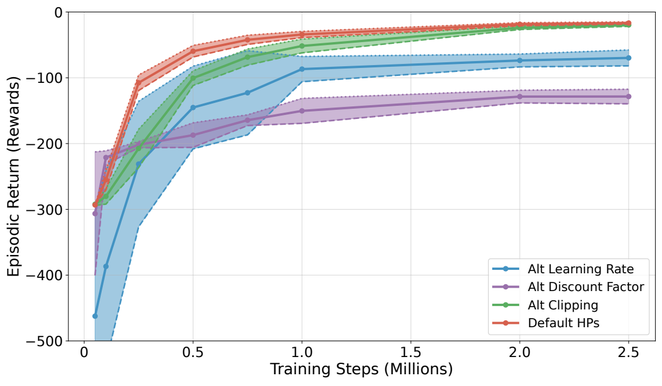
Pitfall Prevalence Criteria

- ☐ **Present:** the pitfall is fully exhibited.
- ☐ **Partial:** exhibited in part, present for some aspects of the pitfall.
- ☐ **Not present:** fully addressed and avoided.

Impact Demonstrated across Three Domains

Autonomous Cyber Defense — MiniCAGE

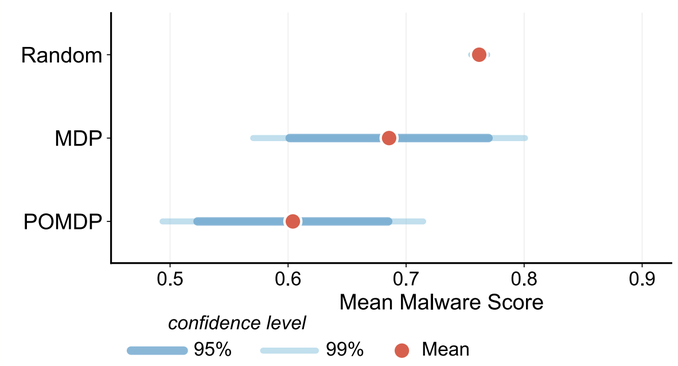
A blue agent defends a 13-host enterprise network against B-Line and Meander attackers targeting a critical server.



↳ Ablates training & deployment assumptions

Adversarial Malware Creation — AutoRobust

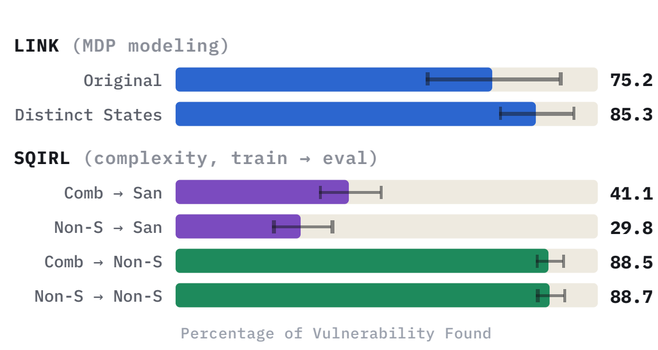
An agent edits sandbox behaviour reports to push malware across the detection boundary while keeping a plausible trace.



↳ Disentangles MDP design from policy performance

Web Security Testing — Link & SQIRL

Agents craft XSS and SQL-injection payloads against deliberately vulnerable web apps.



↳ Exposes the impact of modeling and complexity

Not a critique, but a new standard.

The pitfalls identified in this work are not failures of individual researchers, they are systematic challenges that emerge when adapting DRL to cybersecurity tasks. This paper aims to provide a shared standard, giving authors and reviewers a consistent bar for rigorous, reproducible, and practical DRL4SEC research.

PARTIALLY FUNDED BY National Cyber Security Centre, a part of GCHQ; EPSRC Engineering and Physical Sciences Research Council

AFFILIATIONS King's College London, UCL, KU Leuven, The Alan Turing Institute