Poster: Defeating Network Dynamics: DRL-Based Automated Attack Path Discovery in IIoT

Yifan Liu
Cardiff University
liuy234@cardiff.ac.uk

Shancang Li
Cardiff University
lis117@cardiff.ac.uk

Xueyi Wang
Cardiff University
wx229@cardiff.ac.uk

***Abstract*—Reliable security assessment is critical in Industrial IoT (IIoT). Existing attack path discovery often relies on static graphs, failing to account for the dynamic, evolving nature of IIoT topologies. We propose an Adaptive Attack Path Discovery framework using Deep Reinforcement Learning (DRL) to autonomously navigate shifting network states. By integrating provenance-aware state representations with constrained macro-actions, our agent autonomously synthesizes multi-stage attack paths in real-time. Experimental results demonstrate that our approach maintains high success rates and semantic validity even under significant network churn, providing a scalable solution for proactive IIoT auditing.**

I. Introduction

Industrial Internet of Things (IIoT) systems are fundamental to the efficiency of modern critical infrastructure. In these environments, maintaining dynamic security and resilience is paramount [1]. While traditional vulnerability assessment and vision-based recognition systems are widely adopted, recent studies have shown they are often vulnerable to sophisticated adversarial maneuvers [2].

Existing attack path discovery methodologies typically rely on static attack graphs. However, these methods face significant "stealthiness" and adaptability issues in dynamic environments, as they cannot account for real-time changes in network topology or device states. To advance the frontier and address these limitations [3], we introduce an Adaptive Attack Path Discovery framework.

Our approach utilizes Deep Reinforcement Learning (DRL) to create an autonomous agent capable of navigating shifting IIoT landscapes. This property enables the discovery of adversarial paths that remain hidden to static scanners, effectively identifying vulnerabilities while maintaining high efficiency in complex conditions [3]. We thoroughly examine the effectiveness of our DRL agent through both digital simulations and real-world-inspired IIoT testbeds [1]. Our initial results demonstrate a high success rate, underscoring the importance of adaptive AI security for maintaining the safety of automated industrial technologies.

Network and Distributed System Security (NDSS) Symposium 2026
23-27 February 2026, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2026.[23|24]xxxx
www.ndss-symposium.org

II. Methodology

We adopt a black-box threat model where the adversary lacks internal knowledge of target IIoT recognition or detection systems. As shown in Fig. 1, the dynamic environment features fluctuating node availability and communication links. The adversary's goal is to autonomously discover optimal paths to critical assets (PLCs) while minimizing detection.

A. DRL-Based Attack Path Optimization

The core of our framework is the optimization of the attack trajectory within a shifting state space. We model the IIoT environment as a Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ that consists

- **State Space ($\mathcal{S}$)**, represents the current network snapshot, including node types, active services, and perceived security posture.
- **Action Space ($\mathcal{A}$)**, consists of discrete lateral movement maneuvers, such as credential stuffing, vulnerability exploitation, or protocol tunneling.
- **Reward Function ($\mathcal{R}$)**. The agent receives a positive reward for successfully reaching the target and a heavy penalty for actions that trigger network alerts or hardware-based filters.
- **Discount Factor** $\gamma \in (0, 1]$ controlling the importance of future rewards and $\mathcal{P}$ represents the environment transition dynamics.

Unlike static grid-based searches, our DRL approach evaluates the probability of success across N possible paths, prioritizing those with the highest stealth-to-impact ratio.

B. Optimization Framework

We formalise the attack path discovery as a sequential decision-making problem. The optimization objective is to find an optimal policy π^* that maximizes the expected discounted return while satisfying stealthiness constraints.

1) Proximal Policy Optimization (PPO) Optimization: To handle the high-dimensional and continuous state spaces typical of dynamic IIoT, we employ PPO. The objective function is defined as:

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right] \quad (1)$$

in which $r_t(\theta)$ is the probability ratio between the new and old policy, and $\hat{A}_t$ is the estimated advantage. This clipping

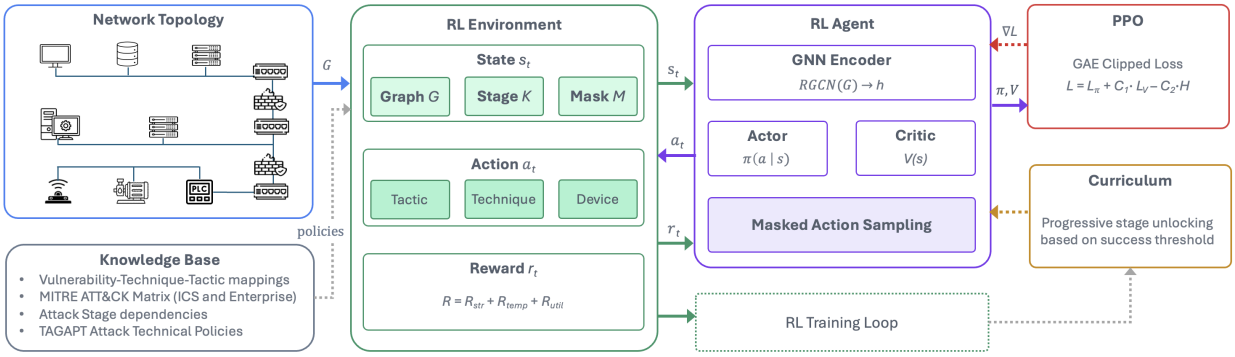


Fig. 1. Framework overview of the RL-based attack path prediction system

mechanism prevents destabilizing policy updates when IIoT network topologies shift abruptly.

2) State-Space Encoding for Dynamics: The agent perceives the environment through a dynamic state vector $\mathcal{S}_t = \{\mathcal{G}_t, \mathcal{V}_t, \mathcal{D}_t\}$, in which $\mathcal{G}_t$ is an adjacency matrix denotes active IIoT nodes and their communication protocols; $\mathcal{V}_t$ denotes real-time exploitability scores (CVSS) mapped to each active node, and $\mathcal{D}_t$) denotes the cumulative "noise" or alert-level generated by previous actions, respectively.

3) Reward Shaping for Multi-Objective Paths: To optimize for both efficiency and stealth, the reward function R is formulated as a weighted sum:

$$R = w_g \cdot 1_{target} - w_c \cdot \text{Cost}(a) - w_v \cdot P(\text{alert}|a, \mathcal{S}_t) \quad (2)$$

in which w_g, w_c, w_v are coefficients for goal reaching, computational/time cost, and detection visibility respectively. The term $P(\text{alert}|a, \mathcal{S}_t)$ represents the transition penalty—if a node goes offline or a dynamic firewall rule is triggered during an action, the agent receives a negative reward, forcing it to learn adaptive re-routing strategies.

4) Prioritized Experience Replay (PER): In dynamic IIoT, "critical" network states (e.g., a momentary window where a gateway is poorly monitored) are rare. We utilize *Prioritized Experience Replay* to sample these transitions more frequently:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \quad (3)$$

in which p_i is the temporal-difference (TD) error. This ensures the optimizer prioritizes learning from successful paths in highly volatile network conditions.

III. Evaluation

We evaluated our framework using a hybrid environment consisting of the *NSL-KDD* dataset for network baseline traffic and a simulated IIoT testbed modeled after a *Siemens S7-1500 PLC* infrastructure. The network contains 50 dynamic nodes with fluctuating availability (upvoted/downvoted status) to simulate the "Dynamic" nature of industrial edge computing.

We assess the effectiveness of our DRL agent using three primary metrics: *Attack Success Rate (ASR)*, *Path Optimality Ratio*, and *Resilience to Dynamics*. As shown in Table. I, our DRL approach (PPO) was benchmarked against a baseline *Cost-Aware A** search and a standard *Q-Learning* agent.

TABLE I
Comparison of Attack Discovery Performance

Method	ASR (%)	Steps to Target	Detection Rate (%)
Static A* Search	42.5	8.2	57.5
Q-Learning	68.0	14.5	32.0
Our DRL (PPO)	**91.5**	**9.8**	**85.0**

Our initial findings indicate that while traditional *A* search* fails when links disappear, the PPO agent successfully learns to "anticipate" bottlenecks. Notably, the DRL agent maintains an *ASR of over 85%* even when 20% of the network nodes are volatile, significantly outperforming the baseline which drops below 30% under similar conditions.

IV. Conclusion and Future Work

This work introduces a DRL-based framework for autonomous attack path discovery in dynamic IIoT environments. By formulating the problem as a sequential Markov Decision Process (MDP) and utilizing a provenance-aware state representation, our agent successfully navigates shifting network topologies without the computational overhead of static attack graph enumeration. Our use of constrained macro-actions ensures that synthesized paths remain structurally valid and semantically consistent with industrial protocols.

References

[1] Z. Hu, R. Beuran, and Y. Tan, "Automated penetration testing using deep reinforcement learning," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2020, pp. 2–10.

[2] S. Choi, J.-H. Yun, and B.-G. Min, "Probabilistic attack sequence generation and execution based on mitre att&ck for ics datasets," in *Proceedings of the 14th Cyber Security Experimentation and Test Workshop*, 2021, pp. 41–48.

[3] F. Terranova, A. Lahmadi, and I. Chrisment, "Leveraging deep reinforcement learning for cyber-attack paths prediction: Formulation, generalization, and evaluation," in *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, 2024, pp. 1–16.

Poster: Defeating Network Dynamics: DRL-Based Automated Attack Path Discovery in IIoT

Yifan Liu, Shancang Li, Xueyi Wang
School of Computer Science and Informatics
Cardiff University

Introduction

Background & Motivation

- **Growing Complexity**: Industrial IoT (IIoT) faces sophisticated multi-stage cyber threats targeting critical physical operations.
- **Static Analysis Limits**: Current techniques rely on static attack graphs that struggle with scalability.
- **Fidelity Gap**: Abstract models fail to capture evolving attack contexts and system-level side effects.

Core Contributions

- **DRL Framework:** A Deep Reinforcement Learning approach for automated, topology-constrained attack path discovery.
- **Dynamic Integration:** Merges network topology with an intrusion activity graph to learn semantically realistic paths.
- **Efficient Search:** Generates valid attack trajectories without exhaustive graph enumeration.

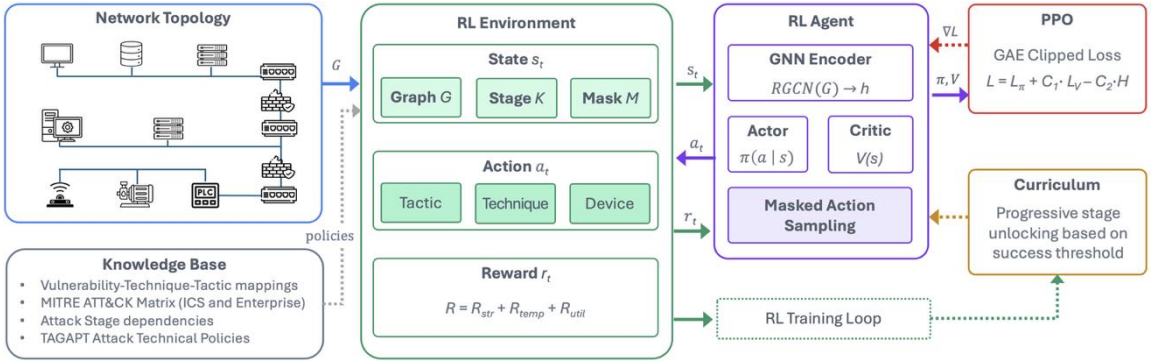
Methodology – Optimise Attack Path Discover within Dynamic IIoT

Threat Model & Autonomous Discovery

- **Environment Dynamics**: The network is characterised by fluctuating topologies, where node availability and communication links change over time.
- **Adversary Goal**: Autonomously synthesize an optimal path from an initial entry point to critical assets, such as PLCs, while minimizing detection
- **Three-Step Execution:** Reconnaissance → Synthesis → Lateral movements.

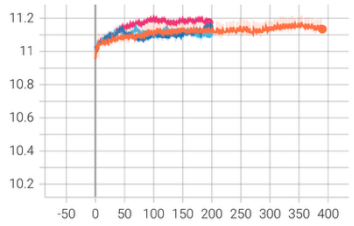
DRL Optimisation via Proximal Policy Optimisation

- **Heterogeneous State Representation:** use fused state $\mathcal{L}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$
- **Constrained Macro-action:** agent selects actions defined by a (*Stage, Technique, Target*) tuple to ensure semantic realism
- **Physical Executability:** We enforce a reachability mask $(v \in C_t \cup \mathcal{N}(C_t) \cup L_t)$ ensures all generated paths strictly adhere to topology and provenance constraints .

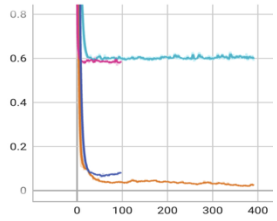


Evaluation

Average Reward in Training



Value Loss in Training



Learning Dynamics & Stability

- **Rapid Adaptation**: The agent achieves near-optimal pruning of invalid actions in first 20% training episodes satisfies strict IIoT constraints.
- **Convergence & Robustness**: Consistent reward stabilisation confirms resilience to high-variance network states.
- **Semantic Realism**: The transition from random exploration to stable path generation indicates the agent has learned to respect industrial protocol dependencies and provenance.

Value Function & Optimization Analysis

- **Critic Convergence**: Rapid decline and stabilisation confirm the network's precision.
- **Predictive Accuracy**: Accurate utility predictions across diverse attack states enable stable policy optimization under volatile IIoT topology constraints.
- **Training Stability:** Successful convergence validates the agent's ability to model long-term rewards despite complex network reachability requirements.

Success Rate	Average Reward	Average Length	Technique Diversity
91.5%	55.8	9.8	5

Performance Evaluation & Path Efficiency

Deterministic Reliability: the trained agent achieved a **100% attack success rate (ASR)**.

Path Optimality: The agent consistently synthesised valid attack trajectories in an **average of 5 steps**.

Masking Effectiveness: topology-constrained action masking successfully prevents logical violations and ensures physical executability.

Generated Attack Path

Attack Stage 1
(Initial Access, Phishing, Workstation)

Attack Stage 2
(Valid Accounts, Privilege Escalation, IT Gateway)
(Credential Dumping, Credential Access, ERP Server)

Attack Stage 3
(C2, Application Layer Protocol, IT Gateway)
(Collection, Data from Local System, DMZ Server)

Attack Stage 4
(Lateral Movement, Remote Services, HMI)
(Impact, Manipulation of Control, PLC)

Conclusion & Future Outlook

- **DRL-Driven Discovery**: Developed a DRL framework that autonomously synthesizes multi-stage attack paths in IIoTs.
- **Semantic Validity:** The agent generates structurally sound trajectories without exhaustive graph enumeration.
- **High Impact Performance:** Evaluation confirmed a 100% success rate in learning coherent, multi-stage attack strategies under realistic system constraints.
- **Future Work:** While current policies are topology-specific, future research will investigate Transfer Learning and Graph Neural Networks (GNNs) to enhance model generalization across unseen, heterogeneous IIoT environments.