Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ANCHOR: Adaptive Nash–Commitment Defense against Learning Attackers in Smart Environments

Mohammed Nurudeen

mohammed.nurudeen@upsamail.edu.gh

University of Professional Studies, Accra

Godfred Yaw Koi-Akrofi

University of Professional Studies, Accra

Hannah Ayaba Tanye

University of Professional Studies, Accra

Yvonne Gyasi Hayford

University of Professional Studies, Accra

Research Article

Keywords: Adaptive cyber defense, Game-theoretic control, Learning attackers, Intelligent environments, Multi-objective optimization

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11599637/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ANCHOR: Adaptive Nash–Commitment Defense against Learning Attackers in Smart Environments

Mohammed Nurudeen[1*], Godfred Yaw Koi-Akrofi[1], Hannah Ayaba Tanye[1], Yvonne Gyasi Hayford[1]

[1*]Department of Information Technology Studies,University of Professional Studies, Accra,Accra,Ghana.

*Corresponding author(s). E-mail(s): mohammed.nurudeen@upsamail.edu.gh;
Contributing authors: godfred.akrofi@upsamail.edu.gh; hannah.tanye@upsamail.edu.gh; 10296436@upsamail.edu.gh;

Abstract

Reliable smart environments require cyber-defense policies that respond to learning attackers without continuously incurring high latency, energy, and reconfiguration costs. This study develops ANCHOR (Adaptive Nash–Commitment Hybrid for Operational Resilience), an interpretable, campaign-stage-aware controller that coordinates low-cost, Static, Nash, and commitment-derived protection anchors. A temporal risk signal determines the requested posture, while asymmetric hysteresis permits immediate escalation and stabilizes subsequent de-escalation. We evaluate ANCHOR in a multi-stage simulation containing stationary attacker learning, cross-regime transfer, drift, and extreme operating conditions. Following disjoint calibration, the frozen policy is tested on 3,300 held-out policy–scenario–seed bundles and a further 600-bundle mechanism-ablation study. ANCHOR reduces effective risk by 34.16% relative to Static and total operational burden by 19.81% relative to Nash and 20.79% relative to full commitment. Paired inference and Pareto analysis establish a non-dominated security–burden operating point among the evaluated policies. Ablations identify Nash anchoring, campaign-stage awareness, and hysteresis as complementary mechanisms: strategic anchoring determines escalation strength, stage awareness guides its timing, and hysteresis reduces reconfiguration oscillation. The results demonstrate how interpretable temporal orchestration can translate strategic cyber protection into an operationally efficient adaptive defense policy.

Keywords: Adaptive cyber defense, Game-theoretic control, Learning attackers, Intelligent environments, Multi-objective optimization

1 Introduction

Reliable smart environments coordinate connected sensing, automated decisions, and physical services while facing attackers that learn from repeated interaction. Their security controllers must respond to changing threats, yet stronger controls also consume latency and energy budgets and may interrupt operation when repeatedly reconfigured. The resulting research question is a joint decision problem: how should a defender allocate strategic protection over an evolving attack campaign while preserving service efficiency?

Moving-target defense and learning-based cyber defense offer adaptive protection, and game-theoretic models characterize strategic incentives

[1–7]. These approaches suggest a useful distinction between the strength of a defensive posture and the time at which that posture is deployed. In a multi-stage intrusion, strong controls may be essential after an attacker gains privileges, while maintaining the same controls throughout earlier interactions imposes unnecessary operating costs. The timing of escalation and de-escalation therefore becomes a scientific mechanism in its own right.

We introduce ANCHOR (Adaptive Nash–Commitment Hybrid for Operational Resilience), a stage-aware defense architecture with interpretable low-cost, Static, Nash, and commitment-derived operating anchors. A temporal risk signal combines the current threat state with recent attacker effectiveness, campaign stage determines admissible postures, and asymmetric hysteresis allows immediate escalation while stabilizing subsequent de-escalation. The attacker learns online and is evaluated under stationary learning, cross-regime transfer, and drift. The architecture is designed to establish how strategic defense and temporal switching control jointly shape security risk and operational burden.

Evaluation is organized around independent development and confirmation evidence. The frozen controller is assessed in 3,300 held-out policy–scenario–seed bundles and a separate 600-bundle mechanism-ablation study. In the primary comparisons, ANCHOR reduces effective risk by 34.16% relative to Static while reducing modeled operational burden by 19.81% relative to Nash and 20.79% relative to full commitment. Its non-dominated operating point is supported by paired comparisons, and ablations distinguish the contributions of strategic anchoring, campaign-stage awareness, and hysteresis.

The paper makes three connected contributions: (i) a finite-grid, stage-aware security policy that selects among strategic anchors; (ii) a precise model of attacker learning, campaign progression, simulated security costs, and defender observations; and (iii) held-out evidence that explains the resulting risk–burden trade-off, including the mechanism-specific effects of the controller's main components. The governing claim is that security capability and the continuous operating cost of that capability can be decoupled through interpretable temporal orchestration.

2 Related Work

2.1 Adaptive and Moving-Target Defense

MTD research addresses the predictability of static defensive configurations by changing attack-surface characteristics over time. Surveys by Cho et al., Sengupta et al., Navas et al., Zheng et al., and Sun et al. identify timing, scope, effectiveness, and reconfiguration cost as recurring design questions [1–5]. These questions are particularly important in IoT systems, where energy and latency constraints can make frequent security adaptation costly [8–11]. Work on smart-home security similarly emphasizes the breadth and heterogeneity of security risks in intelligent environments [12].

ANCHOR addresses a related but different problem: its alternatives are not merely different MTD configurations, but qualitatively different security operating regions ranging from low-cost and Static protection to Nash and commitment-derived postures.

Adaptive IoT policy research also places security control within a dynamic operating environment [13]. The wider complex-systems simulation literature motivates explicit transition models when security decisions interact with evolving adversarial states [14].

2.2 RL-Based Cyber Defense

RL and deep RL support sequential intrusion detection, adaptive response, MTD triggering, and security/energy optimization [15–19]. Game-aware RL has also been applied to security and network-control settings [20–23]. Most such work places the learning algorithm on the defender side. ANCHOR instead keeps the defender structurally interpretable and makes persistent attacker learning a primary stress condition.

Attacker uncertainty in dynamic cyber-system defense has also motivated reinforcement-learning formulations that treat adversarial behavior as a changing part of the control environment [24].

2.3 Learning and Nonstationary Attackers

Adversarial-ML studies show that security mechanisms can be progressively exploited by adversaries

that learn how a target responds [25–27]. Concept-drift studies likewise show that security-relevant distributions need not remain stationary [28, 29]. Transfer-learning work provides a systems basis for asking whether knowledge acquired in one operating context remains useful in another [30, 31]. ANCHOR operationalizes these ideas on the attacker side through stationary learning, warm-start cross-regime transfer, and continuous learning under drift.

2.4 Game-Theoretic Strategic Defense

Game-theoretic approaches represent attacker and defender decisions as strategically interdependent. Nash and Stackelberg formulations capture simultaneous and defender-first interaction, while stochastic and dynamic games extend this reasoning over time [6, 7, 20, 21, 32]. The central distinction in ANCHOR is that these strategic solutions are not treated as whole-system policies that must be maintained continuously; they become *anchors* within a higher-level temporal controller.

2.5 Research Gap

The scientific opportunity is to integrate five established concerns into one controlled comparison: (i) persistent attacker learning with transfer and drift, (ii) campaign-sensitive admissibility of low-cost defenses, (iii) Nash and commitment-derived strategic anchors, (iv) asymmetric de-escalation hysteresis, and (v) joint risk–burden evaluation. Table 1 summarizes the positioning.

3 Problem Formulation and Simulation Method

3.1 Defender observations, actions and objective

At each decision step t, the simulated defender observes $o_t = (\bar{r}_t, \eta_t, \sigma_t, z_t, \theta_{t-1}, k_t)$: the current stored aggregate risk, recent attacker effectiveness, campaign stage, perceived strategic regime, previous deployed control, and elapsed decisions since the last control change. The defender selects a normalized intensity $\theta_t \in [0, 1]$ from the posture choices admissible at that stage. The attacker selects a tactic m_t and intensity x_t, subject to stage validity. The environment then realizes detection, attack success, security-state transitions, operating cost, and the next observation. This temporal sequence is the system plant against which ANCHOR is evaluated.

The co-primary policy outcomes are effective risk and total operational burden. Over the specified scenario and random-seed distribution, a policy $\pi : o_t \mapsto \theta_t$ is assessed through $(\mathbb{E}[R(\pi)], \mathbb{E}[B^{\mathrm{op}}(\pi)])$. A policy is Pareto efficient among the evaluated alternatives when no comparator has both lower risk and lower burden, with at least one strict improvement. This bi-objective formulation motivates a selective controller positioned between low-burden Static defense and stronger, more expensive Nash and commitment postures.

3.2 Smart-environment dynamics and threat model

At decision step t, the defender selects a normalized security-control intensity

$$\theta_t \in [0, 1], \tag{1}$$

where larger values represent stronger monitoring, filtering, validation, hardening, or containment. The attacker then chooses a stage-valid action, after which the environment realizes detection and attack success, updates security risk and campaign stage, and returns attacker reward and defender utility.

The strategic operating condition is summarized by $z_t = (\lambda_t, \delta_t)$, where λ_t is the attack-effort cost coefficient and δ_t is the detection-penalty coefficient. Three principal regimes are the defender-favorable ($DF^\star$), contested ($CT^\star$), and attacker-favorable ($AF^\star$) conditions:

$$\begin{aligned} z^{DF^\star} &= (0.85, 0.70), \\ z^{CT^\star} &= (0.55, 0.50), \\ z^{AF^\star} &= (0.25, 0.10). \end{aligned} \tag{2}$$

Additional noisy and extreme variants are used in robustness evaluation. The principal experiments provide the regime context directly to the defender, while dedicated robustness conditions inject controlled regime misspecification. This separates strategic orchestration from regime estimation and

Table 1 Positioning of ANCHOR Relative to Representative Research Streams

Stream	Representative contribution	Scientific emphasis	ANCHOR contribution
Adaptive/MTD defense	Proactive reconfiguration reduces predictability; timing and overhead matter [1, 3, 9].	Persistent attacker learning and campaign-aware strategic escalation are not jointly central.	Uses low-cost, Static, Nash, and commitment-derived operating anchors and selects them temporally.
RL cyber defense	RL/DRL supports sequential intrusion detection and adaptive control [15, 17].	Learning is predominantly defender-centered.	Keeps the defender interpretable while making online attacker learning a primary stress condition.
Learning and nonstationary attackers	Adversarial learning, drift, and transfer establish time-varying threats [25, 28, 30].	Often focuses on classifier evasion or distribution shift rather than strategic control orchestration.	Evaluates stationary learning, warm-start transfer, and continuous drift in a multi-stage campaign.
Game-theoretic defense	Nash, Stackelberg, and stochastic games characterize strategic interaction [6, 7].	Strong strategic postures can be operationally costly when maintained continuously.	Embeds strategic solutions as escalation anchors inside a stage-aware temporal controller.

permits direct evaluation of ANCHOR's control mechanism.

The attacker occupies one of four campaign stages,

$$\sigma_t \in \{\text{external}, \text{foothold}, \text{privileged}, \text{impact_ready}\}. \tag{3}$$

Its nominal action space contains six tactic families at three intensities $x_t \in \{0.35, 0.65, 1.00\}$ plus abstention, yielding 19 actions. Tactic families are reconnaissance, initial access, spoofing/manipulation, lateral/persistence, collection/exfiltration, and disruption/impact. Only stage-valid actions are admissible. A successful undetected action may advance the campaign by at most one stage, while detection may probabilistically roll it back.

For tactic-specific detectability χ_{m_t}, the simulated event probabilities are

$$p_t^{\text{det}} = \text{sigmoid}\big(-1.25 + 3.2\theta_t + 1.2\delta_t + 0.65\chi_{m_t}x_t\big), \tag{4}$$

$$p_t^{\text{succ}} = \text{sigmoid}\big(0.55 + 1.55x_t + 0.22\,\text{ord}(\sigma_t) - 2.2\theta_t - 0.75\delta_t\big). \tag{5}$$

Here $\text{sigmoid}(u) = 1/(1+e^{-u})$, $\text{ord}(\sigma_t)$ is the campaign-stage index, and $D_t \in \{0,1\}$ is the sampled detection indicator from p_t^{det}. Attack success is sampled separately from p_t^{succ}. After detection, a one-stage rollback has probability $0.12 + 0.25\theta_t$; undetected successful actions may advance the campaign by at most one stage. Thus θ_t affects both the probability of stopping an attack and the subsequent campaign dynamics. The stage-valid action structure and risk profiles are detailed in the supplementary file (Online Resource 1).

Each tactic has a confidentiality–integrity–availability (CIA)-specific impact profile. The environment maintains confidentiality, integrity, and availability risks $\mathbf{r}_t = [r_t^C, r_t^I, r_t^A]^T$, aggregated as

$$\bar{r}_t = 0.30r_t^C + 0.35r_t^I + 0.35r_t^A. \tag{6}$$

The controller uses the current stored risk $\bar{r}_t$ before selecting θ_t. The environment then samples attack detection and success, advances the three CIA risk components to $\mathbf{r}_{t+1}$, and computes the transition-level effective-risk outcome R_t for evaluation. The updated aggregate risk $\bar{r}_{t+1}$ becomes available at the next decision. This ordering matches the executed simulator.

The scalar effective risk used for policy comparison is

$$R_t = \frac{1}{2}(\bar{r}_t + \bar{r}_{t+1}). \tag{7}$$

Operational burden comprises latency, energy, and switching:

$$L_t = 0.035 + 0.20\theta_t + 0.12R_t + 0.06x_t, \tag{8}$$

$$E_t = 0.025 + 0.35\theta_t^2 + 0.04x_t, \tag{9}$$

$$S_t = |\theta_t - \theta_{t-1}|, \tag{10}$$

$$B_t^{\text{op}} = L_t + E_t + S_t. \tag{11}$$

The realized defender utility is

$$U_t^D = 1 - 1.35R_t - 0.32L_t - 0.27E_t - 0.24S_t - 0.10x_t(1 - D_t). \quad (12)$$

Effective risk and total operational burden remain the co-primary interpretable outcomes; defender utility is used as a complementary systems metric.

3.3 ANCHOR Control Architecture

3.3.1 Strategic Operating Anchors

ANCHOR organizes control around three principal anchors and one low-cost external posture. The fixed Static anchor is $\theta^S = 0.42$. The Nash anchor θ_t^N is computed on the finite control grid $\Theta_G = \{0, 0.1, \ldots, 1.0\}$ using stage-valid attacker actions. A pure mutual best response is selected when available; otherwise the minimum-total-regret grid pair is used. The commitment anchor θ_t^C anticipates the attacker's utility-maximizing response to each candidate control and selects the defender-favorable candidate. The safety posture is

$$\theta_t^{\text{safe}} = \max\{\theta^S, \theta_t^N, \theta_t^C\}. \quad (13)$$

The low-cost external posture θ_t^L is bounded below Static and receives only a small regime-family adjustment.

3.3.2 Temporal Risk Signal and Stage-Aware Selection

ANCHOR forms the temporal risk signal

$$q_t = \text{clip}(\bar{r}_t + \omega\eta_t, 0, 1), \quad (14)$$

where η_t is recent attacker effectiveness and ω its weight. External-stage decisions use two thresholds: below the first threshold the controller selects the low-cost posture, between the two thresholds it uses Static, and above the second it moves to the Nash region. At foothold, the controller selects Static below the foothold threshold and the Nash region otherwise. At privileged stage, low-cost operation is disabled and the admissible target is restricted to Static or Nash-level operation. At impact-ready stage, or whenever $\bar{r}_t$ exceeds the emergency threshold, the safety posture is invoked. These rules produce the requested target $\tilde{\theta}_t$; the asymmetric hysteresis mechanism in the next subsection determines whether that target is deployed immediately or held temporarily.

The frozen R11 values are: base low-control 0.35, external Static threshold 0.20, external Nash threshold 0.27, foothold Nash threshold 0.19, effectiveness weight 0.18, emergency-risk threshold 0.36, cooldown 5 steps, and minimum de-escalation change 0.05. For the algorithm below, τ_1, τ_2, and τ_3 denote the context-adjusted external-Static, external-Nash, and foothold-Nash thresholds, respectively, and $\rho_{\text{emg}} = 0.36$ denotes the emergency-risk threshold. Regime-family offsets are retained as a secondary contextual adjustment; D2 quantifies their independent contribution.

3.3.3 Asymmetric Hysteresis

If the requested target $\tilde{\theta}_t$ exceeds the deployed control, escalation is immediate. If it is lower, de-escalation is accepted only after the cooldown has expired and the requested change is sufficiently large:

$$\theta_t = \begin{cases} \theta_{t-1}, & k_t < K \text{ or } |\tilde{\theta}_t - \theta_{t-1}| < \epsilon_\theta, \\ \tilde{\theta}_t, & \text{otherwise.} \end{cases} \quad (15)$$

Here k_t counts elapsed decisions since the most recent change in deployed control, with $K = 5$ and $\epsilon_\theta = 0.05$. The executed update is $k_{t+1} = 0$ when $\theta_t \neq \theta_{t-1}$, and $k_{t+1} = k_t + 1$ otherwise. Initially $k_0 = 999$, encoding that no recent switch constrains the first decision; only the comparison $k_t < K$ affects the policy. This asymmetry never delays security escalation. Fig. 1 summarizes the resulting closed-loop orchestration: campaign stage, effective risk, recent attacker effectiveness, and regime context determine the requested operating posture; escalation is immediate, whereas downward transitions are filtered by the cooldown and minimum-change requirements before the deployed control is updated.

3.3.4 Algorithmic Realization and Complexity

Algorithm 1 gives the per-decision realization of ANCHOR. The strategic anchors are computed only over the finite defender grid and the attacker actions admissible at the current campaign stage;

Table 2 Principal notation and abbreviations used in the manuscript.

Notation	Meaning	Abbreviation	Meaning
θ_t	Deployed defender control intensity	ANCHOR	Adaptive Nash–Commitment Hybrid for Operational Resilience
$\theta^S, \theta_t^N, \theta_t^C, \theta_t^L, \theta_t^{\text{safe}}$	Static, Nash, commitment, low-cost, and safety postures	MTD / RL	Moving-target defense / reinforcement learning
$\bar{r}_t, R_t, q_t, \eta_t$	Pre-decision aggregate risk, transition effective risk, temporal signal, and attacker effectiveness	CIA / IoT	Confidentiality–integrity–availability / Internet of Things
$\sigma_t,\ z_t = (\lambda_t, \delta_t)$	Campaign stage and regime context	$DF^\star/CT^\star/AF^\star$	Defender-favorable / contested / attacker-favorable regimes
$L_t, E_t, S_t, B_t^{\text{op}}, U_t^D$	Latency, energy, switching, total burden, and defender utility	D1 / D2	Held-out confirmation / mechanism-ablation campaigns
x_t, D_t, g_t	Attack intensity, sampled detection flag, attacker reward	–	–

Table 3 Frozen ANCHOR-R11 Controller Parameters

Parameter	Value
Base low-control posture	0.35
External Static threshold	0.20
External Nash threshold	0.27
Foothold Nash threshold	0.19
Recent-effectiveness weight	0.18
Emergency-risk threshold	0.36
De-escalation cooldown	5 steps
Minimum de-escalation change	0.05
Static anchor	0.42

the resulting requested target is then filtered by the asymmetric hysteresis rule.

Let $G = |\Theta_G|$ and $A_t = |\mathcal{A}_t|$. Under the direct finite-grid implementation, the Nash and commitment anchors require scanning the defender–attacker utility pairs, giving $O(GA_t)$ time per decision; stage selection, risk-signal formation, and hysteresis are $O(1)$. Hence an episode of T decisions has controller cost $O(TGA_{\max})$. If the payoff grid is materialized, the working memory is $O(GA_t)$; a streaming implementation can reduce this to $O(G + A_t)$. In the frozen experiments, $G = 11$ and $A_t \le 19$, so at most 209 defender–attacker grid pairs are examined at any decision step. This complexity excludes the attacker-learning update used in the adversarial evaluation.

3.4 Learning attacker and nonstationarity

The principal attacker is a tabular softmax Q-learner whose state is

$$s_t = (b_t^\theta, b_t^\eta, b_t^p, \sigma_t, \mu_{t-1}), \tag{16}$$

where the first three components are three-level buckets for current defender control, recent attacker effectiveness, and recent defensive pressure; σ_t is campaign stage and μ_{t-1} is the previous tactic code. The nominal state count is $3 \times 3 \times 3 \times 4 \times 7 = 756$.

The three-level state buckets use control boundaries 0.34 and 0.67, recent-effectiveness boundaries 0.20 and 0.55, and defensive-pressure boundaries 0.25 and 0.60; both recent signals use five-interaction windows. The attacker reward is realized after the environment transition; the supplementary file details the learning protocol.

Writing b_t^A for the realized tactic benefit supplied by the simulated environment and κ_{m_t} for its tactic-specific effort multiplier, the documented reward structure has the form

$$\begin{aligned} g_t = \text{clip}\big(&b_t^A - \lambda_t \kappa_{m_t} x_t^2 \\ &- \delta_t D_t (0.25 + 0.75 x_t), -1, 1\big). \end{aligned} \tag{17}$$

The realized benefit is $b_t^A = I_{m_t} v_{m_t} x_t (0.35 + 0.65 Y_t)(1 - 0.20 D_t)(0.55 + 0.45 R_t)$, where I_{m_t} is the CIA-weighted tactic impact, v_{m_t} the attack-surface multiplier, and $Y_t \in \{0, 1\}$ the sampled attack-success indicator. The effort multiplier κ_{m_t} is tactic-specific. Thus g_t is computed from the completed transition, whereas the defender's

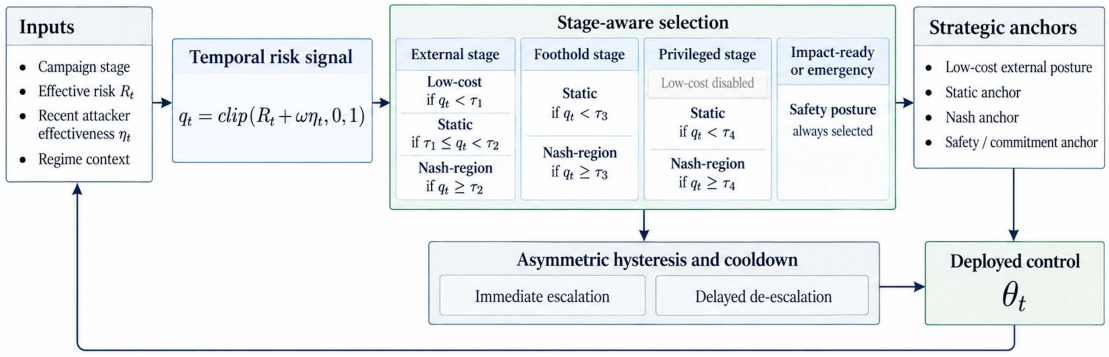


Fig. 1 Stage-aware temporal orchestration of ANCHOR. Strategic operating anchors feed the stage-aware selector, which combines campaign stage with the temporal risk signal; asymmetric hysteresis and cooldown then determine the deployed control, and recent outcomes feed the next decision.

strategic-anchor calculations use anticipated attack and defender utilities before the action is realized. These quantities are distinct by design. The complete ex-ante payoff model is specified in the supplementary file.

The primary Q-learning update is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big[g_t + \gamma \max_{a' \in \mathcal{A}(\sigma_{t+1})} Q(s_{t+1}, a') - Q(s_t, a_t) \Big], \tag{18}$$

with $\alpha = 0.15$ and $\gamma = 0.90$. Stage-valid actions are sampled through a Boltzmann policy with temperature $\tau = 0.25$ and exploration floor 0.05. A five-interaction rolling memory defines recent effectiveness and defensive pressure. Under drift, per-update forgetting uses $\phi = 0.03$; stationary and transfer conditions use $\phi = 0$.

Three forms of history dependence are evaluated. In stationary learning, the attacker improves within a fixed regime. In transfer, it trains for 20 episodes in a source regime and warm-starts a 10-episode target phase, covering all six directed transitions among $DF^\star$, $CT^\star$, and $AF^\star$. Under drift, it learns continuously over a 600-step trajectory with abrupt, gradual, or recurring regime change. SARSA and Double Q-learning are used as robustness checks.

3.5 Simulation Development, Confirmation, and Statistical Analysis

3.5.1 Development–Confirmation Separation

An initial exploratory campaign established the design need for a controller that couples adaptation with explicit security–burden management. The resulting selective controller is retained as a *developmental selective baseline* that documents the progression toward ANCHOR's stage-aware orchestration.

ANCHOR development used three disjoint 10-seed rounds. D0 (seeds 1000–1009) evaluated profiles P01–P12 under a stringent pointwise security constraint, and D0R2 (1100–1109) evaluated Q01–Q12 using an interpolated Nash–Static risk budget. These rounds established strong security but concentrated operation near the higher-burden strategic region, motivating the D0R3 structural redesign. D0R3 (1200–1209) introduced temporal stage-aware orchestration and produced R11 as the lowest-burden profile among those satisfying the pre-specified $\geq 80\%$ paired-security criterion. A same-seed development frontier then placed R11 on a non-dominated operating point between Static and Nash/full commitment.

Before confirmation, R11, the controller source, the D1/D2 manifests, protocol, and D1 analysis code were hash-locked and implementation equivalence was verified. The exploratory foundation and D0–D0R3 rounds define the development phase,

Algorithm 1 ANCHOR control update at decision step t

Require: Current aggregate risk $\bar{r}_t$, recent attacker effectiveness η_t, stage σ_t, perceived regime z_t, previous control θ_{t-1}, cooldown state k_t

Form the stage-valid action set $\mathcal{A}_t = \mathcal{A}(\sigma_t)$ and apply the frozen regime-family modulation to the R11 low posture and thresholds.
Evaluate defender/attacker utilities on $\Theta_G \times \mathcal{A}_t$.
Obtain θ_t^N from a pure mutual best response when one exists; otherwise choose the minimum-total-regret grid pair.
For each $\theta \in \Theta_G$, let $a^*(\theta) = \arg\max_{a \in \mathcal{A}_t} U_t^A(\theta, a)$; set $\theta_t^C = \arg\max_{\theta \in \Theta_G} U_t^D(\theta, a^*(\theta))$.
$\theta_t^{\text{safe}} \leftarrow \max\{\theta^S, \theta_t^N, \theta_t^C\}$ and $q_t \leftarrow \text{clip}(\bar{r}_t + \omega\eta_t, 0, 1)$.
if $\sigma_t =$ impact_ready or $\bar{r}_t \geq \rho_{\text{emg}}$ **then**
    $\tilde{\theta}_t \leftarrow \theta_t^{\text{safe}}$
else if $\sigma_t =$ external **then**
    $\tilde{\theta}_t \leftarrow \theta_t^L$ if $q_t < \tau_1$; θ^S if $\tau_1 \leq q_t < \tau_2$; otherwise $\max\{\theta^S, \theta_t^N\}$
else if $\sigma_t =$ foothold **then**
    $\tilde{\theta}_t \leftarrow \theta^S$ if $q_t < \tau_3$; otherwise $\max\{\theta^S, \theta_t^N\}$
else ▷ privileged stage
    $\tilde{\theta}_t \leftarrow \max\{\theta^S, \theta_t^N\}$
end if
if $\tilde{\theta}_t > \theta_{t-1}$ **then**
    $\theta_t \leftarrow \tilde{\theta}_t$ ▷ immediate escalation
else if $k_t < K$ or $|\tilde{\theta}_t - \theta_{t-1}| < \epsilon_\theta$ **then**
    $\theta_t \leftarrow \theta_{t-1}$ ▷ hold during cooldown / small requested decrease
else
    $\theta_t \leftarrow \tilde{\theta}_t$
end if
if $\theta_t \neq \theta_{t-1}$ **then**
    $k_{t+1} \leftarrow 0$
else
    $k_{t+1} \leftarrow k_t + 1$
end if
return (θ_t, k_{t+1})

while D1 and D2 form fully disjoint held-out confirmation campaigns. The frozen configuration was carried unchanged into confirmation, preserving a clean separation between design and evaluation.

3.5.2 D1 Held-Out Confirmation

D1 uses seeds 2000–2029 and five policies: ANCHOR, the developmental selective baseline, Static, Nash, and full commitment. It contains 450 stationary bundles, 900 transfer bundles, 1,350 drift bundles, and 600 extreme-condition bundles, for a total of 3,300.

The three pre-specified directional hypotheses are

$$H_1: \quad E[R^A - R^S] < 0, \tag{19}$$

$$H_2: \quad E[B^A - B^N] < 0, \tag{20}$$

$$H_3: \quad E[B^A - B^F] < 0, \tag{21}$$

where A, S, N, F denote ANCHOR, Static, Nash, and full commitment.

The inferential unit is the seed, not the timestep. Scenario values are first averaged within each of the four scenario classes, and those class means receive equal weight, producing one block-balanced value per policy and seed. Policy comparisons use common random numbers. Each primary contrast reports the paired mean difference, a 10,000-resample paired bootstrap 95% confidence interval, and a one-sided Wilcoxon signed-rank test. Holm correction controls family-wise error across H_1–H_3 at $\alpha = 0.05$. Pareto efficiency is assessed jointly on block-balanced effective risk and total operational burden.

3.5.3 D2 Held-Out Mechanism Ablation

D2 uses seeds 3000–3019 and 600 bundles. Five variants are crossed with six conditions: full ANCHOR, no hysteresis, no Nash anchor, no regime adjustment, and no stage awareness. Each ablation is paired with full ANCHOR using the same seed and condition. Four primary mechanism metrics—risk, total burden, defender utility, and switching burden—are tested with two-sided Wilcoxon tests, paired bootstrap intervals, and rank-biserial effects; Holm correction is applied across all 16 mechanism tests. The supplementary file (Online Resource 1) provides the complete implementation specification, development and pre-confirmatory lock records, and expanded D1/D2 numerical and robustness results.

4 Results

4.1 Held-Out Security–Burden Frontier

Table 4 reports the block-balanced D1 means. Full commitment and Nash provide the lowest risk but

the largest burden. Static provides the lowest burden but substantially higher risk. ANCHOR lies between these extremes.

Relative to Static, ANCHOR reduces effective risk by 34.16%. Relative to Nash and full commitment, it reduces total operational burden by 19.81% and 20.79%, respectively. These simultaneous gains move ANCHOR beyond the developmental selective baseline and onto the Pareto-optimal security–burden frontier. Fig. 2 shows the corresponding held-out operating geometry.

4.2 Confirmatory Hypotheses

The mean ANCHOR–Static risk difference is -0.040589 with 95% paired-bootstrap CI $[-0.041391, -0.039844]$. The ANCHOR–Nash burden difference is -0.089647 with CI $[-0.091073, -0.088248]$, and the ANCHOR–full-commitment burden difference is -0.095225 with CI $[-0.096638, -0.093842]$. The Wilcoxon statistic is $W = 0$ for all three comparisons. All remain significant after Holm correction ($p_{\text{Holm}} < 0.001$). Thus the intended intermediate operating position is supported both geometrically and inferentially.

4.3 Operational Interpretation

The operational-burden comparison reveals why temporal selection produces a non-dominated policy. Nash and full commitment maintain stronger defense over more interactions, which compounds the model's convex θ_t^2 energy term and the level-dependent latency term. ANCHOR reduces those sustained costs by using strategic protection selectively. Reconfiguration introduces additional $S_t = |\theta_t - \theta_{t-1}|$, but the latency and energy savings outweigh the switching increment in the held-out aggregates (Fig. 3). The policy therefore changes *when* a costly defense region is occupied, rather than relying on a weaker equilibrium throughout the campaign.

4.4 D2 Mechanism Attribution

Removing the Nash anchor causes the largest security deterioration: effective risk rises from 0.077396 to 0.111223 (paired delta +0.033827). Removing stage awareness raises risk by 0.016248 and lowers defender utility. Removing hysteresis raises risk by 0.021513 and switching burden by 0.028733, corresponding to a 76.2% increase in switching relative to full ANCHOR. Without the de-escalation waiting rule, downward and subsequent upward movements occur more readily around changing requested targets; the additional posture reversals accumulate in the absolute-control-difference switching metric. The same ablation slightly lowers total burden even as switching rises, because the other burden terms also change. This distinction shows why transition stability, effective risk, and aggregate operating cost must be read together. All of these primary contrasts remain significant after the 16-test Holm correction.

The ablation evidence concentrates the performance contribution in Nash anchoring, campaign-stage awareness, and hysteresis. Regime-threshold adjustment has a comparatively small aggregate effect: its burden change is close to zero ($p_{\text{Holm}} = 0.6742$), indicating that the core ANCHOR gains arise primarily from the three structural mechanisms. Fig. 4 summarizes these held-out mechanism deltas relative to full ANCHOR.

Viewed together, the ablations reveal a layered control architecture rather than redundant components. The Nash anchor determines the strength of strategic escalation, stage awareness determines when that strength becomes admissible, and hysteresis governs how safely the controller returns to lower-cost operation. Their complementary roles explain why full ANCHOR preserves a favorable risk–burden position: strategic strength is available when needed without sustaining its full operational cost throughout the trajectory. This mechanism-level evidence directly connects the D1 frontier to the controller design.

5 Discussion

The principal scientific result is that a stage-aware controller can occupy a non-dominated security–operations region by separating strategic protection strength from the duration and frequency of its deployment. This operating principle is relevant to smart environments, where security controls share resources with latency-sensitive services and constrained energy budgets. ANCHOR therefore converts strong strategic protection from a continuously paid cost into a selectively invoked resource, directly aligning security intensity with operational need. Recent systems-security work on

Table 4 Block-Balanced Held-Out D1 Results

Policy	Risk	Total burden	Utility	Latency	Energy	Switching	Pareto
Full commitment	0.019136	0.458142	0.833294	0.212146	0.242387	0.003609	Yes
Nash	0.019665	0.452563	0.834007	0.210772	0.238080	0.003712	Yes
ANCHOR	0.078240	0.362917	0.775088	0.173583	0.148156	0.041177	Yes
Static	0.118829	0.271763	0.741980	0.163268	0.106745	0.001750	Yes
Developmental selective baseline	0.132285	0.291755	0.718194	0.161792	0.103484	0.026480	No

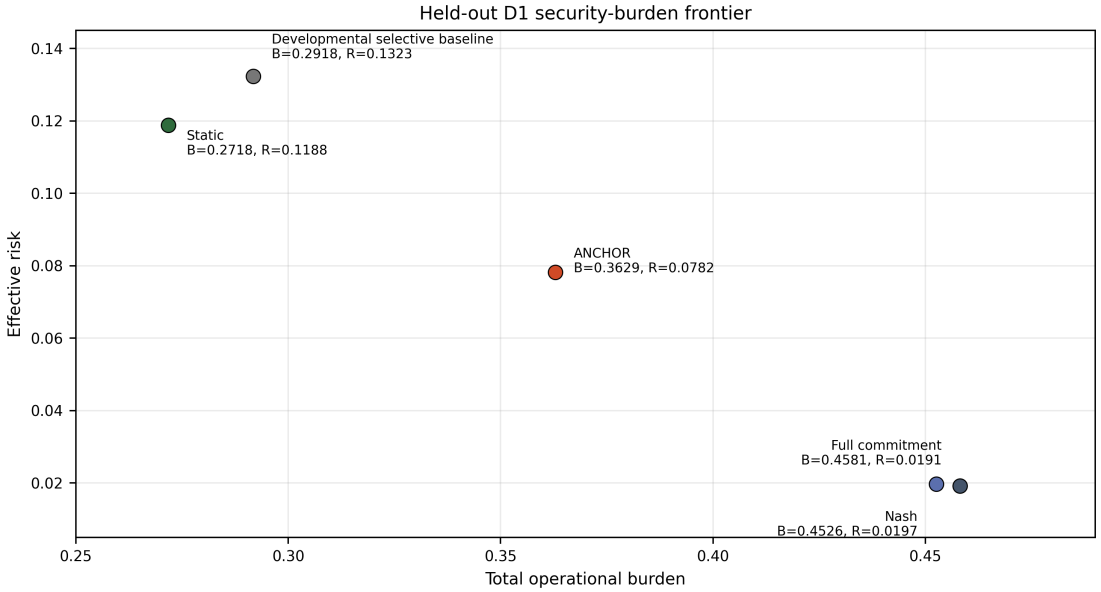


Fig. 2 Held-out D1 security–burden frontier. Lower risk and lower burden are preferred.

Table 5 Pre-Specified D1 Confirmatory Tests

Hypothesis	Comparison	Mean difference	95% paired-bootstrap CI	Decision
H_1	ANCHOR risk − Static risk	−0.040589	$[-0.041391, -0.039844]$	Supported
H_2	ANCHOR burden − Nash burden	−0.089647	$[-0.091073, -0.088248]$	Supported
H_3	ANCHOR burden − Full-commitment burden	−0.095225	$[-0.096638, -0.093842]$	Supported

harmonizing security and performance reinforces the relevance of this joint perspective [33].

ANCHOR's advantage is produced by the structure of its orchestration. Low-cost operation is reserved for campaign states that permit it; campaign progression progressively removes economical choices; the Nash anchor supplies a strategically strong escalation destination; the commitment-derived anchor provides an emergency safety posture; and asymmetric hysteresis stabilizes de-escalation. Together, these mechanisms create the measured security–burden balance rather than relying on frequent switching alone.

D2 clarifies the division of labor among the core mechanisms. The Nash anchor is the largest individual strategic-security contributor. Stage awareness adds progression information beyond the instantaneous risk score, allowing ANCHOR to distinguish externally similar risk levels that carry different campaign implications. Hysteresis provides security stability by suppressing premature downward transitions; its removal increases both switching and effective risk. These effects explain why the combined architecture sustains the observed Pareto-optimal operating point.

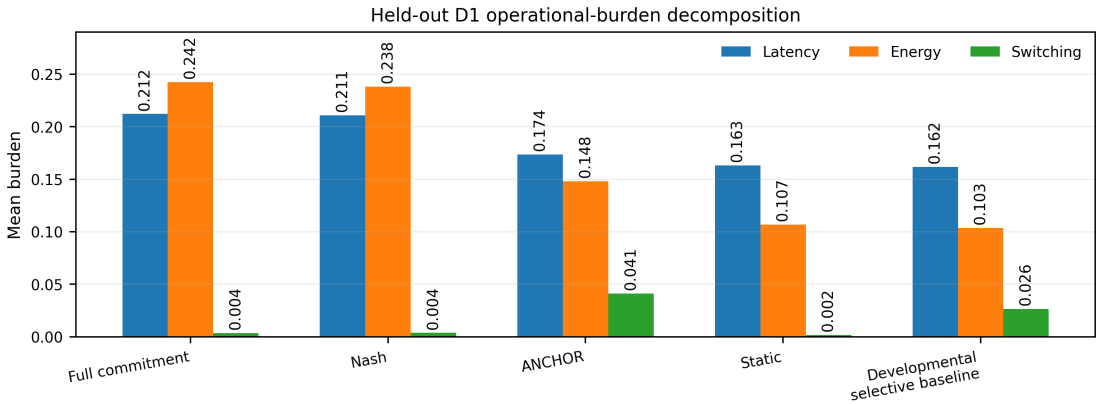


Fig. 3 Held-out D1 operational-burden decomposition. ANCHOR accepts greater switching cost in exchange for lower sustained latency and energy burden than Nash or full commitment.

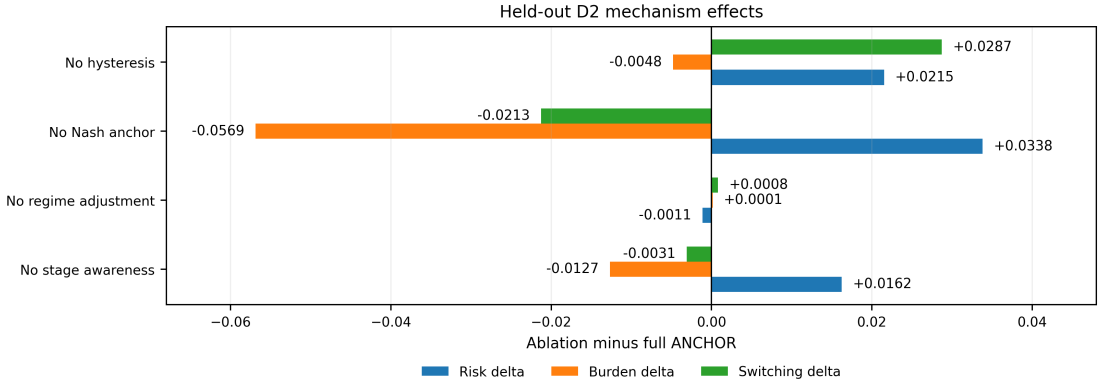


Fig. 4 Held-out D2 mechanism effects. Values are ablation minus full ANCHOR; positive risk indicates worse security.

The regime-adjustment ablation further sharpens the architecture. Its comparatively small aggregate contribution shows that the dominant gains are carried by Nash anchoring, campaign-stage awareness, and hysteresis. This finding supports a streamlined deployment path in which contextual offsets can be simplified or learned while preserving the core control structure.

Methodologically, the development–confirmation separation strengthens the evidence for ANCHOR. D0 and D0R2 established the high-security region of the design space, while D0R3 introduced the stage-aware structural redesign that moved R11 toward substantially lower burden. Freezing R11 before D1/D2 then converted this development trajectory into a genuinely held-out test of the final controller. The resulting D1 and D2 evidence therefore evaluates a fixed architecture under independent seeds rather than another design iteration.

6 Research Implications and Future Extensions

The evaluation uses a controlled smart-environment simulation to isolate the joint effects of effective risk, latency, energy, and reconfiguration burden under reproducible attacker-learning conditions. This setting supports precise mechanism attribution across stationary learning, transfer, drift, extreme conditions, and 3,900 held-out D1/D2 bundles. Hardware-in-the-loop and real-system deployments form the natural next validation stage for calibrating these quantities to platform-specific units.

The attacker design combines an interpretable tabular softmax Q-learner with SARSA and Double-Q robustness checks across six tactic families, three intensities, and four campaign stages. This provides a transparent multi-stage adaptive adversary while leaving a clear path to deeper, recurrent, model-based, and multi-agent attacker models and richer enterprise attack graphs.

The finite strategic grid keeps Nash and commitment anchors computationally tractable and auditable, while the regime context enables direct study of strategic orchestration and controlled misspecification. Future deployment-oriented extensions can incorporate finer or continuous strategic solvers and learned regime estimation. The paired seed design and common random numbers provide high-precision confirmation in the present study, and additional cross-platform Monte Carlo campaigns can further broaden external validation across smart-environment verticals.

7 Conclusions

This paper introduced ANCHOR, an Adaptive Nash–Commitment Hybrid for Operational Resilience designed for smart environments facing learning, multi-stage attackers. ANCHOR organizes defense around low-cost, Static, Nash, and commitment-derived safety anchors and moves among them according to effective risk, recent attacker effectiveness, campaign progression, and asymmetric de-escalation hysteresis.

In the 3,300-bundle held-out D1 campaign, ANCHOR reduced effective risk by 34.16% relative to Static while reducing total operational burden by 19.81% relative to Nash and 20.79% relative to full commitment. All three pre-specified paired comparisons remained significant after Holm correction, establishing ANCHOR as a Pareto-optimal operating point within the evaluated policy set. The independent 600-bundle D2 study further identified Nash anchoring, campaign-stage awareness, and hysteresis as the dominant structural mechanisms behind this balance.

ANCHOR therefore provides a multi-objective defense architecture that balances security improvement with operational efficiency along the Pareto frontier. Strong strategic protection remains continuously available as an escalation resource, while stage-aware selection and hysteresis avoid paying its full latency and energy cost throughout benign or lower-risk periods. This combination makes ANCHOR a practical operating optimum in the Pareto sense within the evaluated design space. Future work will extend the architecture through hardware-in-the-loop and real smart-environment validation, empirically calibrated operational costs, richer multi-agent learning attackers, learned regime estimation, and finer strategic solution methods.

Statements and Declarations

Funding: No specific external project grant supported this study.

Competing interests: The authors declare that they have no competing interests.

Data and code availability: Supporting experiment records, configurations, manifests and analysis scripts are archived by the authors and are planned for public availability upon publication.

References

[1] Cho, J.-H., Yang, S.J., Khalil, I., Baek, J., Al-Hamadi, Y., Alazab, M., Hong, C.S.: Toward proactive, adaptive defense: A survey on moving target defense. IEEE Communications Surveys & Tutorials **22**(1), 709–745 (2020) https://doi.org/10.1109/COMST.2019.2963791

[2] Sengupta, S., Chowdhary, A., Sabur, A., Alshamrani, A., Huang, D., Kambhampati, S.: A survey of moving target defenses for network security. IEEE Communications Surveys & Tutorials **22**(3), 1909–1941 (2020) https://doi.org/10.1109/COMST.2020.2982955

[3] Navas, R.E., Alcaraz, J.M., Lopez, J.: MTD, where art thou? a systematic review of moving target defense techniques for IoT. IEEE Internet of Things Journal **8**(10), 7818–7832 (2021) https://doi.org/10.1109/JIOT.2020.3040358

[4] Zheng, Y., Li, Z., Xu, X., Zhao, Q.: Dynamic defenses in cyber security: Techniques, methods and challenges. Digital Communications and Networks **8**(4), 422–435 (2022) https://doi.org/10.1016/j.dcan.2021.07.006

[5] Sun, R., Zhu, Y., Fei, J., Chen, X.: A survey on moving target defense: Intelligently affordable,

optimized and self-adaptive. Applied Sciences **13**(9), 5367 (2023) https://doi.org/10.3390/app13095367

[6] Kiennert, C., Ismail, Z., Debar, H., Leneutre, J.: A survey on game-theoretic approaches for intrusion detection and response optimization. ACM Computing Surveys **51**(5) (2019) https://doi.org/10.1145/3232848

[7] Temghart, A.A., Marwan, M., Baslam, M.: Stackelberg security game for optimizing cybersecurity decisions in cloud computing. Security and Communication Networks **2023**, 2811038 (2023) https://doi.org/10.1155/2023/2811038

[8] Zhou, Y., Cheng, G., Zhao, Y., Chen, Z., Jiang, S.: Toward proactive and efficient DDoS mitigation in IIoT systems: A moving target defense approach. IEEE Transactions on Industrial Informatics **18**(4), 2734–2744 (2022) https://doi.org/10.1109/TII.2021.3090719

[9] Osei, A.B., Yeginati, S.R., Mtawa, Y.A., Halabi, T.: Optimized moving target defense against DDoS attacks in IoT networks: When to adapt? In: Proceedings of the IEEE Global Communications Conference (GLOBECOM), pp. 2782–2787 (2022)

[10] Rehman, Z., Gondal, I., Ge, M., Dong, H., Gregory, M., Tari, Z.: Proactive defense mechanism: Enhancing IoT security through diversity-based moving target defense and cyber deception. Computers & Security **139**, 103685 (2024) https://doi.org/10.1016/j.cose.2023.103685

[11] Swati, Roy, S., Singh, J., Mathew, J.: Securing IIoT systems against DDoS attacks with adaptive moving target defense strategies. Scientific Reports **15**, 9558 (2025) https://doi.org/10.1038/s41598-025-93138-7

[12] Yuan, B., Wan, J., Wu, Y.-H., Zou, D.-Q., Jin, H.: On the security of smart home systems: A survey. Journal of Computer Science and Technology **38**(2), 228–247 (2023) https://doi.org/10.1007/s11390-023-2488-3

[13] Halgamuge, M.N., Niyato, D.: Adaptive edge security framework for dynamic IoT security policies in diverse environments. Computers & Security **148**, 104128 (2025) https://doi.org/10.1016/j.cose.2024.104128

[14] Tang, Y., Li, L., Liu, X.: State-of-the-art development of complex systems and their simulation methods. Complex System Modeling and Simulation **1**(4), 271–290 (2021) https://doi.org/10.23919/CSMS.2021.0025

[15] Nguyen, T.T., Reddi, V.J.: Deep reinforcement learning for cyber security. IEEE Transactions on Neural Networks and Learning Systems **34**(8), 3779–3795 (2023) https://doi.org/10.1109/TNNLS.2021.3121870

[16] Palmer, G., Parry, C., Harrold, D.J.B., Willis, C.: Deep Reinforcement Learning for Autonomous Cyber Operations: A Survey (2023). https://doi.org/10.48550/arXiv.2310.07745 . https://arxiv.org/abs/2310.07745

[17] Jamshidi, S., Amirnia, A., Nikanjam, A., Nafi, K.W., Khomh, F., Keivanpour, S.: Self-adaptive cyber defense for sustainable IoT: A DRL-based IDS optimizing security and energy efficiency. Journal of Network and Computer Applications **239**, 104176 (2025) https://doi.org/10.1016/j.jnca.2025.104176

[18] Tharewal, S., Ashfaque, M.W., Banu, S.S., Uma, P., Hassen, S.M., Shabaz, M.: Intrusion detection system for industrial internet of things based on deep reinforcement learning. Wireless Communications and Mobile Computing **2022**, 9023719 (2022) https://doi.org/10.1155/2022/9023719

[19] Kim, S., Yoon, S., Cho, J.-H., Kim, D.S., Moore, T.J., Free-Nelson, F., Lim, H.: DIVERGENCE: Deep reinforcement learning-based adaptive traffic inspection and moving target defense countermeasure framework. IEEE Transactions on Network and Service Management **19**(4), 4834–4846 (2022) https://doi.org/10.1109/TNSM.2021.3139928

[20] Benaddi, H., Ibrahimi, K., Benslimane, A., Jouhari, M., Qadir, J.: Robust enhancement of intrusion detection systems using deep

reinforcement learning and stochastic game. IEEE Transactions on Vehicular Technology **71** (2022) https://doi.org/10.1109/TVT.2022.3186834

[21] Liu, X., Zhang, H., Dong, S., Zhang, Y.: Network defense decision-making based on a stochastic game system and a deep recurrent Q-network. Computers & Security **111**, 102480 (2021) https://doi.org/10.1016/j.cose.2021.102480

[22] Bello, Y., Hussein, A.R.: Dynamic policy decision/enforcement security zoning through stochastic games and meta learning. IEEE Transactions on Network and Service Management **22**(1), 807–821 (2025) https://doi.org/10.1109/TNSM.2024.3481662

[23] Ding, T., Liu, N., Yan, Z.-M., Liu, L., Cui, L.-Z.: An efficient reinforcement learning game framework for uav-enabled wireless sensor network data collection. Journal of Computer Science and Technology **37**(6), 1356–1368 (2022) https://doi.org/10.1007/s11390-022-2419-8

[24] Dutta, A., Chatterjee, S., Bhattacharya, A., Halappanavar, M.: Deep Reinforcement Learning for Cyber System Defense under Dynamic Adversarial Uncertainties (2023). https://doi.org/10.48550/arXiv.2302.01595 . https://arxiv.org/abs/2302.01595

[25] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., Poor, H.V.: Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey. IEEE Communications Surveys & Tutorials **25**(4), 2245–2298 (2023) https://doi.org/10.1109/COMST.2023.3319492

[26] McCarthy, A., Ghadafi, E., Andriotis, P., Legg, P.: Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. Journal of Cybersecurity and Privacy **2**(1), 154–190 (2022) https://doi.org/10.3390/jcp2010010

[27] Zhang, C., Costa-Pérez, X., Patras, P.: Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. IEEE/ACM Transactions on Networking **30**(3), 1294–1311 (2022) https://doi.org/10.1109/TNET.2021.3137084

[28] Shyaa, S., Raju, C.K., Prasad, D.V.S.N.V.: Evolving cybersecurity frontiers: Concept drift and feature dynamics aware machine learning in intrusion detection systems. Engineering Applications of Artificial Intelligence **137**, 109143 (2024) https://doi.org/10.1016/j.engappai.2024.109143

[29] Moriano, P.A., Hespeler, S.C., Li, M., Mahbub, M.: Adaptive anomaly detection for identifying attacks in cyber-physical systems: A systematic literature review. Artificial Intelligence Review **58**(9), 283 (2025) https://doi.org/10.1007/s10462-025-11292-w

[30] Cody, T., Beling, P.A.: A systems theory of transfer learning. IEEE Systems Journal **17**(1), 26–37 (2023) https://doi.org/10.1109/JSYST.2022.3224650

[31] Cody, T., Adams, S.C., Beling, P.A.: Empirically measuring transfer distance for system design and operation. IEEE Systems Journal **16**(3), 4962–4973 (2022)

[32] Liu, X., Shi, L.: A dynamic game model for assessing risk of coordinated physical-cyber attacks in an AC/DC hybrid transmission system. Frontiers in Energy Research **10**, 1082442 (2023) https://doi.org/10.3389/fenrg.2022.1082442

[33] Li, W.-T., Wang, Z.-X., Gu, J.-Y., Xia, Y.-B., Zang, B.-Y.: Harmonizing security and performance in microkernel file servers. Journal of Computer Science and Technology **40**(2), 464–481 (2025) https://doi.org/10.1007/s11390-025-3762-3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf