Research on the Evolution of Oil and Gas Pipeline Network Attack and Defense Based on Reinforcement Learning and Game Theory

Lidong Jia[1], Fei Song[1], Weichun Hei[1], Jian Wang[1], Yinggang Xie[2]*

[1] PipeChina North Pipeline Company, China
[2] Beijing Information Science and Technology University, China
jiald@pipechina.com.cn, songfei@pipechina.com.cn, heiwc@pipechina.com.cn, wangjian09@pipechina.com.cn, xieyinggang@bistu.edu.cn

Abstract

As a critical national infrastructure, the oil and gas pipeline network faces more complex and dynamically evolving cybersecurity risks during the process of informatization and intelligentization. Attackers usually have the advantage of information asymmetry and can flexibly adjust the attack path in the multi-stage chain structure, which poses higher requirements for the adaptability of the pipeline network defense system. The article proposes an intelligent defense modeling method that integrates game modeling, deep reinforcement learning, and strategy evolution mechanisms. A dual-agent confrontation structure is constructed to simulate real attack and defense behaviors, enabling the self-optimization of defense strategies in a dynamic environment. The model introduces the replicator dynamic equation in evolutionary game theory to enhance the stability of strategy updating and the expressive ability of diversity, thus making up for the robustness deficiency of traditional methods in incomplete information scenarios. The experimental results show that this method reduces the attack success rate to 17.2% and enhances system stability to 89.1% in typical attack and defense scenarios, exhibiting superior strategy adaptability and robustness in complex environments.

Keywords: Cyber situational awareness, OT/ICS (SCADA) security, Spatiotemporal graph/Transformer, Closed-loop defense

1 Introduction

With the continuous improvement of the intelligence and informatization levels in the oil and gas industry, the network system and control system of the oil and gas pipeline network have gradually evolved into a highly integrated, networked and distributed collaborative system [1]. Industrial Control System (ICS) [2] are widely deployed in core processes such as data acquisition, remote monitoring and automatic execution, forming a typical Cyber-Physical System (CPS) [3]. This integration trend has significantly improved the efficiency of oil and gas transportation, the collaborative ability of equipment and the level of remote control. However, it also exposes the system to the risks of more complex and covert cyberattacks. Especially in the context of wide-area networking, multi-source access, and protocol heterogeneity, key elements such as communication paths, node permissions, and protocol stack structures in oil and gas pipeline networks are increasingly exposed. Attackers can achieve in-depth control and interference of the system through a chain of steps such as information reconnaissance, privilege acquisition, denial of service, lateral penetration, and sensitive data theft, which significantly increases the complexity and uncertainty of network security protection.

Current existing research has conducted extensive exploration on the security protection mechanisms of oil and gas pipeline networks. Among them, traditional methods mostly build defense strategies based on rule-driven or static configuration. However, such methods lack the ability to model the dynamic threat environment and thus have difficulty in coping with the continuous evolution and path switching of attack behaviors. Attackers usually have the advantage of information asymmetry. They can quickly adjust their behavioral strategies based on the detected information and conduct multi-round and multi-path attack interactions targeting key nodes. If the defender lacks the capabilities of environmental perception and feedback regulation, it is highly likely to fall into a passive response state, making it difficult to ensure the stable operation of the system under disturbances and the efficiency of resource utilization. As the security requirements shift from "post-event response" to "real-time adaptation", there is an urgent need to construct an intelligent defense system with strategy evolution capabilities to enhance the collaborative robustness of the defense model under attack dynamics, uncertainties, and resource constraints.

Against this background, game theory and reinforcement learning methods have gradually become important theoretical tools for intelligent defense modeling. Game theory reveals the optimal decision-making path under incomplete information by characterizing the strategic game relationship between attackers and defenders under multi-objective constraints [4]. Reinforcement learning leverages environmental feedback to drive the continuous

*Corresponding Author: Yinggang Xie; Email: xieyinggang@bistu.edu.cn
DOI: https://doi.org/10.70003/160792642026032702011

optimization of defense strategies and the accumulation of experience, providing an effective framework for dynamic strategy adjustment in complex environments [5]. Some studies have combined game theory with deep reinforcement learning for multi-step attack-defense modeling [6], joint strategy evolution, and long-term benefit estimation, showing certain effectiveness in static or semi-dynamic network scenarios. However, most of the existing models have the following shortcomings:

1) It is difficult to capture the actual strategic interaction laws under the conditions of multiple agents coexisting, local observation, and information delay in oil and gas pipeline networks.

2) The phased evolution characteristics of the attack chain are not considered, and the strategy design and defense feedback are not aligned with the entire process of the chained attack.

3) There is a lack of systematic evaluation of the adaptability, convergence, and resource allocation efficiency of the models in highly dynamic environments.

To address the above issues, this research proposes an intelligent network attack-defense modeling method that integrates a game reasoning mechanism, a deep reinforcement learning structure, and a strategy evolution and update mechanism. This method constructs a dual-agent confrontation structure, realizes the adaptive evolution of defense strategies through reinforcement learning, and introduces evolutionary game theory to characterize the dynamic nonlinear mechanism of the strategy distribution adjustment process, thereby enhancing the system's response robustness under unknown state transitions and attack disturbances. The model supports continuous optimization under incomplete information conditions, which is suitable for typical oil and gas pipeline network scenarios with chained attack characteristics.

At the methodological validation level, this article constructs a multi-dimensional simulation platform based on real control network structures, and designs offensive and defensive confrontation experiments covering the entire process including information reconnaissance, attack progression, and system interference. The experimental evaluation encompasses the system's performance across multiple dimensions such as critical node protection rates, resource scheduling efficiency, and policy convergence quality. Furthermore, by conducting multiple sets of comparative and ablation studies, the article analyzes the collaborative value and practical efficacy of each component module within the model. The research aims to provide a structural modeling approach and quantifiable validation pathways for intelligent security protection mechanisms in the complex network environments of oil and gas pipeline systems, thereby advancing the cybersecurity defense capabilities and adaptive evolution levels of critical infrastructure under highly adversarial and high-risk conditions.

2 Related Work

Driven by the industrial Internet, the control systems of oil and gas pipeline networks are continuously evolving into highly coupled cyber-physical systems. The realization of their functions increasingly depends on the collaborative support of complex communication networks and intelligent computing platforms. This deep integration promotes the integration of data collection, remote control, and task execution. However, it also significantly expands the attack surface. As a result, network security issues have shifted from static vulnerability identification to the defense against dynamic, multi-stage, and chained attacks, which are characterized by concealed attack paths, high strategic uncertainty, and strong destructive coupling. In scenarios involving high-value assets and realtime control in the oil and gas industry, attackers often interfere with the system operation process through multi-round resource infiltration and control-right enhancement strategies. Traditional static defense mechanisms have delayed responses and inflexible scheduling when facing continuously evolving attack behaviors, making it difficult to effectively ensure the security and stability of critical nodes [7].

To address the above-mentioned issues, existing research has proposed risk protection mechanisms based on attack graph modeling [8], abnormal behavior identification, and rule-driven detection. These methods typically achieve rapid trigger responses by constructing rule bases and attack behavior templates [9]. They have the advantages of high recognition accuracy and low implementation cost under known attack models. Some methods combine control instruction sequences with traffic time features to quickly locate abnormal link activities. However, such mechanisms are highly dependent on attack patterns and system states and lack the ability to update dynamically. They struggle to adapt to complex scenarios where attackers frequently adjust their strategies and there is strong interference from incomplete information. The system's robustness and generalization ability are still insufficient.

To better depict the offensive-defensive confrontation relationship, game theory has gradually become an important theoretical tool for network security modeling. The Stackelberg game model [10] can reflect the asymmetric information game between the attacker and the defender through the leader-follower structure. It is widely used to model task scenarios such as attack path selection, defense resource scheduling, and denial-of-service interference control. The evolutionary game model introduces a distribution evolution mechanism of historical round payoffs, which is suitable for depicting the sensitivity and stability feedback of participants to environmental changes during the strategy adjustment process. Related research has achieved the theoretical construction of self-adaptive strategy models in aspects such as static topology structure optimization, threat propagation path blocking, and key control link protection.

Although game models have strong interpretability and strategy derivation capabilities in theoretical analysis, they often face three types of challenges in industrial-level deployment environments. Firstly, the modeling process relies on precise prior information such as state-transition matrices and payoff functions. However, real-world oil

and gas control networks have strong non-linearity and feedback delay characteristics, making it difficult to meet the input requirements for modeling [11]. Secondly, constructing a strategy space in a high-dimensional state space requires a large amount of computing resources, leading to a significant increase in training costs and a decrease in strategy convergence. Thirdly, in dynamic interaction rounds, the game structure lacks a flexible control mechanism for strategy adjustment trajectories and payoff feedback. As a result, the strategy evolution path struggles to adapt to the complex behavioral feedback of the system [12].

Reinforcement learning methods offer a new technical approach for the adaptive update and long-term optimization of defense strategies. These methods continuously obtain state feedback and reward signals through interaction with the environment and gradually approach the optimal strategy function. They have the potential to handle unknown attacker behaviors, complex state-space changes, and dynamic interaction structures. Existing research has applied methods such as Q-learning [13], Deep Q-Network (DQN) [14], and Actor-Critic [15] to scenarios like intrusion detection, resource scheduling, and communication link optimization. These applications have effectively enhanced the flexibility of defense strategies and the training efficiency. Some models have demonstrated good attack suppression and node recovery capabilities on simulated platforms.

In industrial environments, reinforcement learning still has certain limitations. Problems such as low sample efficiency, slow training convergence, frequent policy oscillations, and limited adversarial capabilities are still hard to circumvent in long-term deployments. Particularly in scenarios with complex policy structures, delayed state feedback, and diverse attack targets, the stability, interpretability, and transferability of reinforcement learning models are inadequate, which impacts their practical application outcomes [16].

In light of the complementary strengths of game models and reinforcement learning methods, some studies have begun to explore their integration. By constructing the strategy interaction map in game models and the strategy update path in reinforcement learning, these studies aim to achieve dynamic optimization of strategies in complex interactive environments [17].

Theoretically, this approach can enhance the model's adaptability to attackers' behaviors and improve the convergence effect of defenders' strategies when facing uncertain game payoffs and state noise interference. The integrated models usually adopt a two-agent structure. Independent strategy update modules are designed for the attacker and the defender respectively, and joint training is conducted in a shared environment. These models exhibit good scalability and flexibility [6].

Although the integration approach has demonstrated certain research value, most current work remains limited to general communication networks or abstract security testing environments. A systematic interaction structure, state modeling, and reward design mechanism tailored to the specific scenarios of oil and gas industrial control systems [18] have not yet been established. In tasks such as modeling chained attack paths [19], balancing resource consumption and benefits, and scheduling the priorities of critical nodes, there is still a lack of a defense strategy design framework with strong adaptability, good interpretability, and high training efficiency. As a result, it fails to effectively meet the dual requirements of security and practical deployment capabilities.

3 Method

In the complex environment of oil and gas pipeline networks, the game process between attackers and defenders typically exhibits characteristics such as high dynamics, multidimensional strategy spaces, and phased evolution of strategies [20]. Attack behaviors are not executed as static strategies but instead manifest as a task-chain-based phased progression, encompassing steps like reconnaissance scanning, vulnerability exploitation, lateral movement, service disruption, privilege control, and sensitive data exfiltration. These attack paths are complex and unpredictable [7]. Meanwhile, defenders often operate within environments characterized by information lag and unobservable states, requiring timely responses based on current conditions. Consequently, they face significant challenges related to strategy uncertainty and adaptability [21].

To address this issue, the article proposes a multi-agent offensive and defensive modeling method that integrates a game reasoning mechanism, a reinforcement learning structure, and an evolutionary update mechanism. An intelligent defense model, namely the Hybrid-based Reinforcement Learning (HBRL) model, is constructed. This model features responsive flexibility, learning stability, and deployment generalization ability. The model employs a dual-agent modeling approach for attackers and defenders and incorporates an evolutionary game and a deep reinforcement learning structure to achieve dynamic game modeling under multi-stage decision-making. As shown in Figure 1, the overall structure of the HBRL model comprises an attacker agent, a defender agent, a state-space update module, and a reward calculation module. Both the attacker and the defender output their current actions based on their respective strategy distributions, which drives the evolution of the system state. Through reward feedback [22], they continuously optimize their strategic behaviors, thus closing the loop of dynamic games and adaptive evolution.

To formally characterize this structure, based on non-zero-sum game modeling, this research abstracts the offensive-defensive interaction into the following five-tuple:

$$G = \langle A, D, S, (U_A, U_D), T \rangle \quad \textbf{(1)}$$

Where A and D are the action sets of the attacker and the defender, respectively, S, (U_A, U_D) represents the system state space, and T are the corresponding payoff functions, and is the state transition function.

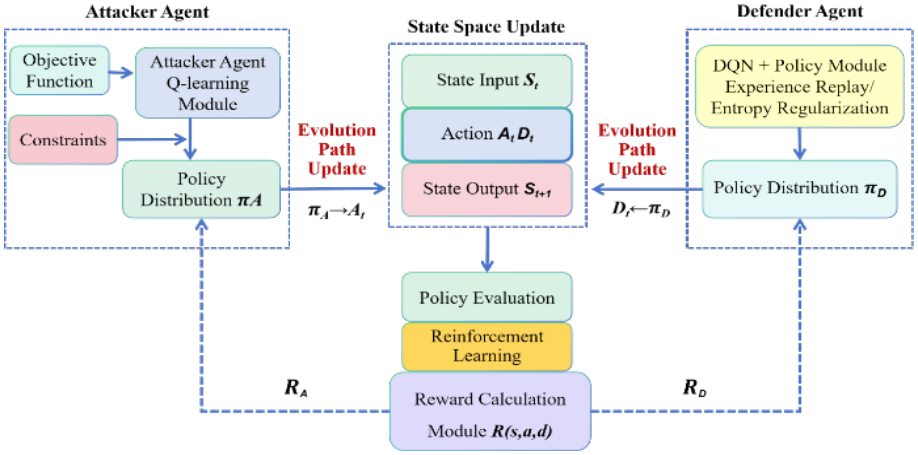

Figure 1. HBRL model structure

To characterize the uncertainty in strategy output and the mixed-behavior structure in reality, probabilistic strategy distribution modeling is introduced, which satisfies the normalization constraint of mixed strategies:

$$\pi_i(a_j) \in [0,1], \quad \sum_{j=1}^{|A_i|} \pi_i(a_j) = 1, \quad i \in \{A, D\} \tag{2}$$

The system state transition function is driven by the joint actions of the attacker and the defender. The probability of state transition is:

$$P\left(s' \middle| s, a, d\right) = Pr\left(S_{t+1} = s' \middle| S_t = s, A_t = a, D_t = d\right) \tag{3}$$

where $P \in \mathbb{R}^{|S|\times|S|\times|A|\times|D|}$

Based on the Bayesian expectation theory of mixed strategies [23], the game-level modeling of the profit function is carried out:

$$\mathbb{E}[U_A] = \sum_{s\in S}\sum_{a\in A}\sum_{d\in D} \pi_A(a)\pi_D(d)\cdot U_A(s,a,d) \tag{4}$$

To demonstrate that this model can logically derive the conclusion of "adaptive optimization of defense strategies," we must prove that a unique Nash equilibrium exists within the mixed strategy space and that this equilibrium is asymptotically stable. Therefore, we make the following assumptions:

Attacker Strategy Set $A = \{a_1, a_2, \ldots, a_m\}$ and defender Strategy Set $D = \{d_1, d_2, \ldots, d_m\}$; Both parties employ mixed strategies, where the attacker chooses strategy a_i with probability p_i , and the defender chooses strategy d_j with probability p_j, and satisfying $\sum p_i = 1, \sum q_j = 1$.

The payoff function is a bounded continuous function.

According to the Bayesian expected payoff definition in the original formula (4), the defender's expected payoff function can be rewritten in matrix form:

$$E[U_D(p,q)] = p^T M_D q \tag{5}$$

Among these, M_D is the defense payoff matrix.

We aim to prove that there exists a strategy pair (p^*, q^*) such that for any p, q satisfying:

$$\begin{aligned} E\left[U_D\left(p^*, q\right)\right] &\geq E\left[U_D\left(p, q\right)\right], \\ E\left[U_A\left(p, q^*\right)\right] &\geq E\left[U_A\left(p, q\right)\right] \end{aligned} \tag{6}$$

Define the defender's optimal response correspondence $BR_{D(q)}$ as:

$$BR_{D(q)} = \left\{p \in \Delta(A) \mid p^T M_{Dq} \geq p'^T M_{Dq}, \forall P' \in \Delta(A)\right\} \tag{7}$$

Since ΔA is a nonempty, compact, and convex strategy simplex, and U_D is linear with respect to p (and thus also pseudo-concave), by Kakutani's fixed-point theorem, the mapping $BR_{D(p,q)} = BR_{A(q)} \times BR_{D(p)}$ has at least one fixed point (p^*, q^*).

Therefore, in finite games defined by this quintuple model, a mixed-strategy Nash equilibrium necessarily exists.

To realize the self-evolution process of survival of the fittest for strategies during the game evolution, the replicator dynamic equation is introduced as the core mechanism for strategy evolution:

$$\frac{d\pi_A(a)}{dt} = \pi_A(a)\left[U_A(a, \pi_D) - \sum_{a'} \pi_A(a') U_A(a', \pi_D)\right] \tag{8}$$

When the evolutionarily stable point satisfies the condition of Nash equilibrium (ESS) [24], the target solution for strategy convergence is:

$$\pi_A^*(a) = \arg\max_{\pi}\left[\sum_d \pi_D(d) U_A(s,a,d)\right], \quad \text{subject to } \sum_a \pi(a) = 1 \tag{9}$$

Assume that for a fixed state s and attacker policy π_A, the defender's payoff function is linear, and the fitness function is continuously differentiable.

According to evolutionary game theory, the equilibrium point of the replicator dynamics constitutes a Nash equilibrium. Constructing Lyapunov functions

$$V = \sum_i p_i \ln(p_i / p_i^*) \tag{10}$$

Here, p_i denotes the proportion of the i-th pure strategy in the population, f_i represents the fitness (expected payoff) of that strategy, and $\overline{f}$ is the average fitness of the population. P_i^* is the strategy proportion at the ESS. Taking the derivative of V yields:

$$\frac{dV}{dt} = \sum_i (f_i - \overline{f}) \ln(p_i / p_i^*) \le 0 \tag{11}$$

Moreover, equality holds only when $p = p^*$. Therefore, the system asymptotically stabilizes to the ESS. In the HBRL model, although the defender's strategy update is driven by reinforcement learning, the replicator dynamics can serve as a theoretical analysis tool. This demonstrates that introducing this mechanism enhances the stability of strategy updates and avoids local oscillations. In practice, the attacker's strategy also evolves, but when its rate of change is slower than the defender's learning rate, this convergence property remains approximately valid.

This evolutionary game structure can be further extended to be represented by a Markov Decision Process (MDP) [25]:

$$\mathcal{M} = \langle S, A, P, R, \gamma \rangle \tag{12}$$

where R(s,a,d) represents the immediate reward function, and $\gamma \in (0,1)$ is the future discount factor. The strategic objective is to maximize the cumulative expected reward:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, d_t) \,|\, s_0 = s\right] \tag{13}$$

In actual strategy calculation, a deep reinforcement learning structure is introduced to approximately evaluate the state-action value function (Q-function) [5]. The defender's objective function:

$$Q_D(s,a) = \mathbb{E}_{\pi_D}\left[R(s,a,d) + \gamma \max_{a'} Q_D(s',a') | s, a\right] \tag{14}$$

The attacker adopts the Q-learning algorithm to update the strategy:

$$Q_A(s,a) = \mathbb{E}_{\pi_A}\left[R(s,a,d) + \gamma \max_{a'} Q_A(s',a') | s, a\right] \tag{15}$$

Given that resource constraints in oil and gas pipeline network defense lead to diminishing marginal returns in the objective function, we introduce a regularization term to the original objective function and construct a strictly concave effective objective function:

$$\tilde{U}_D(q) = q^T (M_D^T p) - \lambda \sum_{j=1}^{n} q_j \ln q_j \ (\lambda > 0) \tag{16}$$

Find the second derivative of this function (Hessian matrix):

$$H(\tilde{U}_D) = -\lambda \cdot \mathrm{diag}(1/q_1, 1/q_2, ..., 1/q_n) \tag{17}$$

Since $q_j > 0$ and $\lambda > 0$, then $H(\tilde{U}_D)$ is a negative definite matrix.

After introducing entropy regularization, the defender's optimization problem becomes strictly concave, thereby guaranteeing that the defender's optimal response is unique for a given attack strategy. The same holds true for the attacker's side. Consequently, this game possesses a unique mixed-strategy Nash equilibrium.

To address the issues of strategies getting trapped in local optima or experiencing fluctuations during real-world training, an Experience Replay mechanism [26] and a target network synchronization structure are introduced to mitigate update oscillations and enhance training stability. Meanwhile, to improve the generalization ability and fault-tolerance of the strategy distribution, a strategy entropy regularization term is added to the loss function:

$$J(\pi) = \mathbb{E}_{\pi}[Q(s,a)] + \beta \cdot \mathcal{H}(\pi), \quad \mathcal{H}(\pi) = -\sum_a \pi(a) \log \pi(a) \tag{18}$$

The distribution entropy of the strategy output can adjust the strategy stability and generalization ability of the model under conditions of uncertain states, data loss, and dynamic fluctuations. Especially in the case of diverse attack chains (such as link interruption, control instruction forgery, abnormal detection traffic, etc.) [27], the model needs to adjust the strategy priority and strategy distribution at high-risk nodes, which poses higher requirements for the robustness and perturbation sensitivity of the strategy distribution. To this end, this article also introduces a local perturbation test mechanism in the construction of the state space to ensure that the strategy network can produce stable response outputs to small-amplitude state perturbations and input loss.

Further considering the resource consumption and the cost of damage to critical nodes caused by strategy execution, three indicators are introduced in the design of the reward function:

$$R_D(s,a,d) = \lambda_1 \cdot C_{\text{survive}} - \lambda_2 \cdot C_{\text{damage}} - \lambda_3 \cdot C_{\text{cost}},\ i \in \{A, D\} \tag{19}$$

where C_{survive} measures the proportion of normal operation of the system's critical nodes. C_{damage} represents the losses caused by link failures and node paralysis triggered by attack behaviors. C_{cost} indicates the degree of resource occupation consumed by defensive response operations.

This reward structure supports the adjustment of defensive behaviors under different attack target priorities and constructs an interpretable and deployable strategic optimization target system.

4 Experiments

4.1 Experimental Setup

To verify the adaptability and effectiveness of the proposed intelligent attack-defense strategy that integrates game modeling and deep reinforcement learning mechanisms in the oil and gas pipeline network scenario, this article conducts an experimental design relying on multi-dimensional industrial control network simulation platform. This platform emulates a real industrial control network structure, based on the prototype of a regional oil and gas pipeline industrial control network. The simulated network comprises 32 core network nodes, as illustrated in Figure 2.

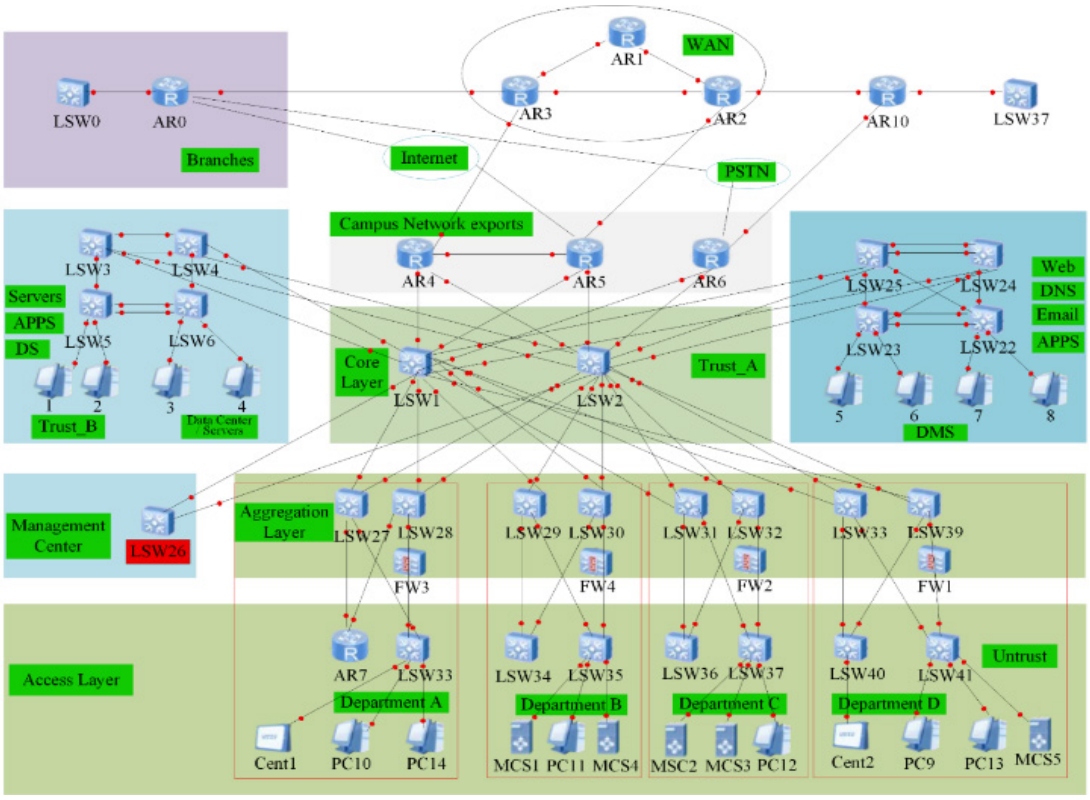

Figure 2. Simulation diagram of a regional oil and gas pipeline industrial control network structure

The configuration of the main firewalls in the network is shown in Table 1.

Table 1. Firewall configuration rules

Device	Port	Security zone	Priority
FW	GE1/0/1	Trust_A	70
	GE1/0/1.2	Trust_B	80
	GE1/0/0	DMZ	Default
	GE1/0/2	Untrust	Default

The attacker has 12 types of typical attack actions, including port scanning, command injection, forged communication, and denial-of-service attacks. The defender has 9 types of response strategies and can implement operations such as communication isolation, redundant switching, traffic throttling, and access reconstruction. The execution effect of the strategies will dynamically affect the system operating state and guide the evolution of the game.

To facilitate comparison, the tests employed a defined attack path, which is depicted in Figure 3.

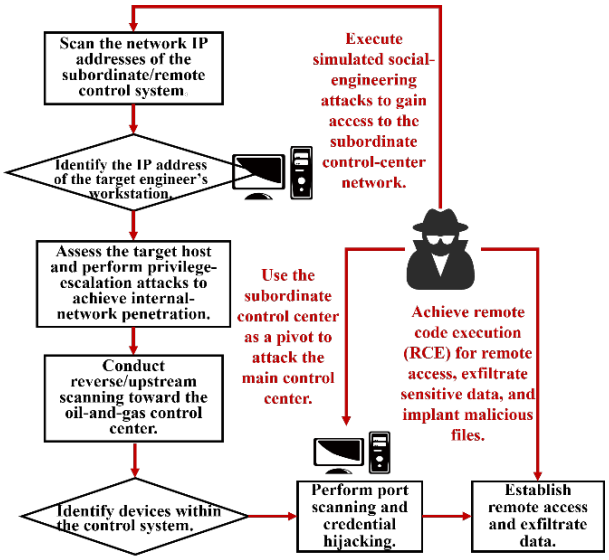

Figure 3. Attack chain diagram

The experimental platform integrates NS3 communication simulation and Python strategy training modules, enabling dual-agent interaction modeling through a state synchronization mechanism. Key hyperparameters during model training are set as shown in Table 2.

Each round of attack-defense simulation is set to 100 steps, with a total of 5,000 training rounds. The experiment adopts 20 independent runs to ensure statistical robustness by averaging the results. The defense agent employs a DQN-based learning architecture, with a neural network containing two hidden layers (128 and 64 neurons, respectively), using the ReLU activation function. The optimizer is Adam with a learning rate of 0.0005. The training process incorporates experience replay and delayed synchronization of the target network to mitigate update fluctuations. To address convergence speed and stability requirements in complex strategy spaces, an entropy regularization term is introduced to adjust the strategy distribution, guiding the defense agent to balance exploration and exploitation in the strategy space.

Table 2. Hyperparameter settings in the training process of the HBRL model

Parameter name	Setting value	Description
Learning rate α	0.0005	Initial learning rate of the Adam optimizer
Discount factor γ	0.9	Control the influence weight of future rewards
Experience replay capacity	5000	Size of the state-action storage buffer
Update frequency	Every 20 steps	Synchronization cycle between the target network and the main network
Batch size	64	Number of replay samples sampled for each update
DQN hidden layer structure	[128,64]	Number of neurons in the hidden layer
Exploration strategy	ϵ-greedy	Strategy to balance exploration and exploitation

4.2 Performance Evaluation Indicators

To quantify the performance of the strategy model from multiple dimensions, this article designs three evaluation metrics: System Stability Rate (SR), Attack Success Rate (ASR) [28], and Defense Reward (DR) [29]. The mathematical definitions of each metric are as follows.

The System Stability Rate is defined as the proportion of critical nodes maintaining operational status per unit time:

$$SR = \frac{1}{T}\sum_{t=1}^{T}\frac{|V_{\text{active}}^{t}|}{|V|} \tag{20}$$

where V_{active}^{t} represents the set of nodes in normal operational status at time step t, $|V|$ is the total number of all critical nodes, and T is the total number of simulation steps.

The attack success rate reflects the proportion of successfully compromised nodes or completed data infiltration by an attacker throughout the entire experimental cycle, and is defined as:

$$ASR = \frac{N_{\text{attack success}}}{N_{\text{attack attempts}}} \tag{21}$$

The cumulative defense benefit is used to measure the total reward of a strategy throughout the entire training process, and is expressed in the form of the cumulative sum of the reward function R_D(s, a, d) as:

$$DR = \sum_{t=1}^{T} R_D(s_t, a_t, d_t) \tag{22}$$

4.3 Comparative Experiment

To validate the performance advantages of the proposed method, four typical defense strategies were established as comparative models: a Static Rule-based Defense method (SRD), a policy model employing standard DQN network for deep reinforcement learning (PDQN), a GTAD model without reinforcement learning mechanisms that relies solely on static game-theoretic inference structures, and the HBRL model method proposed in this study. All the aforementioned methods were executed on a unified experimental platform while maintaining identical state spaces, network structures, and

hyperparameter configurations to ensure experimental fairness and reproducibility.

Based on the hyperparameter settings in Table 2 above, dynamic simulations were conducted to observe the evolution of network attack-defense situations under the four defense strategies. After three rounds of dynamic simulations, the dataset with the highest infection rate for each strategy was selected.

The results are shown in Figure 4, the horizontal axis represents the simulation time, with the total duration of the attack-defense simulation set to 100 steps. The vertical axis represents the proportion of different types of nodes to the total number of network nodes, with a maximum value of 100% Considering the confidentiality, integrity, and availability of the network information system, it is defined that when the peak proportion of infected nodes (IN) exceeds 10% it is considered a large-scale outbreak of the virus. When the peak proportion of infected nodes (IN) exceeds 50%, it is concluded that the network information system has been paralyzed due to the attack.

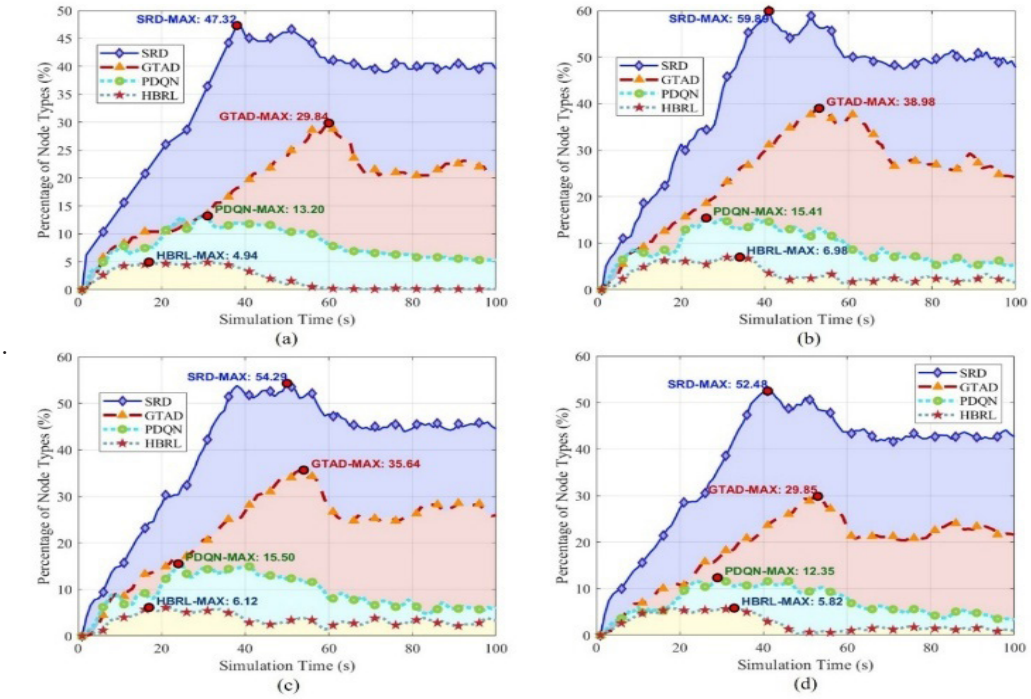

Figure 4. Evolution trend chart of cyber offensive and defensive posture under four types of defense strategies

A comparison of the Attack Success Rate (ASR) and stability among different models is shown in Figure 5. It can be observed that at t = 39, the peak proportion of infected nodes in the SRD model reached 47.32%. The evolutionary trend of the network attack-defense situation in this scenario indicates that the virus caused a large-scale outbreak within a relatively short period, ultimately leading to the paralysis of the network information system.

At t = 62, the peak proportion of infected nodes in the GTAD model reached 29.84%. The evolutionary trend of the network attack-defense situation in this scenario indicates that the virus caused a large-scale outbreak within a certain period, potentially resulting in significant damage to the network information system.

When t = 32, the peak proportion of infected nodes in the PDQN model reached 12.97%. The evolutionary trend of the network attack-defense situation in this scenario indicates that the virus outbreak occurred on a certain scale, with the scale remaining within a controllable range, potentially causing some damage to the network information system.

At t = 18, the HBRL model proposed in this article achieved apeak infected node proportion of 4.94%. The evolutionary trend of the network attack-defense situation in this scenario demonstrates that the virus did not trigger a large-scale out-break, causing only minor damage to the network information system.

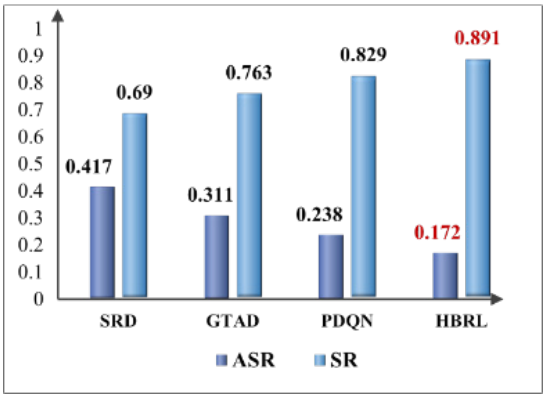

Figure 5. ASR and stability comparison across models

In terms of game information, this article analyzes network attack-defense interactions based on incomplete information game theory, where both attackers and

defenders can be of multiple types. For situational evolution analysis, drawing on infectious disease dynamics theory, we propose a definition and analytical method for network attack-defense situational evolution. Regarding experimental scenarios, the NetLogo multi-agent simulation tool is employed to dynamically simulate the evolution of network attack-defense situations over time, with a large-scale network node setting in the simulation environment. Compared to existing literature, our approach better aligns with the reality of incomplete information possessed by both parties in network conflicts. It enables the analysis and interpretation of macro-level situational evolution trends from the perspective of micro-level attack-defense behaviors, offering advantages such as applicability to large-scale scenarios and intuitive situational visualization.

Table 3 presents the average experimental results of each model across three metrics: SR, ASR, and DR.

Table 3. Comparative results of various models

Model	SR	ASR	DR
SRD	0.690	0.417	211.6
GTAD	0.763	0.311	298.2
PDQN	0.829	0.238	372.5
HBRL	0.891	0.172	458.3

From the experimental results, it can be observed that the HBRL model significantly outperforms the comparative models across all three-evaluation metrics, fully demonstrating the synergistic effects brought by the integration of game-theoretic reasoning structures and deep reinforcement learning mechanisms. A system stability metric of nearly 0.9 indicates that the proposed method maintains strong key node preservation capabilities under sustained interference environments. The attack success rate dropping to 0.172 shows that the defense strategy can effectively restrict the progressive advancement of attack chains, particularly exhibiting stronger containment effects in scenarios with unobservable states and frequent policy transitions.

It is noteworthy that the advantages of the HBRL model are not only manifested in policy performance but also reflected in training efficiency and policy generalization capabilities. The evolutionary path perturbation mechanism introduced in this model drives the policy distribution to continuously converge toward the desired evolutionary direction, while the experience replay and target network architecture of the DQN framework mitigate the fluctuation issues commonly observed in traditional deep reinforcement learning during early training stages, thereby accelerating convergence and enhancing policy robustness.

In contrast, while PDQN demonstrates certain advantages in defensive efficiency, it exhibits a notable decline in policy convergence speed under high-dimensional state perturbations. The model's training strongly depends on initial values, resulting in relatively poor policy stability. Although the GTAD model constructs a comprehensive policy space based on game theory, its lack of a state-action feedback mechanism prevents its response strategies from dynamically adapting to environmental changes. This leads to a higher attack success rate compared to both PDQN and HBRL. The SRD model performs the worst across all metrics, as its static response mechanism struggles to adapt to the dynamic evolution of task chains, showing significant performance degradation under high-frequency multi-stage attack chains.

4.4 Ablation Experiment

To further investigate the specific impact of key modules in the HBRL model on overall performance, three ablation experiments were designed: removing the game-theoretic modeling structure (w/o Game), eliminating the replicator dynamics mechanism (w/o Replicator), and simplifying the reward function structure to a single factor (w/o Reward Structure). These experiments aimed to analyze the differential roles of each component in strategy generation and defensive performance. The ablation models maintained consistent parameter configurations with the full model in terms of training epochs, state space, and network architecture, with the results shown in Table 4.

Table 4. Results of module ablation experiments

Configuration	SR	ASR	DR
HBRL	0.892	0.172	461.8
w/o Game	0.816	0.263	371.2
w/o Replicator	0.843	0.229	389.5
w/o Reward Struct	0.791	0.286	328.7

After removing the game-theoretic modeling structure, the evolutionary path of model strategies lacks the constraints of game equilibrium, which easily leads to deviations from reasonable expectations of adversarial behavior during strategy updates. This manifests as a significant increase in attack success rate and a noticeable decline in cumulative rewards. The removal of the replicator dynamics mechanism prevents the strategy evolution from capturing the dynamic feedback of historical utility, exacerbating the influence of initial distribution on the strategy convergence process. When facing high-frequency disturbance scenarios, the model exhibits a certain degree of lag, resulting in reduced adaptability of defensive behaviors. Simplifying the reward function structure to a single objective causes the model to lose its ability to perceive multi-state feedback from the system. Consequently, the strategy convergence process lacks comprehensive balancing between damage suppression and resource constraints, leading to the lowest levels of both system stability and defensive rewards.

The performance comparisons demonstrate that each component module in the HBRL model critically impacts its final performance. The synergistic interactions among the modules form a stable, adaptive, and multi-objective balanced evolutionary path for defensive strategies, providing structural support for cybersecurity protection in complex oil and gas pipeline network environments.

5 Conclusion

In the process of informatization and intelligentization, the functions carried by the oil and gas pipeline network control system continue to expand, and the network security threats it faces are increasingly characterized by high dynamics, diversity, and strategic evolution. Against this backdrop, how to construct intelligent defense strategies with reasoning ability, adaptive update mechanisms, and attack prediction capabilities has become a key technical issue in the security of industrial control networks. To address this challenge, the article proposes an intelligent attack-defense modeling method that integrates game modeling, deep reinforcement learning, and strategy evolution mechanisms, aiming to optimize defense strategies and model their dynamic evolution in complex and uncertain environments.

In terms of theoretical design, the method proposed in this research integrates the non-zero-sum game structure and the Markov decision process to construct a unified modeling framework that can depict strategic interactions, system state transitions, and the optimization of expected returns. By introducing the replicator dynamic evolution mechanism, the study realizes the self-adaptation and response capabilities of defense strategies when facing continuously changing attack behaviors. Further, combined with the deep Q-network structure, it completes the reinforcement learning modeling of the high-dimensional strategy space. In the design of the reward function, this article fully considers the multi-objective attributes of defense strategies. Factors such as the survival rate of key nodes, the degree of network damage, and resource scheduling costs are introduced to construct a composite revenue function that is measurable and has practical deployment reference value.

To validate the practical effectiveness of the proposed model, this research constructs a simulation experiment platform based on the real control network architecture of the oil and gas industry and conducts multiple sets of comparative experiments and module ablation tests. The experimental results show that the model proposed in this article exhibits superior performance in terms of system stability, attack suppression ability, and long-term defense benefits. When compared with three types of typical control models, this method consistently outperforms in terms of the stable operation ability of key nodes and attack suppression effects. The cumulative strategy revenue is significantly higher than that of other models, indicating that it achieves a more reasonable balance between protection effectiveness and resource efficiency. The ablation experiments further verify the positive contributions of the game reasoning structure, strategy evolution mechanism, and composite reward design to the overall model performance, demonstrating the independent effectiveness and collaborative value of each module in the strategy update path.

Combining the current experimental verification results and the scalability of the model framework, future research can further conduct in-depth exploration on the collaborative game mechanism of multi-agent linked defense strategies, the optimization path of transfer reinforcement learning based on historical scenarios, and the deployment feasibility on real industrial platforms.

Acknowledgement

This work is supported by National Key Research and Development Program Project (Grant No. 2023YFB3107700); Provincial Science and Technology Plan Project of Hebei Province (Grant No. 253A7634D).

References

[1] M. Y. Zemenkova, E. L. Chizhevskaya, Y. D. Zemenkov, Intelligent monitoring of the condition of hydrocarbon pipeline transport facilities using neural network technologies, *Journal of Mining Institute*, Vol. 258, pp. 933–944, December, 2022. https://doi.org/10.31897/PMI.2022.105

[2] D. Bhamare, M. Zolanvari, A. Erbad, R. Jain, K. Khan, N. Meskin, Cybersecurity for industrial control systems: A survey, *Computers & Security*, Vol. 89, Article No. 101677, February, 2020. https://doi.org/10.1016/j.cose.2019.101677

[3] W. Duo, M. C. Zhou, A. Abusorrah, A survey of cyber attacks on cyber physical systems: Recent advances and challenges, *IEEE/CAA Journal of Automatica Sinica*, Vol. 9, No. 5, pp. 784–800, May, 2022. https://doi.org/10.1109/JAS.2022.105548

[4] E. N. Barron, *Game Theory: An Introduction*, John Wiley & Sons, 2024.

[5] S. E. Li, Deep reinforcement learning, in: *Reinforcement Learning for Sequential Decision and Optimal Control*, Springer Nature Singapore Pte Ltd., 2023, pp. 365–402. https://doi.org/10.1007/978-981-19-7784-8_10

[6] Y. Li, M. Cheng, C. J. Hsieh, T. C. M. Lee, A review of adversarial attack and defense for classification methods, *The American Statistician*, Vol. 76, No. 4, pp. 329–345, 2022. https://doi.org/10.1080/00031305.2021.2006781

[7] G. Stergiopoulos, D. A. Gritzalis, E. Limnaios, Cyber-attacks on the oil & gas sector: A survey on incident assessment and attack patterns, *IEEE Access*, Vol. 8, pp. 128440–128475, July, 2020. https://doi.org/10.1109/ACCESS.2020.3007960

[8] A. Presekal, A. Stefanov, V. S. Rajkumar, P. Palensky, Attack graph model for cyber-physical power systems using hybrid deep learning, *IEEE Transactions on Smart Grid*, Vol. 14, No. 5, pp. 4007–4020, September, 2023. https://doi.org/10.1109/TSG.2023.3237011

[9] M. A. S. P. Dayarathne, M. S. M. Jayathilaka, R. M. V. A. Bandara, V. Logeeshan, S. Kumarawadu, C. Wanigasekara, Mitigating cyber risks in smart cyber-physical power systems through deep learning and hybrid security models, *IEEE Access*, Vol. 13, pp. 37474–37492, February, 2025. https://doi.org/10.1109/ACCESS.2025.3545637

[10] M. Tang, H. Liao, X. Wu, A Stackelberg game model for large-scale group decision making based on cooperative incentives, *Information Fusion*, Vol. 96, pp. 103–116, August, 2023. https://doi.org/10.1016/j.inffus.2023.03.013

[11] K. H. Lee, R. Baldick, Solving three-player games by the matrix approach with application to an electric power market, *IEEE Transactions on Power Systems*, Vol. 18, No. 4, pp. 1573–1580, November, 2003. https://doi.org/10.1109/TPWRS.2003.818744

[12] M. Wang, Y. Li, Z. Cheng, C. Zhong, W. Ma, Evolution and equilibrium of a green technological innovation system: Simulation of a tripartite game model, *Journal of Cleaner Production*, Vol. 278, Article No. 123944, January, 2021. https://doi.org/10.1016/j.jclepro.2020.123944

[13] A. Kumar, A. Zhou, G. Tucker, S. Levine, Conservative Q-learning for offline reinforcement learning, *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2020, pp. 1179–1191. https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf

[14] A. Oroojlooyjadid, M. R. Nazari, L. V. Snyder, M. Takáč, A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization, *Manufacturing & Service Operations Management*, Vol. 24, No. 1, pp. 285–304, January-February, 2022. https://doi.org/10.1287/msom.2020.0939

[15] C. A. Cheng, T. Xie, N. Jiang, A. Agarwal, Adversarially trained actor critic for offline reinforcement learning, *Proceedings of the 39th International Conference on Machine Learning*, *PMLR*, Vol. 162, Baltimore, Maryland, USA, 2022, pp. 3852–3878. https://proceedings.mlr.press/v162/cheng22b/cheng22b.pdf

[16] O. Dogru, J. Xie, O. Prakash, R. Chiplunkar, J. Soesanto, H. Chen, K. Velswamy, F. Ibrahim, B. Huang, Reinforcement learning in process industries: Review and perspective, *IEEE/CAA Journal of Automatica Sinica*, Vol. 11, No. 2, pp. 283–300, February, 2024. https://doi.org/10.1109/JAS.2024.124227

[17] A. Rajeswaran, I. Mordatch, V. Kumar, A game theoretic framework for model based reinforcement learning, *Proceedings of the 37th International Conference on Machine Learning*, *PMLR*, Vol. 119, Virtual Event, 2020, pp. 7953–7963. https://proceedings.mlr.press/v119/rajeswaran20a/rajeswaran20a.pdf

[18] T. R. Wanasinghe, R. G. Gosine, L. A. James, G. K. I. Mann, O. de Silva, P. J. Warrian, The internet of things in the oil and gas industry: A systematic review, *IEEE Internet of Things Journal*, Vol. 7, No. 9, pp. 8654–8673, September, 2020. https://doi.org/10.1109/JIOT.2020.2995617

[19] H. Jmal, F. B. Hmida, N. Basta, M. Ikram, M. A. Kaafar, A. Walker, SPGNN-API: A Transferable Graph Neural Network for Attack Paths Identification and Autonomous Mitigation, *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp. 1601-1613, 2024. https://doi.org/10.1109/TIFS.2023.3338965

[20] A. S. Mohammed, P. Reinecke, P. Burnap, O. Rana, E. Anthi, Cybersecurity challenges in the offshore oil and gas industry: An industrial cyber-physical systems (ICPS) perspective, *ACM Transactions on Cyber-Physical Systems*, Vol. 6, No. 3, pp. 1–27, July, 2022. https://doi.org/10.1145/3548691

[21] Y. Cui, N. Quddus, C. V. Mashuga, Bayesian network and game theory risk assessment model for third-party damage to oil and gas pipelines, *Process Safety and Environmental Protection*, Vol. 134, pp. 178–188, February, 2020. https://doi.org/10.1016/j.psep.2019.11.038

[22] E. Bates, V. Mavroudis, C. Hicks, Reward shaping for happier autonomous cyber security agents, *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, Copenhagen, Denmark, 2023, pp. 221–232. https://doi.org/10.1145/3605764.3623916

[23] A. Vehtari, A. Gelman, T. Sivula, P. Jylänki, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, C. Robert, Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data, *Journal of Machine Learning Research*, Vol. 21, No. 17, pp. 1–53, 2020. https://jmlr.org/papers/v21/18-817.html

[24] J. Jo, J. Yu, J. Park, Incentive design of shared ESS energy trading game, *IEEE Transactions on Control Systems Technology*, Vol. 33, No. 1, pp. 408–415, January, 2025. https://doi.org/10.1109/TCST.2024.3483440

[25] E. Altman, *Constrained Markov Decision Processes*, Routledge, Abingdon, 1999.

[26] W. Fedus, P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, W. Dabney, Revisiting fundamentals of experience replay, *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Vol. 119, Virtual Event, 2020, pp. 3061–3071. https://proceedings.mlr.press/v119/fedus20a/fedus20a.pdf

[27] D. Deyannis, E. Papadogiannaki, G. Chrysos, K. Georgopoulos, S. Ioannidis, The diversification and enhancement of an IDS scheme for the cybersecurity needs of modern supply chains, *Electronics*, Vol. 11, No. 13, Article No. 1944, July, 2022. https://doi.org/10.3390/electronics11131944

[28] B. Ma, C. Zhao, D. Wang, B. Meng, DIHBA: Dynamic, invisible and high attack success rate boundary backdoor attack with low poison ratio, *Computers & Security*, Vol. 129, Article No. 103212, June, 2023. https://doi.org/10.1016/j.cose.2023.103212

[29] X. Zhang, Y. Ma, A. Singla, X. Zhu, Adaptive reward-poisoning attacks against reinforcement learning, *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Vol, 119, Virtual Event, 2020, pp. 11225–11234. https://proceedings.mlr.press/v119/zhang20u/zhang20u.pdf

Biographies

Lidong Jia graduated from China University of Petroleum (East China) and worked for PipeChina North Pipeline Company with a bachelor's degree. He has experience in industrial control system network security. He has participated in multiple scientific research projects and has rich experience in industrial control system network security.

Fei Song is currently working at PipeChina North Pipeline Company, where his research interests include information security of intelligent manufacturing systems, digital control. He has participated in several related project studies, with a core focus on the design of industrial internet security architecture .

Weichun Hei works at PipeChina North Pipeline Company, mainly engaged in the fields of industrial control system network security, with a solid theoretical foundation and rich practical experience. In terms of industrial control network security, it can implement risk assessment, vulnerability protection, and security reinforcement plans.

Jian Wang received a Bachelor's degree in Measurement & Control Technology and Instruments from China University of Petroleum (East China) in Shandong, China In 2009. Currently, he serves as the Director of the Instrumentation Automation Department at the Technical Support Center of PipeChina North Pipeline Company. His current research interests network security technology for industrial control systems.

Yinggang Xie received a Doctor's degree in Control Theory and Control Engineering from the University of Science and Technology Beijing, Beijing, China, In 2007. Currently, he serves as a professor in the Department of Internet of Things at Beijing Information Science and Technology University, Beijing, China. His current research interests include artificial intelligence algorithms, and Internet of Things application technologies.