Intelligent and Converged Networks
2023, 4(2): 116−126

ISSN 2708-6240
DOI: 10.23919/ICN.2023.0012

Q-learning based strategy analysis of cyber-physical systems considering unequal cost

Xin Chen, Jixiang Cheng, Luanjuan Jiang*, Qianmu Li, Ting Wang, and Dafang Li

Abstract: This paper proposes a cyber security strategy for cyber-physical systems (CPS) based on Q-learning under unequal cost to obtain a more efficient and low-cost cyber security defense strategy with misclassification interference. The system loss caused by strategy selection errors in the cyber security of CPS is often considered equal. However, sometimes the cost associated with different errors in strategy selection may not always be the same due to the severity of the consequences of misclassification. Therefore, unequal costs referring to the fact that different strategy selection errors may result in different levels of system losses can significantly affect the overall performance of the strategy selection process. By introducing a weight parameter that adjusts the unequal cost associated with different types of misclassification errors, a modified Q-learning algorithm is proposed to develop a defense strategy that minimizes system loss in CPS with misclassification interference, and the objective of the algorithm is shifted towards minimizing the overall cost. Finally, simulations are conducted to compare the proposed approach with the standard Q-learning based cyber security strategy method, which assumes equal costs for all types of misclassification errors. The results demonstrate the effectiveness and feasibility of the proposed research.

Key words: cyber security; Q-learning; policy selection; unequal cost; misclassification interference

1 Introduction

The advent of the Internet of things (IoT), enabling real-time perception of the physical world and providing data support for intelligent decision-making[1, 2], has spurred the deployment of intelligent cyber-physical systems (CPS) that exploit wireless networking paradigms[3, 4]. CPS integrates physical processes, and computational and communication elements to create systems that can monitor and control physical processes in real-time. These systems range from small-scale devices, such as sensors and actuators, to large-scale infrastructures, such as smart factories and transportation systems[5]. Cyber attacks on CPS can cause physical damage and financial losses, and compromise personal data, making the security of CPS essential for the integrity and privacy of IoT devices and the infrastructure they manage[6]. Given the backbone role CPS played in industrial process and production control (smart factory), particularly in the context of the IoT[7], keeping CPS security has taken its top priority for most business units and attracted the largest share of attention from researchers and engineers in many fields. However, the classical defense strategies for CPS security including anti-viruses, firewalls, and intrusion detection systems (IDS)[8] tools are facing more serious challenges due to the rapidly evolving environment, especially the dynamic correlations between the cyber layer and the physical layer of a CPS[9]. An important objective for CPS security is to design effective strategies that can prevent cyberattacks more proactively.

A large volume of studies, have been conducted by Lalropuia and Gupta[10], Jin et al.[11], Huang et al.[12],

• Xin Chen, Jixiang Cheng, Ting Wang, and Dafang Li are with the School of Management Science and Engineering, Nanjing University of Finance and Economics, Nanjing 210023, China. E-mail: njuechx@163.com; airenchan2000@gmail.com; tingwang@nufe.edu.cn; df.li@nufe.edu.cn.

• Luanjuan Jiang and Qianmu Li are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210014, China. E-mail: njgesang@hotmail.com; qianmu@njust.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2023-03-31; revised: 2023-04-30; accepted: 2023-05-30

etc., on the CPS defense strategies analysis from various perspectives. However, most of them only assume that the potential loss caused by different defense strategies is equal regardless of the real cost to CPS. Actually, in many real-world applications, different errors in strategy selection may result in different levels of system losses. For example, the defender will lose only the strategy deployment cost when CPS is not at risk of being attacked, however, the loss could be much larger if the defender chooses a non-defense strategy when CPS is really at risk of being attacked. This introduces the concept of unequal cost, which needs to be addressed in order to accurately evaluate the effectiveness of different strategies. Although some studies are focusing on the cost-sensitive issues in the field of machine learning[13], such as the varying costs of different types of misclassification errors in the missing healthcare data prediction[14, 15], to our knowledge, little attention has been paid to this stream in the literature regarding the unequal cost of different defense strategies in CPS security management. To address this issue, a modification to the standard Q-learning algorithm is proposed which takes into account unequal costs by introducing a weight parameter that adjusts the associated cost based on the types of misclassification errors. This modified Q-learning algorithm can be applied to a wide range of CPS that requires efficient decision-making strategies for optimal performance[16]. An intuitive research motivation diagram is given in Fig. 1.

The main contributions of this paper are summarized as follows:

(1) We develop a novel model for the cyber security of CPS based on the Markov decision process (MDP). In the presence of an adversary, the cyber layer of the CPS system subsumes devices such as firewall, web server, database server, etc.

(2) We characterize precisely but intuitively when a defense strategy is reasonable in the presence of misclassifications and show how unequal cost is to optimize the combination of cyber security defense actions for CPS to obtain a more realistic defense strategy.

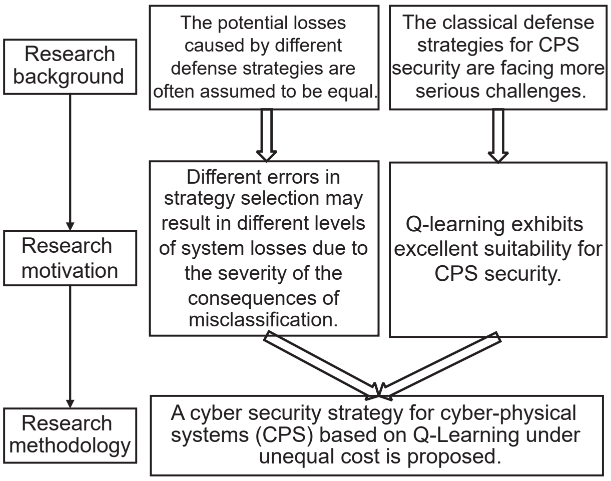

Research background
The potential losses caused by different defense strategies are often assumed to be equal.
The classical defense strategies for CPS security are facing more serious challenges.
Research motivation
Different errors in strategy selection may result in different levels of system losses due to the severity of the consequences of misclassification.
Q-learning exhibits excellent suitability for CPS security.
Research methodology
A cyber security strategy for cyber-physical systems (CPS) based on Q-Learning under unequal cost is proposed.

Fig. 1 Research motivation and methodology.

(3) We propose a modified Q-learning algorithm that incorporates unequal costs, which is important for optimizing the performance of CPS where system loss may vary depending on different errors in strategy selection, demonstrated through simulations.

The remainder of this paper is organized as follows: Section 2 discusses the pioneering work that has been done on CPS. Section 3 presents a model of a CPS with a cyber layer and a physical layer and defines unequal cost. A cyber security strategy for CPS based on Q-Learning under unequal cost is introduced in Section 4. Numerical experiments and analysis are presented in Section 5. And finally, in Section 6, the paper is concluded and an outlook for future work is given.

2 Related work

In this section, an overview pertaining to the modeling aspects of the security of CPS is provided, as well as the motivations driving research in this area. In recent years, several researchers have explored the use of game-theoretic models and reinforcement learning (RL) techniques to develop novel security decision-making approaches for CPS.

2.1 Game-theoretic models for CPS

Ma et al.[17] and Orojloo and Azgomi[18] have developed game theoretic models to predict the interactions between a cyber-physical attacker and a defender in different phases of intrusion and disruption. Huang et al.[19] have proposed a Markov attack-defense differential game model to analyze multi-stage continuous attack-defense processes, while Sun et al.[20] have proposed a method to quantify the benefits of attack and defense strategies, considering the misdetection defects of the defense detection system in CPS. Guo et al.[21] employed a game-theoretic approach to analyze the interactions between attackers and defenders in CPS, as well as the interdependency between the cyber and physical layers in CPS, enabling the quantification of the impact of cyber attacks on physical damage in CPS, and facilitated the development of effective defense strategies. However, these studies have not taken into account that some CPS environments are so complex that it is different to model them. For example, it is necessary that generate data first, and then only the dataset is used to train the model, which is one aspect that makes it difficult to simulate and solve such complex CPS environments. Therefore, reinforcement learning is employed to approach these questions.

2.2 Reinforcement learning for CPS

Previous researchers have investigated methods for optimizing the performance of CPS based on RL. Some studies have focused on incorporating different environmental factors or constraints into reinforcement learning, such as energy consumption or communication delay. Other studies have explored the use of deep learning techniques[22] to improve the performance of reinforcement learning in CPS[1]. Gai and Qiu[23] utilized RL to achieve intelligent content-centric services and obtain highly accurate quality of experience (QoE) in resource allocations for IoT, resolving the contradiction between performance and strategy generation. Huang et al.[12] considered both the cyber and physical layers of CPS through quantitative vulnerability analysis and time-based unified payoff quantification and used RL to generate optimal defense strategies in the absence of complete game parameter knowledge. This approach could minimize system losses caused by cyber-attacks in real-world CPS. Khoury and Nassar[24] proposed to frame CPS security on two different levels, strategic and battlefield, by meeting ideas from game theory and multi-agent reinforcement learning (MARL). Cong et al.[25] designed and implemented RL-based routing schemes combined with multi-optimality routing criteria (RLR-M) for CPS, which is more flexible than traditional strategies. Yan et al.[26] introduced a Q-learning based approach to analyze the vulnerability of CPS under sequential topology attacks, taking into account the physical system behaviors, demonstrating the effectiveness and feasibility of Q-learning in enhancing the cyber security of CPS, despite restricted information on opponents[27, 28]. While these approaches address some of the challenges associated with applying RL to CPS and make some progress, few have considered the unequal costs caused by strategy selection errors due to the severity of the consequences of misclassification, which is an essential aspect that we include in our proposed payoff setting.

3 Model

3.1 Attack topology

We present a model of a CPS network layer composed of a firewall, web server, client-server, database server, and file server FTP, as well as a physical layer consisting of physical devices and control components that require monitoring[29]. The attack scenario considered involves the exploitation of vulnerabilities by an adversary to penetrate the firewall and compromise the web server, resulting in consequential damage to the client-server, database server, and file server FTP. The illustration provided in Fig. 2 demonstrates the components of the CPS cyber and physical layers.

Assuming an attacker with minimal access privileges initiates an attack on a CPS system from outside the firewall, their ultimate objective is to gain root access to the database to obtain or destroy sensitive information, which plays a crucial role in many fields[30], such as recommendation[31−35], prediction[36, 37], correlation mining[38, 39], etc. The CPS system state space comprises nine distinct states, each representing a unique system configuration, and the transition process between these states is graphically depicted in Fig. 3. The attack-defense game between the attacker and defender progresses through the state space s_1 until the final state s_9 is reached, at which

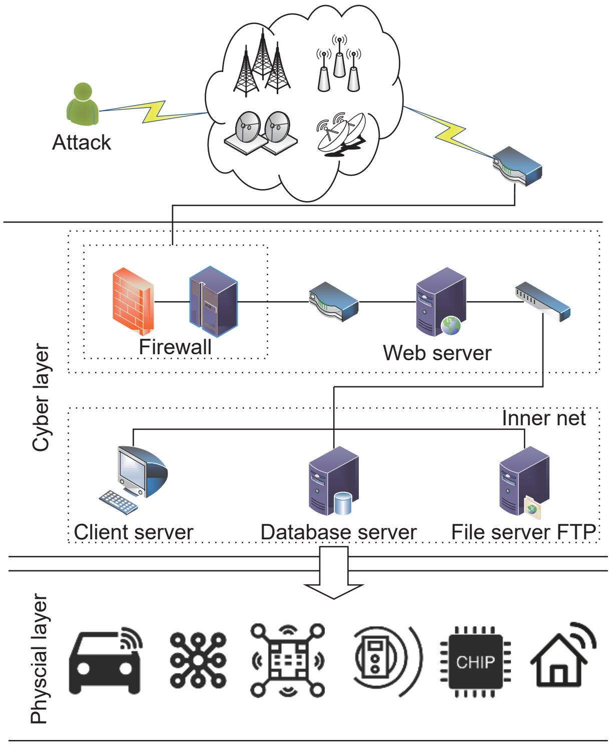


Fig. 2 Topology of the attack process for CPS.

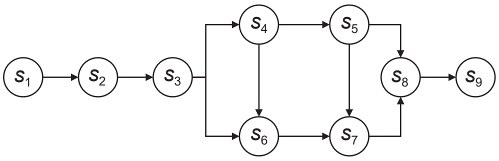


Fig. 3 Cyber system state transition process topology diagram.

point the attacker's goal is either achieved, or the defense has successfully thwarted the attacker's efforts to breach the system's security[19].

3.2 State transitions

Let A denote the set of attack actions available to the attacker in a given stage S_t, and let D denote the set of defense actions available to the defender. In the event that the attacker attempts to invade with an attack strategy a_i, and the defender plans to defend with defense strategy d_j. Then, the probability of successfully transitioning to a new stage S_k can be expressed as Eq. (1).

$$p\left(S_k|S_t,a_i,d_j\right)=\varepsilon\left(S_t,a_i,d_j\right) \tag{1}$$

where $\varepsilon\left(S_t,a_i,d_j\right)$ indicates the stage where the attacker has compromised S_t, and $\varepsilon\left(S_t,a_i,d_j\right)$ is the probability of success of attack, where $\varepsilon\left(S_t,a_i,d_j\right)=1$ signifies that the attack is successfully blocked, and $\varepsilon\left(S_t,a_i,d_j\right)=0$ when it is invalid[20, 40].

3.3 Payoffs

Unequal cost is a concept that refers to the fact that different policy errors may result in different levels of system losses. For instance, consider a cybersecurity scenario, where attackers may attempt to steal sensitive data or cause system failures. In such cases, the losses incurred due to an erroneous defense strategy that incorrectly identifies an attack are significantly greater than the losses that would arise from a failure to detect an attack and the subsequent decision not to defend against it. When dealing with this type of situation, it is important not only to consider the overall number of errors that may occur during the defense process but also to take into account the cost associated with each error. By incorporating unequal cost, the potential cost of each decision error in a CPS can be captured more accurately and the goal is shifted to identify a defense strategy that can minimize the total cost of the errors. This can help to improve the resilience and security of CPS, even in highly dynamic and challenging environments.

Definition 1 System loss $SL\left(S_t,a_i,d_j\right)$ refers to the loss to the system when strategy d_j cannot prevent strategy a_i. $SL\left(S_t,a_i,d_j\right)$ is usually represented by the degree of criticality of target resources (C_t), system authority loss (SAL_t), attack lethality ($AL(a_i)$), and security attribute damage (SAD_t), where C_t reflects the significance of the targeted resource within the overall system, SAL_t quantifies the impact of different levels of cyber access permissions, including access, guest, and root, on system loss, $AL(a_i)$ assesses the level of harm caused by an attack and evaluates its impact on the system, and SAD_t captures the level of damage to cyber security, which encompasses critical aspects such as confidentiality, integrity, and availability[41]. $SL\left(S_t,a_i,d_j\right)$ follows as Eq. (2).

$$SL\left(S_t,a_i,d_j\right)=C_t\times SAL_t\times AL(a_i)\times SAD_t \tag{2}$$

Definition 2 Defend cost (DC) refers to the cost required for the defender to take strategy d_j.

Definition 3 Defense effectiveness $\left(\mu\left(S_t,a_i,d_j\right)\right.$

$\left(0 \leqslant \mu\left(S_t, a_i, d_j\right) < 1\right)\Big)$ denotes the effectiveness of defense, which means how much system loss will be reduced by implementing strategy d_j.

Definition 4 Unequal cost ($C_{i,j}$) denotes the cost incurred due to misclassifying a_i as a_j. There are three types of misclassification scenarios in CPS, namely false acceptance $\lambda_{\rm A,NA}$ (mis-recognizing an effective attack as an ineffective attack), false rejection $\lambda_{\rm NA,A}$ (mis-recognizing an ineffective attack as an effective attack), and false identification $\lambda_{\rm A,A}$ (mis-recognizing between two ineffective attacks), which result in different system losses. Drawing on the relevant concepts of cost-sensitive classification decisions in machine learning, we assume that $\lambda_{\rm A,NA} > \lambda_{\rm A,A} > \lambda_{\rm NA,A}$ in the CPS cyber security defense process. $C_{i,j}$ follows as Formula (3).

$$C_{i,j} = \begin{cases} SL(a_i), \text{ if } a_i = a_j; \\ \lambda_{\rm A,NA} SL(a_i), \text{ if } a_i \text{ is effective and } a_j \text{ is ineffective}; \\ \lambda_{\rm NA,A} SL(a_i), \text{ if } a_i \text{ is ineffective and } a_j \text{ is effective}; \\ \lambda_{\rm A,A} SL(a_i), \text{ if } a_i \text{ and } a_j \text{ are ineffective} \end{cases} \tag{3}$$

Definition 5 Misclassification probability (θ) refers to the probability of misclassifying a_i as a_j.

The formula for calculating total defense gain at the stage t is shown in Formula (4).

$$R_D^t\left(S_t, a_i, d_j\right) = \begin{cases} \left(1-\mu\left(S_t, a_i, d_j\right)\right) \times SL\left(S_t, a_i, d_j\right) + DC\left(d_j\right), \\ \quad \text{if random } p \leqslant \theta; \\ \left(1-\mu\left(S_t, a_i, d_j\right)\right) \times C_{i,j} + DC\left(d_j\right), \\ \quad \text{if random } p > \theta \end{cases} \tag{4}$$

4 Solution of the model

Reinforcement Learning, a branch of machine learning, has been widely applied in security strategy analysis in various domains such as the Internet of vehicles (IoV)[42], mobile social networks (MSNs)[43], vulnerability analysis of smart grid[26], etc. In general, one of the most difficult tasks of the security of CPS is to accurately specify the model parameters, given the limited availability of domain knowledge[44]. Q-learning, proposed by Watkins, is one of the most popular RL methods that seek efficient control policies without the knowledge of an explicit system model, therefore minimizing the effort associated with simulating and solving such complex environments[45]. Furthermore, the characteristic of trial and error in the environment for Q-learning is particularly adaptable and useful in real-time and adversarial environments. Q-learning exhibits excellent suitability for cyber security applications where cyber-attacks are increasingly sophisticated, rapid, and ubiquitous[46, 47]. Its basic process is graphically depicted in Fig. 4.

Bellman equation is the core of Q-learning to update the Q-value. Its core idea is that when making decisions, we consider not only the immediate reward of the current decision but also the future sustainable reward, which is generated by it. The updated formula for the Q-value is shown in Eq. (5).

$$\begin{aligned} Q\left(s_t, a_x, d_y\right) \leftarrow (1-\alpha) Q\left(s_t, a_i, d_j\right) + \\ \alpha\left[R_D^t + \gamma \max_{d \in D} Q(s_{t+1}, a, d)\right] \end{aligned} \tag{5}$$

where $Q\left(s_t, a_x, d_y\right)$ denotes the Q-value when the state is s_t, attack strategy is a_x, and defend strategy is d_y. R_D^t denotes the immediate reward when the state transitions. $\max\limits_{d \in D} Q(s_{t+1}, a, d)$ denotes the max value in the Q-value table at state s_{t+1}. α denotes the learning rate that is often used to balance the aggressiveness and conservatism of the algorithm, where $\alpha = 1$ signifies that the agent will primarily learn from current and future rewards, potentially resulting in forgetting previously learned knowledge and divergence, and $\alpha = 0$ means that the agent has no learning ability. γ denotes decay factor.

Based on the above analysis, we design the optimal defense strategy selection algorithm for the cyber security of CPS (see Algorithm 1).. This algorithm incorporates unequal cost into Q-learning, enabling the training of a robust model capable of effectively defending against attacks, even in the presence of

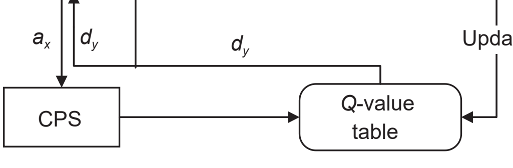


Fig. 4 Basic process of Q-learning.

Algorithm 1 Defense strategy selection algorithm based on Q-learning under unequal cost

```
Initialize:
    Let the discount factor $\gamma$, the learning rate $\alpha = 1$, $Q(S_t, a_0, a_1, \ldots, a_n) = 0$.
    Give the matrix $\mu\left(S_t, a_i, d_j\right)$, the attack set $A = \{a_1, a_2, \ldots, a_n\}$, the defense set $D = \{d_1, d_2, \ldots, d_m\}$.
Choose an action:
    Attacker：return an action $a_i$ in the stage $S_t$ uniformly at random.
    Defender：return an action uniformly at random with an exploring probability, otherwise return $d_j$ according to the $Q$ table.
Learn:
    (1) After receiving a reward $R_D^t$ in the stage $S_t$ by action $a_i$ and $d_j$.
    (2) Update:
$Q\left(s_t, a_x, d_y\right) \leftarrow (1-\alpha) Q\left(s_t, a_i, d_j\right) + \alpha\left[R_D^t + \gamma \max_{d \in D} Q(s_{t+1}, a, d)\right]$
    (3) $\alpha = \alpha \times decay$
```

misclassifications. By assigning different costs to misclassifications, the defender will promote a more cautious approach, which means that the defender takes into account the potential for misclassifications when making decisions on defense strategies. The goal shifts from solely maximizing the success rate of defense to minimizing the overall cost, considering the potential consequences and expenses associated with misclassifications.

5 Experiment and discussion

To demonstrate the effectiveness and feasibility of our approach, we conducted a simulation experiment on a CPS with misclassification interference. Compared to the performance of the proposed algorithm with traditional Q-learning that does not consider unequal costs, the results indicate that the proposed method significantly improves the efficiency and effectiveness of decision-making in CPS. Specifically, our approach outperforms traditional Q-learning approaches in terms of reducing the impact of errors in strategy selection and total system loss. The experiment is implemented using Python 3.9, with the help of popular data analysis libraries such as Numpy and Pandas.

5.1 Data summary

The CPS state transition process is a complex and dynamic process that involves the movement of a system through various states based on its internal and external conditions. Table 1 provides a detailed description of the various states involved in the CPS state transition process.

In CPS cybersecurity, the attacker and defender strategies are crucial in determining the outcome of a security situation. Tables 2 and 3 provide a detailed description of the various attacker and defender strategies respectively [40, 48].

Q^* table plays a crucial role in the process of defense strategy selection. The defense strategy is selected according to the Q^* table. Here, we give the structure of the Q table. The row index of the Q table is the action a_i in the stage S_t, and the column index is $\{d_0, d_1, \ldots, d_{15}\}$, which denotes the set of defense strategies for the defender at each state. Table 4 displays the attack strategies for each stage and the

Table 1 CPS system state attributes table.

State	State description	*SAD*	*C*	*SAL*
s_1	All nodes are in the normal state	0	0	0
s_2	Firewall root permission is obtained	20	2	0.9
s_3	Web server guest permission is obtained	40	4	0.6
s_4	Client server guest permission is obtained	30	3	0.6
s_5	Client server root permission is obtained	30	3	0.9
s_6	File server FTP guest permission is obtained	30	3	0.6
s_7	File server FTP root permission is obtained	30	3	0.9
s_8	Date server guest permission is obtained	50	5	0.6
s_9	Date server root permission is obtained	50	5	0.9

Table 2 Attack action and attribute description table.

Attack strategy	**Strategy description**	***AL***
a_0	Not attack	0
a_1	Steal account and crack it	10
a_2	Oracle TNS listener	4
a_3	Install trojan	5
a_4	LPC to LSASS process	4
a_5	Shutdown server tenor	6
a_6	THS chunk overflow	7
a_7	Attack address blacklist	5
a_8	Install SQL Listener program	10
a_9	FTP rhost attack	12
a_{10}	Shutdown database server	8

Table 3 Defend action and attribute description table.

Defense strategy	Strategy description	*DC*
d_0	Not defend	0
d_1	Delete account + random frequency	185
d_2	Port enlarging + IP enlarging	155
d_3	Reinstall listener + fixed frequency	135
d_4	Protocol changing+ random frequency	160
d_5	Routing enlarging + fixed frequency	150
d_6	Uninstall Trojan	80
d_7	Protocol changing + IP hopping	65
d_8	Add address blacklist	110
d_9	Storage enlarging	85
d_{10}	IP enlarging + IP hopping	70
d_{11}	Restart database	100
d_{12}	Storage enlarging + fixed frequency	75
d_{13}	Patch SSH on FTP	115
d_{14}	IP enlarging + fixed frequency	90
d_{15}	Storage enlarging + random frequency	70

corresponding matrix of defense strategy effectiveness.

5.2 Comparative analysis for strategy between equal cost and unequal cost

Under the condition of equal cost, where the system losses of false positives and false negatives are equivalent, the defender will adopt d_0 regardless of whether mis-recognizing an effective attack is misclassified as an ineffective attack. And a strategy of good defensive effectiveness will be adopted regardless of whether mis-recognizing an effective attack as the other effective attack or an ineffective attack as an effective attack. However, under the case of equal cost, the defender implements preemptive defense measures even when there is no apparent attack to ensure cyber security. When an attack is detected, the defender considers two possible scenarios: (1) whether the effective attack can be recognized as the other effective attack, and (2) whether the ineffective attack can be misclassified as an effective attack. Based on this assessment, the defender selects a defense strategy that is most likely to achieve a high success rate across all possible scenarios. Overall, this algorithm represents a promising approach to improving the robustness and resilience of machine learning models against adversarial attacks. Set $\lambda_{\text{A,NA}} = 10$, $\lambda_{\text{A,A}} = 5$, $\lambda_{\text{NA,A}} = 1$, $\theta = 0.2$ to train the model. The defense strategy in the case of unequal cost is shown in Table 5.

Table 4 Effectiveness of attack strategies and defense strategies in each stage.

Stage	State transition	Defense strategy effectiveness $\mu\left(S_t,a_i,d_j\right)$
S_1	$s_1 \to s_2$	$\begin{array}{cccc} & d_1 & d_2 & d_3 \\ a_1 & 0.2 & 0.6 & 0.5 \\ a_2 & 0.8 & 0.3 & 0.6 \end{array}$
S_2	$s_2 \to s_3$	$\begin{array}{cccc} & d_4 & d_5 & d_6 \\ a_3 & 0.7 & 0.5 & 0.6 \\ a_4 & 0.8 & 0.7 & 0.5 \end{array}$
S_3	$s_3 \to s_4$	$\begin{array}{cccc} & d_4 & d_6 & d_8 \\ a_3 & 0.7 & 0.6 & 0.1 \\ a_7 & 0.4 & 0.6 & 0.8 \end{array}$
S_4	$s_3 \to s_6$	$\begin{array}{cccc} & d_7 & d_8 & d_9 \\ a_4 & 0.7 & 0.5 & 0.3 \\ a_7 & 0.5 & 0.8 & 0.6 \end{array}$
S_5	$s_4 \to s_5$	$\begin{array}{cccc} & d_1 & d_2 & d_3 \\ a_1 & 0.2 & 0.6 & 0.5 \\ a_2 & 0.8 & 0.3 & 0.6 \end{array}$
S_6	$s_4 \to s_6$	$\begin{array}{cccc} & d_7 & d_8 & d_9 \\ a_4 & 0.7 & 0.5 & 0.3 \\ a_7 & 0.5 & 0.8 & 0.6 \end{array}$
S_7	$s_5 \to s_7$	$\begin{array}{cccc} & d_9 & d_{10} & d_{11} \\ a_5 & 0.6 & 0.3 & 0.4 \\ a_6 & 0.6 & 0.3 & 0.7 \end{array}$
S_8	$s_5 \to s_8$	$\begin{array}{cccc} & d_1 & d_9 & d_{11} \\ a_2 & 0.8 & 0.3 & 0.5 \\ a_6 & 0.3 & 0.6 & 0.7 \end{array}$
S_9	$s_6 \to s_7$	$\begin{array}{cccc} & d_9 & d_{10} & d_{11} \\ a_5 & 0.6 & 0.3 & 0.4 \\ a_6 & 0.6 & 0.3 & 0.7 \end{array}$
S_{10}	$s_7 \to s_8$	$\begin{array}{cccc} & d_6 & d_{12} & d_{13} \\ a_2 & 0.6 & 0.7 & 0.8 \\ a_9 & 0.1 & 0.5 & 0.7 \end{array}$
S_{11}	$s_8 \to s_9$	$\begin{array}{cccc} & d_4 & d_{14} & d_{15} \\ a_8 & 0.6 & 0.5 & 0.1 \\ a_{10} & 0.2 & 0.6 & 0.5 \end{array}$

5.3 Comparative analysis for system loss between equal cost and unequal cost

To verify the feasibility of our algorithm for a dynamic defense environment, we perform a simulation of 1000 attacks and defenses under the condition of equal cost and unequal cost and then compare the system losses.

From Fig. 5, it can be observed that under the case of unequal cost, strategies incur certain defense costs to prevent misclassification errors, resulting in slightly higher system losses compared to the case of equal cost. However, once a misclassification error occurs, the system loss under the case of equal costs becomes difficult to recover, and it is much higher than the loss under the case of unequal costs.

Table 5 Defense action selection under the case of unequal cost.

Stage	Attack strategy	Defense strategy under the case of unequal cost
s_1	a_0	d_3
	a_1	d_2
	a_2	d_1
s_2	a_0	d_4
	a_3	d_4
	a_4	d_4
s_3	a_0	d_6
	a_3	d_4
	a_7	d_8
s_4	a_0	d_8
	a_4	d_7
	a_7	d_8
s_5	a_0	d_3
	a_1	d_2
	a_2	d_1
s_6	a_0	d_8
	a_4	d_7
	a_7	d_8
s_7	a_0	d_{11}
	a_5	d_9
	a_6	d_{11}
s_8	a_0	d_{11}
	a_2	d_1
	a_6	d_{11}
s_9	a_0	d_9
	a_5	d_9
	a_6	d_{11}
s_{10}	a_0	d_{13}
	a_2	d_{13}
	a_9	d_{13}
s_{11}	a_0	d_{14}
	a_8	d_4
	a_{10}	d_{14}

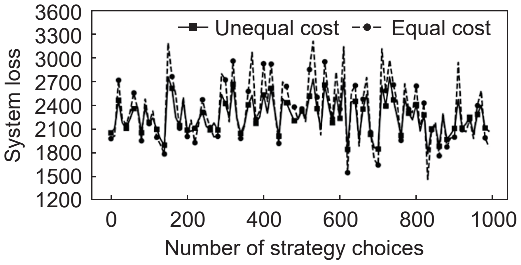


Fig. 5 Comparison of system loss related to strategy selection.

6 Conclusion and future work

This paper proposes a modified Q-learning based strategy analysis approach to address the unequal cost challenges posed by misclassification errors in CPS, aiming to optimize defense strategy selection in CPS. By introducing the concept of unequal cost into the Q-learning algorithm, we demonstrated that a more efficient and low-cost cyber security defense strategy can be obtained in the presence of misclassification interference. The proposed algorithm can be implemented in various CPS applications, including manufacturing, transportation, and healthcare systems. However, the specific form of the cost function may vary depending on the particular environments and systems being analyzed. Careful analysis and tuning are required to ensure that the weighting scheme accurately reflects the true error costs associated with each strategy in the CPS. Future research can be directed to identifying the optimal values of the parameters in the application field of IoT, and to explore how the proposed algorithm can be applied to optimize multiple objectives simultaneously, such as minimizing system loss while maximizing resource utilization.

References

[1] X. Zhou, W. Liang, K. Yan, W. Li, K. I. K. Wang, J. Ma, and Q. Jin, Edge-enabled two-stage scheduling based on deep reinforcement learning for Internet of everything, *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3295–3304, 2022.

[2] Y. Yang, X. Yang, M. Heidari, M. A. Khan, G. Srivastava, M. Khosravi, and L. Qi, ASTREAM: data-stream-driven scalable anomaly detection with accuracy guarantee in IIoT environment, *IEEE Trans. Netw. Sci. Eng.*, doi: 10.1109/TNSE.2022.3157730.

[3] X. Zhou, Y. Hu, J. Wu, W. Liang, J. Ma, and Q. Jin, Distribution bias aware collaborative generative adversarial network for imbalanced deep learning in industrial IoT, *IEEE Trans. Ind. Inform.*, vol. 19, no. 1, pp. 570–580, 2022.

[4] G. Hatzivasilis, I. Papaefstathiou, and C. Manifavas, SCOTRES: secure routing for IoT and CPS, *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2129–2141, 2017.

[5] X. Zhou, X. Xu, W. Liang, Z. Zeng, S. Shimizu, L. T. Yang, and Q. Jin, Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems, *IEEE Trans. Ind. Inform.*, vol. 18, no. 2,

pp. 1377–1386, 2021.

[6] A. R. Sadeghi, C. Wachsmann, and M. Waidner, Security and privacy challenges in industrial Internet of things, in *Proc. 52nd Annu. Design Automation Conf.*, San Francisco, CA, USA, 2015, pp. 1–6.

[7] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, and K. I. K. Wang, Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system, *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9310–9319, 2021.

[8] W. Liang, Y. Hu, X. Zhou, Y. Pan, and K. I. K. Wang, Variational few-shot learning for microservice-oriented intrusion detection in distributed industrial IoT, *IEEE Trans. Ind. Inform.*, vol. 18, no. 8, pp. 5087–5095, 2021.

[9] F. O. Olowononi, D. B. Rawat, and C. Liu, Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS, *IEEE Commun. Surv. Tutor.*, vol. 23, no. 1, pp. 524–552, 2021.

[10] K. C. Lalropuia and V. Gupta, Modeling cyber-physical attacks based on stochastic game and Markov processes, *Reliab. Eng. Syst. Saf.*, vol. 181, pp. 28–37, 2019.

[11] Z. Jin, S. Zhang, Y. Hu, Y. Zhang, and C. Sun, Security state estimation for cyber-physical systems against DoS attacks via reinforcement learning and game theory, *Actuators*, vol. 11, no. 7, p. 192, 2022.

[12] K. Huang, C. Zhou, Y. Qin, and W. Tu, A game-theoretic approach to cross-layer security decision-making in industrial cyber-physical systems, *IEEE Trans. Ind. Electron.*, vol. 67, no. 3, pp. 2371–2379, 2020.

[13] H. Kou, H. Liu, Y. Duan, W. Gong, Y. Xu, X. Xu, and L. Qi, Building trust/distrust relationships on signed social service network through privacy-aware link prediction process, *Appl. Soft Comput.*, vol. 100, p. 106942, 2021.

[14] X. Zhou, Y. Li, and W. Liang, CNN-RNN based intelligent recommendation for online medical pre-diagnosis support, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 18, no. 3, pp. 912–921, 2021.

[15] L. Kong, G. Li, W. Rafique, S. Shen, Q. He, M. R. Khosravi, R. Wang, and L. Qi, Time-aware missing healthcare data prediction based on ARIMA model, *IEEE/ACM Trans. Comput. Biol. And Bioinf.*, doi: 10.1109/TCBB.2022.32050642.

[16] P. Sterner, D. Goretzko, and F. Pargent, Everything has its price: Foundations of cost-sensitive learning and its application in psychology, https://doi.org/10.31234/osf.io/7asgz, 2021.

[17] C. Y. T. Ma, N. S. V. Rao, and D. K. Y. Yau, A game theoretic study of attack and defense in cyber-physical systems, in *Proc. 2011 IEEE Conf. Computer Communications Workshops* (*INFOCOM WKSHPS*), Shanghai, China, 2011, pp. 708–713.

[18] H. Orojloo and M. A. Azgomi, A game-theoretic approach to model and quantify the security of cyber-physical systems, *Comput. Ind.*, vol. 88, pp. 44–57, 2017.

[19] S. Huang, H. Zhang, J. Wang, and J. Huang, Markov differential game for network defense decision-making method, *IEEE Access*, vol. 6, pp. 39621–39634, 2018.

[20] Y. Sun, W. Ji, J. Weng, B. Zhao, Y. Li, and X. Wu, Selection of optimal strategy for moving target defense based on signal game, in *Proc. 2020 Int. Conf. Cyberspace Innovation of Advanced Technologies*, Guangzhou, China, 2020, pp. 28–32.

[21] Y. Guo, Y. Gong, L. L. Njilla, and C. A. Kamhoua, A stochastic game approach to cyber-physical security with applications to smart grid, in *Proc. IEEE INFOCOM 2018 - IEEE Conf. Computer Communications Workshops* (*INFOCOM WKSHPS*), Honolulu, HI, USA, 2018, pp. 33–38.

[22] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart IoT, *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12588–12596, 2021.

[23] K. Gai and M. Qiu, Optimal resource allocation using reinforcement learning for IoT content-centric services, *Appl. Soft Comput.*, vol. 70, pp. 12–21, 2018.

[24] J. Khoury and M. Nassar, A hybrid game theory and reinforcement learning approach for cyber-physical systems security, in *Proc. NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symp.*, Budapest, Hungary, 2020, pp. 1–9.

[25] P. Cong, Y. Zhang, Z. Liu, T. Baker, H. Tawfik, W. Wang, K. Xu, R. Li, and F. Li, A deep reinforcement learning-based multi-optimality routing scheme for dynamic IoT networks, *Comput. Netw.*, vol. 192, p. 108057, 2021.

[26] J. Yan, H. He, X. Zhong, and Y. Tang, Q-learning-based vulnerability analysis of smart grid against sequential topology attacks, *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 1, pp. 200–210, 2017.

[27] K. Chung, C. A. Kamhoua, K. A. Kwiat, Z. T. Kalbarczyk, and R. K. Iyer, Game theory with learning for cyber security monitoring, in *Proc. 2016 IEEE 17th Int. Symp. High Assurance Systems Engineering* (*HASE*), Orlando, FL, USA, 2016, pp. 1–8.

[28] S. Shiva, S. Roy, and D. Dasgupta, Game theory for cyber security, in *Proc. 6th Annu. Workshop on Cyber Security and Information Intelligence Research*, Oak Ridge, TN, USA, 2010, pp. 1–4.

[29] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, Siamese

neural network based few-shot learning for anomaly detection in industrial cyber-physical systems, *IEEE Trans. Ind. Inform.*, vol. 17, no. 8, pp. 5790–5798, 2020.

[30] X. Zhou, W. Liang, K. I. K. Wang, R. Huang, and Q. Jin, Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data, *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 1, pp. 246–257, 2018.

[31] H. Liu, H. Kou, C. Yan, and L. Qi, Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph, *Complexity*, vol. 2020, pp. 1–15, 2020.

[32] H. Kou, J. Xu, and L. Qi, Diversity-driven automated web API recommendation based on implicit requirements, *Appl. Soft Comput.*, vol. 136, p. 110137, 2023.

[33] W. Gong, X. Zhang, Y. Chen, Q. He, A. Beheshti, X. Xu, C. Yan, and L. Qi, DAWAR: diversity-aware web APIs recommendation for mashup creation based on correlation graph, in *Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Madrid, Spain, 2022, pp. 395–404.

[34] L. Qi, W. Lin, X. Zhang, W. Dou, X. Xu, and J. Chen, A correlation graph based approach for personalized and compatible web APIs recommendation in mobile APP development, *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5444–5457, 2023.

[35] Y. Liu, H. Wu, K. Rezaee, M. R. Khosravi, O. I. Khalaf, A. Ali Khan, D. Ramesh, and L. Qi, Interaction-enhanced and time-aware graph convolutional network for successive point-of-interest recommendation in traveling enterprises, *IEEE Trans. Ind. Inf.*, vol. 19, no. 1, pp. 635–643, 2023.

[36] H. Liu, H. Kou, C. Yan, and L. Qi, Link prediction in paper citation network to construct paper correlation graph, *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 1, pp. 1–12, 2019.

[37] Y. Xu, Z. Feng, X. Zhou, M. Xing, H. Wu, X. Xue, S. Chen, C. Wang, and L. Qi, Attention-based neural networks for trust evaluation in online social networks, *Inf. Sci.*, vol. 630, pp. 507–522, 2023.

[38] X. Zhou, X. Yang, J. Ma, and K. I. K. Wang, Energy-efficient smart routing based on link correlation mining for wireless edge computing in IoT, *IEEE Internet Things J.*, vol. 9, no. 16, pp. 14988–14997, 2022.

[39] X. Zhou, W. Liang, K. I. K. Wang, and L. T. Yang, Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations, *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 171–178, 2021.

[40] H. Maleki, S. Valizadeh, W. Koch, A. Bestavros, and M. van Dijk, Markov modeling of moving target defense games, in *Proc. 2016 ACM Workshop on Moving Target Defense*, Vienna, Austria, 2016, pp. 81–92.

[41] M. Azadi, H. Zare, and M. J. Zare, Confidentiality, integrity and availability in electronic health records: An integrative review, in *Proc. 15th Int. Conf. Information Technology – New Generations*, Las Vegas, NV, USA, 2018, 745–748.

[42] Z. Ning, P. Dong, X. Wang, L. Guo, J. J. P. C. Rodrigues, X. Kong, J. Huang, and R. Y. K. Kwok, Deep reinforcement learning for intelligent Internet of vehicles: An energy-efficient computational offloading scheme, *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1060–1072,Dec, 2019.

[43] Q. Xu, Z. Su, and R. Lu, Game theory and reinforcement learning based secure edge caching in mobile social networks, *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3415–3429, 2020.

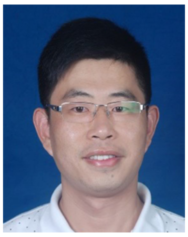
[44] J. Khoury and M. Nassar, A hybrid game theory and reinforcement learning approach for cyber-physical systems security, in *Proc. 2020 IEEE/IFIP Network Operations and Management Symp.*, Budapest, Hungary, 2020, pp. 1–9.

[45] A. Uprety and D. B. Rawat, Reinforcement learning for IoT security: A comprehensive survey, *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8693–8706, 2021.

[46] C. J. C. H. Watkins and P. Dayan, Q-learning, *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[47] T. T. Nguyen and V. J. Reddi, Deep reinforcement learning for cyber security, *IEEE Trans. Neural Netw. Learn. Syst.* ,doi: 10.1109/TNNLS.2021.3121870.

[48] J. L. Tan, C. Lei, H. Q. Zhang, and Y. Q. Cheng, Optimal strategy selection approach to moving target defense based on Markov robust game, *Comput. Secur.*, vol. 85, pp. 63–76, 2019.



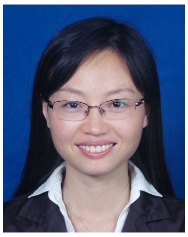
Xin Chen obtained the PhD degree from Nanjing University of Aeronautics and Astronautics, China, in 2007. He is currently an associate professor at Nanjing University of Finance and Economics, China. His main research interests include data analysis and decision support.

Jixiang Cheng obtained the BS degree from Nanjing University of Finance and Economics, China, in 2022. He is currently pursuing the MS degree at Nanjing University of Finance and Economics. His current research interests include reinforcement learning and cyber security.

Luanjuan Jiang is currently pursuing the PhD degree in soft engineering at Nanjing University of Science and Technology, China.

Qianmu Li obtained the PhD degree from Nanjing University of Science and Technology, China, in 2005. He was elected as a foreign academician of the Russian Academy of Sciences in 2021. He is currently a professor at Nanjing University of Science and Technology and the director of the Information Construction and Management Office. His current research interests include artificial intelligence security, industrial Internet of things (IIoT) security, and data security.

Ting Wang obtained the PhD degree from Nanjing University, China, in 2018. She is currently a lecturer at Nanjing University of Finance and Economics, China. Her current research interests include operations research and data-driven robust optimization.



Dafang Li obtained the PhD degree from Tongji University, China, in 2013. She is currently a lecturer at Nanjing University of Finance and Economics, China. Her current research interests include information management and quality and safety management