A Learning-Based Solution for an Adversarial Repeated Game in Cyber–Physical Power Systems

Shuva Paul, *Member, IEEE*, Zhen Ni, *Member, IEEE*, and Chaoxu Mu, *Senior Member, IEEE*

***Abstract*— Due to the rapidly expanding complexity of the cyber–physical power systems, the probability of a system malfunctioning and failing is increasing. Most of the existing works combining smart grid (SG) security and game theory fail to replicate the adversarial events in the simulated environment close to the real-life events. In this article, a repeated game is formulated to mimic the real-life interactions between the adversaries of the modern electric power system. The optimal action strategies for different environment settings are analyzed. The advantage of the repeated game is that the players can generate actions independent of the previous actions' history. The solution of the game is designed based on the reinforcement learning algorithm, which ensures the desired outcome in favor of the players. The outcome in favor of a player means achieving higher mixed strategy payoff compared to the other player. Different from the existing game-theoretic approaches, both the attacker and the defender participate actively in the game and learn the sequence of actions applying to the power transmission lines. In this game, we consider several factors (e.g., attack and defense costs, allocated budgets, and the players' strengths) that could affect the outcome of the game. These considerations make the game close to real-life events. To evaluate the game outcome, both players' utilities are compared, and they reflect how much power is lost due to the attacks and how much power is saved due to the defenses. The players' favorable outcome is achieved for different attack and defense strengths (probabilities). The IEEE 39 bus system is used here as the test benchmark. Learned attack and defense strategies are applied in a simulated power system environment (PowerWorld) to illustrate the postattack effects on the system.**

***Index Terms*— Adversarial game, Markov decision process (MDP), reinforcement learning (RL), repeated game, smart grid (SG) security.**

NOMENCLATURE

S	Status of the transmission lines.
k	Order of the contingency or attack.
t	Attack timescale.
S_A	Attacker's target set. A predefined set of power system transmission lines.
S_D	Defender's target set. A predefined set of power system transmission lines.
G	Contraction mapping.
X	Maximum time step.
H	Value iteration operator.
$N(S_A)$	Size of the attacker's target set.
$N(S_D)$	Size of the defender's target set.
TLC	Total loading capacity of the system (MW).
C_w	Cost of not attacking.
C_a	Cost of attacking (consumed resource from the system).
C_m	Cost of defending (holds the same property as C_a).
B_A	Attacker's allocated budget (defined by a number of actions).
B_D	Defender's allocated budget (holds the same property as B_D).
σ	Attacker's probability of attack or strength.
δ	Defender's strength (the probability of defending).
U_A	Attacker's mixed strategy payoff (contains the mixed strategy payoffs for multiple repetitions).
U_D	Defender's mixed strategy payoff (holds the same property as the U_A).
U_a	Attacker's immediate mixed strategy payoffs in the repetitions.
U_d	Defender's immediate mixed strategy payoffs in the repetitions.
C_i	Immediate damage caused by the attack, the unit is in MW.
P_a	Total loss by the attack (MW).
ϵ	Exploration rate.
γ	Discount factor of long- or short-term reward.
F	Total number of runs in the game.
Q	Quality of the state s, associated with action a and d.
T	Transition function between the states.
$V_a(s')$	Value at the next state s' for the attacker.
$V_d(s')$	Value at the next state s' for the defender.
Pr	Probability of a state-action pair (s, a, d).
C	Total number of times the state s is visited for the specific action a, and d.
π	Action selection probability.
Π	Repeated game function.

Manuscript received January 27, 2019; revised July 10, 2019 and November 2, 2019; accepted November 19, 2019. This work was supported in part by the National Science Foundation under Grant 1949921 and Grant 1833005. The work of C. Mu was supported by the National Science Foundation of China under Grant 61773284. *(Corresponding author: Zhen Ni.)*

S. Paul is with the Electrical Engineering and Computer Science Department, South Dakota State University, Brookings, SD 57007 USA (e-mail: shuva.paul@jacks.sdstate.edu).

Z. Ni is with the Department of Computer, Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431 USA (e-mail: zhenni@fau.edu).

C. Mu is with the Tianjin Key Laboratory of Process Measurement and Control, School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: cxmu@tju.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2955857

R	Reward assigned for action a and d.
Rep	Repetition of a repeated game.
a	Action taken by the attacker.
d	Action taken by the defender.
Z	Represents active status of a line l at time t.

I. Introduction

MACHINE learning methods are becoming popular because of the increasing complexity of the critical infrastructures and their exposure to the security threats. Several machine learning (ML) algorithms are being used for detecting events and anomalies, finding critical elements of a power system, and implementing the interactions between the adversaries in a power system. These ML algorithms include classification and clustering methods, reinforcement learning (RL) algorithms, artificial neural network, and so on [1]–[3]. Securing the modern electric power grid is one of the most challenging issues nowadays. The modern electric power grid or the smart grid (SG) is an interconnected network of a large number of heterogeneous devices. These complex interconnections of SG possibly expose the whole network to severe security threats. High integration of information and communication technologies (cyber layer) with the SG (physical layer) results in a complex and efficient cyber–physical power system (CPPS) [4]–[8].

The interaction between the authority and its adversary can be treated as a game. Game theory helps the power system operators to understand the attackers' motives and moves, and thus they can strengthen the security of the CPPS. Many research works are going on for improving the stability, security, and resiliency of SG under different uncertain conditions [9]–[16]. Different forms of games are formulated in the CPPS. One-shot attack [17], multistage attack [18], simultaneous attack [19], sequential attack [20], malicious false data injection attack [21], and coordinated attack are different types of attack schemes applied to the power grid. Proper solutions to the SG security game are important for the authorities to increase the protection and avoid the damages. In recent literature, various neural networks structures, approximate/adaptive dynamic programming, RL, and deep RL show the potentials to solve such games and provide promising learning and optimization performance [22]–[29].

Pourahmadi *et al.* [30] used a static game-theoretic solution concept to represent coalition formation game theory in assessing the components' criticality for a system's overall reliability. Wei *et al.* [31], Liao *et al.* [32], and Wang *et al.* [33] implemented a stochastic two-player zero-sum game for power grid protection against malicious attacker. Q-learning was applied to find the optimal policies for the attacker. Ni and Paul [18] studied a multistage game between the adversaries that used a unique utility function to find the optimal attack sequences from the attacker's perspective. In this work, the defender's action was passive and predefined throughout the learning process of the attacker. Wang *et al.* [34] modeled a strategic honeypot game for distributed denial-of-service (DoS) attacks in the SG. They analyzed the interactions between the attacker and the defender in the SG communication network. Farraj *et al.* [35] used an iterated/repeated game (one-shot process) to analyze cyber-switching attacks and mitigation in SG systems based on zero-determinant strategies. Zhu *et al.* [36] and Yan *et al.* [37] used RL for sequential attacks against power grid networks where only the attacker's action was considered. Ashok and Govindarasu [38] proposed a single-stage game that was a zero-sum game in nature for risk modeling and mitigation for the SG security. In this work, the authors also mentioned the multistage repeated game and Bayesian game as potential extensions of the work. From the aforementioned works, the Markov decision processes (MDPs) and game theories are increasingly used in the SG security to conduct the interdisciplinary research of the ML and power system. In CPPS security, very few articles have reported two active players in the repeated games. This has been a challenge in the SG to coordinate the actions of two active adversarial players and map their actions to the actual power system states. In addition, the existing games do not consider the costs of actions, budgets, and strengths as the factors for the payoff evaluation. The impact of the learned attack policies on the power system has also been neglected but is very important to the physical system.

Motivated by the above-mentioned studies, we propose a repeated game between the active adversaries and use the RL-based solution to overcome the existing limitation. The main contributions of this article are summarized in the following.

1) We design an RL algorithm for the two-person repeated game that helps to achieve the outcome in favor of the players. The main advantage of using a repeated game is that it does not require any additional information except for what is shared between the players. Unlike most of the existing studies, the repeated game conducts multiple repetitions instead of a one-shot game and uses individual utility functions to evaluate the payoffs. The policies that result in favorable outcome to the players are significantly important for proposing a protection scheme and identifying the critical elements. Favorable outcome means achieving higher payoff compared to the other player. Both players participate actively in the game and take action independently. They adopt a sequential action scheme and learn the policies based on the repeated game process. We also provide alternative action choices for the adversaries for different attack and defense strengths.
2) Realistic and crucial factors, such as costs, allocated budgets, and strengths of the adversaries, have been neglected in the existing literature. In our work, costs are parameterized as a certain percentage of the total capacity of the power system. Budgets are parameterized as a number of limited actions. Strengths of the players are parameterized as a value ranging from 0 to 1. Then, these parameters are incorporated as constraints to calculate the mixed strategy payoffs of the adversaries.
3) After analyzing the interactions and learning the policies, we illustrate the impact of the learned attacker's policies on the CPPS. Very few existing works reported

TABLE I
DIFFERENT RL ALGORITHMS USED IN THE EXISTING MULTIAGENT RL RELATED WORKS

Environment type	Game type	RL algorithm
Adversarial	Zero-sum	Minimax-Q [44], [45]
Adversarial	General-sum	Nash-Q or FFQ [46], [47]
Collaborative	General-sum	Joint action learner (JAL)[48]
Adversarial collaborative	General-sum	Joint action learner (JAL)[48]

the postattack effects (e.g., voltage, current, and power) based on the learned policies from the adversarial game. We simulate the learned attack policies in the PowerWorld simulator. For the evaluation of the impact, voltage violation is considered as the evaluation metric. The results provide insight into the damage caused by the attacks on the power system.

The rest of this article is organized as follows. Section II explains the theoretical background of the repeated game framework and detailed discussions about the threat and attack model. Section III provides detailed discussions about the algorithms and the overall picture of this work, Nash equilibrium (NE) solution of this game, design parameters, and the depiction of the mixed strategy payoffs. In Section IV, the results from the simulation case studies are explained. Section V, shows the effects of attacks in the power system. Finally, Section VI concludes this article with discussions and future works.

II. GAME FORMULATION AND THREAT AND ATTACK MODELING

A. Background of Repeated Game

A repeated game refers to a situation where the same stage game (strategic form game or one-shot game) is repeated for some duration. These type of games are also called "supergames." In other words, a repeated game is a set of multiple one-shot games. It is modeled to evaluate the logic of long-term interactions between the players. The basic idea of this game is that a player will take into account the effect of his/her current behavior on the opponent players' future behavior. Any history, except what is shared between the two players, can be disregarded. In CPPS, the adversaries (the attacker and the defender) do not need to know the full history of the opponent (limited access to the information), so the game can still reach the optimality (NE) with limited information about the opponent. Also, given the repetition of the one-shot game, the environmental information is updated from the previous repetition, so no information is missing about the attacker–defender interactions from the previous repetitions. In [39]–[43], repeated game was used in different areas of study. The repeated game can be expressed as

$$\Pi_i^X = \left\{C_m^X, C_i^X, C_a^X, P_a, U_a, U_d\right\} \in \{S_A, S_D\}. \tag{1}$$

Table I shows different RL algorithms used for solving different multiagent RL problems. From (1), we get the repeated game function Π_i^X. The game model is adopted from [43]. The payoff matrix of the repeated game can be formulated as

$$\begin{bmatrix} \sum_{x=1}^{X}\sum_{i}\left(P_a - C_a^x, U_i - C_i^x\right) & \sum_{x=1}^{X}\sum_{i}\left(-U_a, U_i - C_m^x\right) \\ \sum_{x=1}^{X}\sum_{i}\left(C_w^x, U_i\right) & \sum_{x=1}^{X}\sum_{i}\left(C_w^x, U_i - C_m^x\right) \end{bmatrix}.$$

To solve the game from this payoff matrix, the attacker's and the defender's mixed strategy payoffs can be derived as

$$\begin{aligned} U_A &= \sum_{x=1}^{X}\sum_{i}\left(P_a - C_a^x\right)(1-\delta_i)\sigma_i + (-U_a)\delta_i\sigma_i + C_w^x(1-\sigma_i) \\ &= \sum_{x=1}^{X}\sum_{i}\left(P_a - C_a^x\right)(1-\delta_i)\sigma_i - U_a\delta_i\sigma_i \end{aligned} \tag{2}$$

and

$$\begin{aligned} U_D &= \sum_{x=1}^{X}\sum_{i}\left(U_d - C_i^x\right)(1-\delta_i)\sigma_i + U_i(1-\delta_i)(1-\sigma_i) \\ &\quad + (U_i - C_m^x)\delta_i\left(1-\sigma_i\right) \\ &= \sum_{x=1}^{X}\sum_{i} C_i^x\delta_i\sigma_i + U_d - C_m^x\delta_i - C_i^x\sigma_i \end{aligned} \tag{3}$$

where δ_i and $(1-\delta_i)$ are the probabilities of defending and not defending, and σ_i and $(1-\sigma_i)$ are the probabilities of attacking and not attacking. C_w is the cost of not attacking, which is considered as zero for our research. The total loss caused by the attacks is defined as

$$P_a = \sum_{x=1}^{X}\sum_{i} C_i^X. \tag{4}$$

B. Threat and Attack Model

The threat and attack model is adopted from [18]. Line switching attack is used as the attack scheme. Transmission lines are the commonly used targets for initiating attacks in the critical infrastructures, such as an SG [49], [50]. Hackers/cyber attackers may successfully gain access to the control center of the U.S. power grid, from where they can manipulate any control command (including circuit breaker and relay) or inject false data to create massive damage. Transmission line index is given as input for this model. The output of this model is generation loss and cascaded line outages. The model executes the line switching actions and considers cascaded failures due to the overloading of the transmission lines. Based on the switched transmission lines and the cascaded failures, the generation loss and number of total line failures are calculated. The generation loss is used in the measurement of the utilities of the players. The simulation is initialized with precontingency power flow, which ensures that the system is $n-1$ secured. Then, the $n-k$ contingency is applied (k is the order of contingency) by switching k transmission line to out-of-order status. This application of contingencies may result in separating the grid into several islands. Then, the generators' ramp rates are modified to balance the demand and supply.

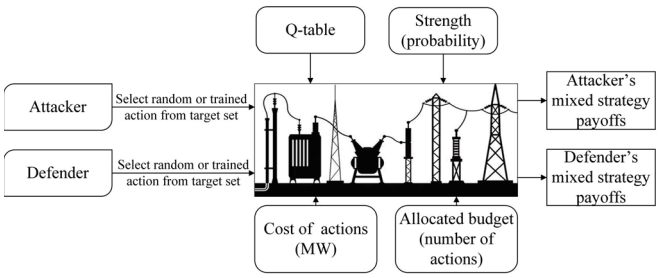


Fig. 1. Attacker–defender interaction in one repetition of a repeated game.

Next, the generation is compared to the load demand. If the generation is higher than the demand, the generators in the islands are tripped until the generation and the demand are balanced. After this, if there is less generation than the demand, the load is shed to balance the generation and demand. Then, the standard dc power flow is applied and checked for the overloads. If there are no overloads, the simulation is terminated. For each transmission line, a time-delayed overcurrent relay determines if an overcurrent tripped a transmission line. In this research, we study the simulation-based impact of attacks and disconnect transmission lines from the system. Under these circumstances, the outcome of this research will provide the following strategic benefits to the power system operators: 1) establishing the necessity of enforcing additional security measures such as remedial action scheme (islanding, microgrid, and so on) for the vulnerable transmission lines and 2) protecting the components that can cause higher damage than the other components.

III. Proposed Algorithm for the Adversarial Repeated Game

The interaction between the adversaries in the power system in one repetition is shown in Fig. 1. The adversaries take their actions from their associated target sets (S_A and S_D). Their strengths (probabilities of attacking and defending), cost of actions, and allocated budgets' information are provided to calculate their mixed strategy payoffs.

Fig. 1 visualizes the agent–environment interactions in the RL framework. The attacker and the defender represent the agents, and the power system represents the environment. The mixed strategy payoffs are the evaluation feedback from the environment. Based on this feedback, the reward is assigned to the agents. The overall diagram of the gaming process using learning theory is presented in Fig. 2 showing the process of solving a bilevel problem. In the upper level, the learning is helping to find a favorable outcome. In the lower level, gaming is conducted between the adversaries. The main idea is to use RL to learn the behavior of the adversaries in order to achieve favorable outcome of the repeated game.

A. Proposed RL-Based Solution

We formulate the repeated game as a zero-sum game and use Minimax-Q algorithm (an RL algorithm) to solve it. RL is used to learn the optimal action strategies for favorable outcomes. The quality of the state for this game is given by

$$Q(s,a,d) = R + \gamma \sum_{s'} T(s,a,d,s')V_a(s') \tag{5}$$

where γ ranges from 0 to 1. The value of γ close to 0 focuses on short-term reward, and the value of γ close to 1 gives more emphases on the long-term reward. $T(s,a,d,s')$ is considered equal for all state transitions. The value of the state for the attacker can be calculated by

$$V_a(s') = \max_{\pi\in\Pi} \min_{d} \sum_{a} Q(s',a,d)\pi(a). \tag{6}$$

Similarly, the value of the state for the defender can be given by

$$V_d(s') = \min_{\pi\in\Pi} \max_{a} \sum_{d} Q(s',a,d)\pi(d). \tag{7}$$

Since this is a zero-sum game with weak duality, $V_a^*(s') = V_d^*(s')$. In this game, the aim is to maximize the difference between the players' utilities, depending on the game's desired outcome. The problem in (6) can be solved using the similar value iteration algorithm as that in [51], [52]

$$\begin{aligned} &\max_{\pi} \ V_a(s') \\ &\text{s.t.} \sum_{a\in S_A} Q(s,a,d)\pi \geq V_a(s') \\ &\sum_{a\in S_A} \pi(a) = 1 \\ &\pi(a) \geq 0 \quad \forall a \in S_A. \end{aligned} \tag{8}$$

The optimal policy is found by

$$\pi'(s) = \arg\max_{a} \ Q_\pi(s,a,d). \tag{9}$$

The probabilities of the state-action pairs are updated following the formula below:

$$\Pr(s,a,d) = \frac{C(s,a,d)}{\sum_{a\in A, d\in D} C(s,a,d)}. \tag{10}$$

The attacker's mixed strategy for a given state s will be

$$\pi_A(s) = [\Pr\{a(s) = a_1\}, \ldots, \Pr\{a(s) = a_N\}] \tag{11}$$

where

$$\sum_{i=1}^{N} \Pr\{a(s) = a_i\} = 1 \tag{12}$$

where $\Pr\{a(s) = a_i\}$ is the probability of choosing attack action a_i in state $s \in S_A$, and $\pi_A(s)$ represents the probability distribution over the attacker's action space associated with the state s. Similarly, we can define the probability for the defender as well.

Algorithm 1 gives the steps for a repeated game in one run. The repetition continues until the resource expenditure by the players exceeding the allocated budget or until there are no targets available in the target set. In every repetition, the adversaries interact and their associated utilities are calculated. Their actions are recorded, and their probabilities are updated in every repetition.

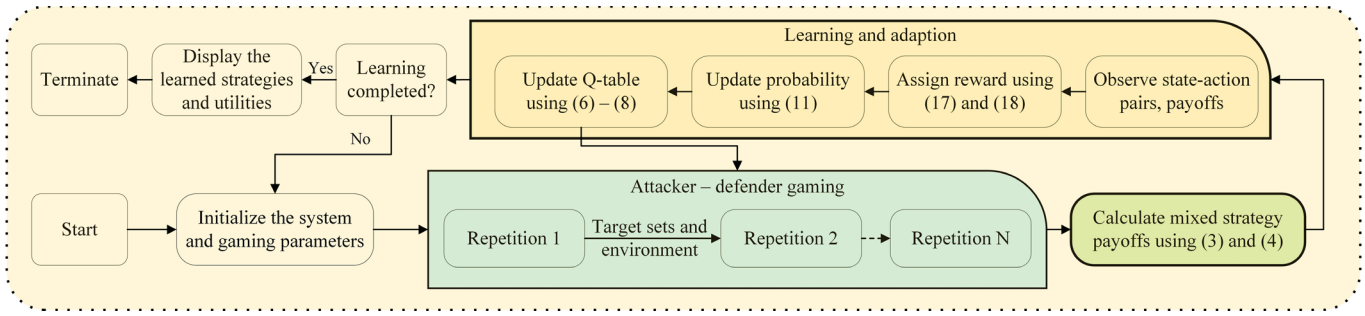


Fig. 2. Overall block diagram of the RL-based solution of a repeated game in the CPPS security. The game showed in the lower block is repeated for transiting from one steady state to the next steady state. At the end of repetition N, the game transits to the next run $n = n + 1$. The detailed interactions between the adversaries in one repetition are explained in Fig. 1. The attacker's target set, the defender's target sets, and the power system environment are updated from one repetition to another repetition. At the end of the repetitions, the history of the actions, payoffs, rewards, and so on from previous runs are updated for the learning process.

Algorithm 1 Adversarial Repeated Game in CPPS Security

Input : Test case, attacking probability, defending probability, budgets of the players, costs of attack and defense, attack and defense sets
Initialize the Q-table, action counter, policies for each state-action pairs with uniform probability distribution;
for *The given system, and maximum number of run* **do**
  Initialize the players' resource spent;
  for *Maximum number of repetition and till the players' resource spent* $\leq$ *Budget* **do**
    Take actions from the agents' associated sets based on **Minimax-Q algorithm** from Algorithm 2;
    Calculate the costs of actions using (18) ;
    Update the resources spent from budget;
    Calculate the utilities using (2) and (3);
    Update the action sets from (14);
    if *There are available resources for the attacker and the defender* **then**
      Go to next repetition;
    else
      Exit the repetition loop and go to next run;
    end
  end
end
Output: The optimal actions with their associated probabilities.

Algorithm 2 Minimax-Q for Repeated Game

Input : Attack and defense action sets
for *Maximum number of run* **do**
  Initialize the attacker's and the defender's state;
  for *Maximum number of repetitions* **do**
    if *Prob* $> \epsilon$ **then**
      Take a random action from the available actions in the sets for both players;
    else
      Take an attack action using (6);
      Take a defense action using (7);
    end
    Execute the action using (15);
    Update action counter, $C = C + 1$ and associated probabilities using (10);
    Calculate the utilities;
    Assign the reward based on (16);
    Update the Q-value using (5);
  end
end
Output: Output the optimal action policies for the attacker and the defender with their associated probabilities.

Algorithm 2 represents the Q-learning algorithm that is used to solve the repeated game. This Q-learning algorithm is learning the behavior of the players from their actions on the top layer of the gaming framework. In the bottom layer, the repeated game is a multistage game. Algorithm 2 is initialized with the players' target sets as the input. The expected outputs of this algorithm are the optimal action sequences for the players along with their probabilities. Then, according to the exploration probability, actions are executed by the players either randomly or following the policy.

With the execution of the actions, a transmission line is switched to out of service and another transmission line is defended from being attacked. After execution of the actions, the utilities for the players are calculated, and their associated Q values are updated. The probabilities for the associated state-action pair are also updated. This process is continued for the maximum number of runs.

B. Design Parameters

Design parameters of this repeated game solution based on RL between the adversaries in the electric power grid are explained briefly in this section.

1) Attack and Defense Sets as the Input of the Threat and Attack Model: The attack and defense sets are the collection of targets for the attacker and the defender. Since we use line switching attack (explained in Section II-B), transmission lines are considered as the target elements for attacking and

defending. Every repetition of the game is considered as the states of the game and represented by S and

$$S = \{\text{Rep}_1, \text{Rep}_2, \ldots, \text{Rep}_N\} \tag{13}$$

where Rep_N is the repetition and N is the number of total repetitions. The attacker and the defender action spaces are represented by S_A and S_D

$$\begin{aligned} S_A &= \{a_1, a_2, \ldots, a_N\} \\ S_D &= \{d_1, d_2, \ldots, d_N\} \end{aligned} \tag{14}$$

where a_N and d_N are the attack and the defense actions in the Nth repetition, respectively. These actions are used as the input to the threat and attack model to quantify loss or damage described in Section II-B.

2) Reward: After each attack and defense action, rewards are assigned by evaluating the feedback of the actions from the environment. $R_A(s, a, d)$ represents the attacker's expected reward associated with the state $s \in S$ and attack and defense actions $a \in S_A$ and $d \in S_D$, respectively. Similarly, $R_D(s, a, d)$ represents the defender's expected reward. The reward is designed based on the requirement. The reward is assigned based on the mixed strategy payoffs of the players. The calculation of the mixed strategy payoffs requires generation loss, which is one of the outcomes of the threat and attack model. The states of this game are presented as a combination of the elements of the target sets. Whenever, an attack or defense action is triggered, the index of that transmission line in the target set switches to zero. The line status in a target set can be represented by $s_t(l)$

$$s_t(l) = \begin{cases} Z, & \text{if line } l \text{ is in-service at time } t \\ 0, & \text{if line } l \text{ is out-of-service at time } t \end{cases} \tag{15}$$

where Z represents the transmission line number. In this zero-sum game, the reward is assigned opposite to each other. Thus, if the attacker's and the defender's mixed strategy payoffs are U_A and U_D, then the reward is assigned following the conditions given below. When the desired outcome of the game is in favor of the attacker

$$R_A(s, a, d) = \begin{cases} +1, & \text{if } U_A > U_D \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

and

$$R_D(s, a, d) = \begin{cases} -1, & \text{if } U_A > U_D \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

When the game is conducted expecting the outcome in favor of the defender, the reward assignment is just in the opposite way of (16) and (17).

3) Allocated Budget and the Costs: The allocated budgets for the attacker and the defender in this game are defined by B_A and B_D. The budget is the amount of resource that limits the action space for the players. To the best of our knowledge, there is no generic way to define the cost of actions and total budget. The total budget also depends on the attackers' available resources and how much access the attacker gained to the system control. In addition, if the system security is weaker, the attacker will have a higher budget and lower cost of actions. In some cases, the cost can be monetized; in some cases, the cost can be considered in terms of the generation power absorbed from the system. In this game, we considered a limited number of actions according to the allocated budget for the players. For demonstration purposes, we consider that the attacker and the defender cannot take more than three actions. Cost is the amount of resource that the players consume while executing the actions. For simplicity, we consider the cost of the attack and defense action as a certain percentage of the total loading capacity of the system. In this game, the costs of attack and defense actions are defined by C_a and C_m, respectively

$$\begin{aligned} C_a &= 0.1 \times \text{TLC} \\ C_m &= 0.1 \times \text{TLC} \end{aligned} \tag{18}$$

where TLC represents the total loading capacity of the system. The attacking and defending probabilities σ and δ represent the strengths of the attacker and the defender and are used to calculate the mixed strategy payoffs of the players.

C. Depicting the Attacker's and the Defender's Mixed Strategy Payoffs and NE

NE is one of the central concepts in game theory and economics. With the NE point, each agent's policy is the best policy against other agent's policy. In NE, we compare the attacker's mixed strategy payoffs with the defender's mixed strategy payoffs. These payoffs represent how much power is saved due to the defense actions and how much power is lost due to the attack actions. Due to the rewarding condition $R_D(a, d, s) = -R_A(a, d, s)$, the proposed repeated game is a zero-sum game [53]. To solve a two-player stochastic game in the normal form, one popular solution is the closed-loop NE. Nash's theorem guarantees that the NE exists for static games. However, for stochastic games, the possible number of strategies is infinite [31], [51], [54]. Reference [55] shows that the optimal defense strategy can be achieved by adopting a game-theoretic actor–critic neural network for optimal defense and worst attack policy. There might be multiple optimal strategies for the repeated games in the case of the finite-time horizon. While the attacker converges to its optimal action policies maximizing its payoffs, the defender converges to its optimal policies minimizing its payoffs.

For an RL-based solution of a repeated game, the existence of the NE can be confirmed through the decision making process over time. Players in the repeated game play repeatedly, following the gaming rules, until the end. In this type of game, both players reach the NE point (converges to optimal policy). As mentioned in [18], for a multistage game, Q-learning is one of the ideal ways to find the player's optimal policy. In this repeated game, the optimal strategies (outcomes of the game) are in favor of the players. Based on the success of the game objective, the rewarding policy maximizes the discounted sum of the future reward. Thus, based on the desired outcome, the attacker's action or the defender's action helps the game to converge to NE policies. Hence, the convergence analysis of this repeated game will ensure the findings of the NE point. The convergence of Q-learning (Minimax) algorithm

follows by relating the convergence proof described in [53], [56]–[58].

Derived from the general Bellman equation for Q-learning, we get the following Q equation by substituting $\alpha = 1$, the learning parameter:

$$Q(s, a, d) = [R(s, a, d) + \gamma V^*(s')]. \tag{19}$$

We use x as the time step where, $x = 1, 2, 3, \ldots, X$ and X is the maximum time step (repetition). The rewritten Bellman equation converges to the optimal $Q^*(s, a, d)$ values if the system has a finite state and action space and the following.

1) $\sum_{x=0}^{X} \alpha(s, a, d) < \infty$ and $\sum_{x=0}^{X} \alpha^2(s, a, d) < \infty$ uniformly with probability 1.
2) Var$\{R(s, a, d)\}$ is bounded.
3) If $\gamma = 1$, all policies lead to a cost-free terminal state with probability 1.

By subtracting $Q^*(s, a, d)$ from both sides of (19) and define $G(s, a, d) = Q(s, a, d) - Q^*(s, a, d)$ together with

$$G(s, a, d) = R(s, a, d) + \gamma V^*(s') - Q^*(s, a, d) \tag{20}$$

where $V^*(s')$ is the value of the next state s' and the agent's Q value is optimized based on the game's desired outcome. $G(s, a, d)$ is a contraction mapping with respect to some max–min norm that perfectly fits for the Minimax Q-learning algorithm. This is done by

$$\begin{aligned}
&\max_{a\in A}\min_{d\in D} |G(s, a, d)| \\
&\quad = \gamma \max_{a\in A}\min_{d\in D} |\sum_{s'\in S}[V(s') - V^*(s')]| \\
&\quad \leq \gamma \max_{a\in A}\min_{d\in D} \sum_{s'\in S} \max |Q(s', a', d') - Q^*(s', a', d')| \\
&\quad = \gamma \max_{a\in A}\min_{d\in D} \sum_{s'\in S} V^{\Delta} \\
&\quad = H(V^{\Delta})
\end{aligned} \tag{21}$$

where H is the value iteration operator (the cost associated with each state is zero). If $\gamma < 1$, the contraction property of H and, thus, G can be seen directly from the above-mentioned formulas. When the future costs are not discounted ($\gamma = 1$) but the chain is absorbing and all policies lead to the terminal state with probability 1, there still exists a weighted max–min norm. H is a contraction mapping [59] with respect to this norm. $G(s, a, d)$ depends on $Q(s, a, d)$ at most linearly. Thus, the variance of $G(s, a, d)$ is within the bounds of theorem [53], [56]–[58], and the variance of $R(s, a, d)$ is bounded. The contraction mapping of $G(s, a, d)$ is maximizing $|Q(s', a', d') - Q^*(s', a', d')|$ due to the attack action a. Thus, the converged optimal $Q^*(s, a, d)$ is the solution for game, while the objective is to find the optimal action strategies in favor of the attacker. Similarly, we can prove the convergence to the optimal strategies, while the game is conducted in favor of the defender. The contraction mapping for the game where the outcome is desired to be in favor of the defender can be represented as

$$\min_{a\in A}\max_{d\in D} |G(s, a, d)| = H(V^{\Delta}). \tag{22}$$

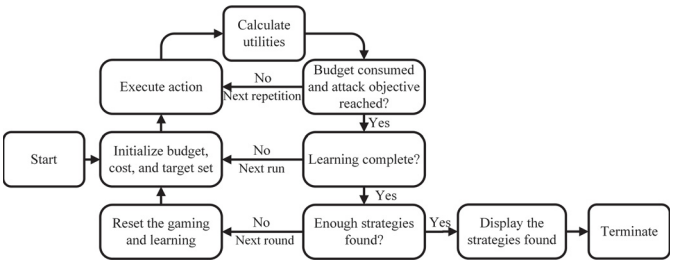


Fig. 3. Simulation loops are explained in this figure. The simulation contains repetitions, runs, and round loops.

Here, the contraction mapping of $G(s, a, d)$ will maximize $|Q(s', a', d') - Q^*(s', a', d')|$ due to the defense action d. Hence, the convergence of the players to their optimal action policies (or NE point) for the repeated game is analyzed.

IV. Simulation Results' Analysis

The simulation is conducted using MATLAB R2018a on a standard PC with an Intel i7-6700 CPU running at 3.40 GHz and 24-GB RAM. There are three loops in the simulation: rounds, runs, and repetitions. Fig. 3 shows the implementation loops for the learning and gaming for this adversary game. The repetition loop ends when the allocated budget is fully consumed by the players. The run loop is the main loop where the players (the attackers and the defenders) and the environment (power system) interact to converge to the optimal policies. As the number of runs increases, the players tend to learn the optimal policies. At the end of the run loop, the players reach the NE point. The run loop is repeated in several rounds in the round loop to get different optimal policies. The number of repetitions also represents the number of actions that the attacker and the defender take. The number of runs represents the number of trials required in the learning process. A Q-table is required in the learning process that uses the Q-learning algorithm. Q-table is defined with the attacker's and the defender's state-action pairs, action counters, probabilities, and Q values. The discount factor γ is applicable for both of the players' learning procedures. Thus, both players decide the value of γ. The attacker and the defender are allowed to take only one action in a repetition.

The outcome of this repeated game, solved by RL, is analyzed mainly from two different perspectives. First, we analyze the outcome from the game-theoretic viewpoint where the attacker's and the defender's utilities are compared. Next, we analyze the postattack effects in the simulated power system. In this section, we analyze the outcome of the game from the game-theoretic viewpoint. We conduct different case studies under different conditions to find out the favorable outcome for the attacker and the defender.

Fig. 4 shows the case studies and the conditions associated with the repeated game solution using RL. The aim of these case studies is to analyze the behavior of the agents and learn the strategies for different settings of the game (such as in attacker's favor and defender's favor). These case studies provide optimal strategies of the players (including alternative action choices) in different game settings. Branches in the

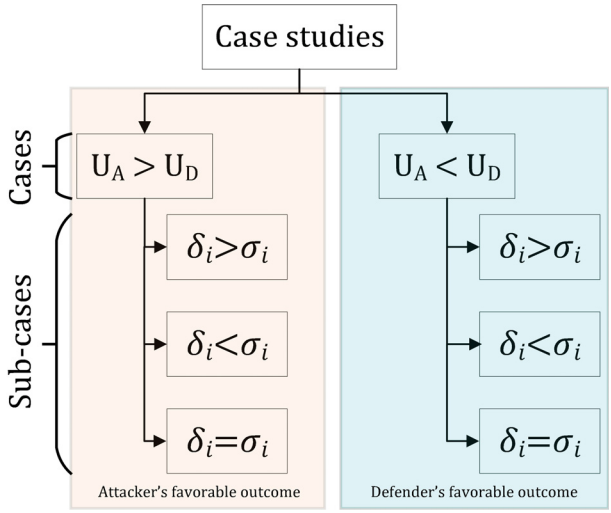


Fig. 4. Case studies for the solution of the repeated game using RL on the IEEE 39 bus system. Basically, we solve for the attackers' and the defenders' favorable outcome for their different strengths.

TABLE II
VALUE OF THE PARAMETERS USED IN THIS RESEARCH

Parameter	Value	Parameter	Value
Total branches	46	ϵ	0.8
B_A	5 actions	γ	0.9
B_D	5 actions	F	5000
$N(S_A)$	10	Avg	10
$N(S_D)$	10	TLC	6150 MW
C_a	$0.01 \times TLC$	C_m	$0.01 \times TLC$

left-hand side under the case studies are the conditions, where the outcome of the game is in favor of the attacker ($U_A > U_D$). Branches in the right-hand side are the conditions, where the outcome of the game is in favor of the defender ($U_A < U_D$). δ_i and σ_i represent the probabilities (strength) of defending and attacking, respectively. For different attacking and defending strengths, the game simulation is conducted.

U_A and U_D represents the attacker's and the defender's utilities, respectively. During the gaming, the adversarial game will terminate when one of the players wins. In order to reach this decision, either the attacker has to win ($U_A > U_D$) or the defender has to win ($U_A < U_D$). If $U_A = U_D$, the players will go to the next gaming stage by taking more action. If all the resources are used, then the game will be a draw and the players will have less valuable information to learn from that game. Table II shows some general settings regarding the game, which are commonly used in different case studies. There are ten independent rounds conducted for the game simulation, and each round consists 5000 runs. The initial exploration probability is set to 0.8. It gradually drops to a very small and positive final value ϵ_f until the end. A very small and positive value of ϵ_f ensures the convergence to the optimal policy and still avoids the local optimal point.

Table III is showing the attacker's and the defender's action sets. These action sets are containing ten different transmission lines that are selected randomly. Targets are selected from these target sets according to the budget.

TABLE III
ATTACKER'S AND DEFENDER'S TARGET SETS CONTAINING THE TRANSMISSION LINES FROM THE IEEE 39 BRANCH SYSTEM. EACH OF THE TARGET SETS HAS TEN TRANSMISSION LINES

Parameter	Line index
Attacker's set, S_A	[18 20 5 13 6 33 3 26 43 37]
Defender's set, S_D	[19 21 40 46 36 1 32 10 45 29]

TABLE IV
ATTACK AND DEFENSE ACTIONS WITH THEIR ASSOCIATED PROBABILITIES AND Q VALUE

Time step	Attack line index	Defense line index	Probability	Q-value
1	43	36	0.92	0.66
2	26	21	0.94	0.73
3	20	40	0.95	0.81
4	5	10	0.97	0.90
5	13	1	0.98	1

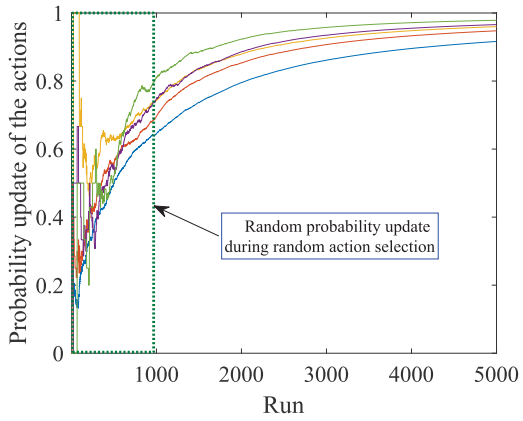


Fig. 5. Probability update for the optimal actions throughout the 5000 runs.

Next, we analyze the outcomes of this game in different conditions for the attacker's and the defender's favorable outcomes.

1) Case I ($U_A > U_D$): In this section, we conduct the game to find the outcome in favor of the attacker. Three different conditions are applied to evaluate the attacker's and defender's strengths.

$\delta_i < \sigma_i$: In this condition, the game is solved in favor of the attacker, and the attacker has higher probability of attacking (higher strength) than the defender. Table IV is showing the optimal actions for the attacker and the defender with their associated probabilities, and Q values. The initial actions (at time step 1) for the attacker and the defender are the attacking transmission line 43 and the defending transmission line 36. These actions have the highest probability (0.92) of being selected among the other available actions. The updating of the probabilities of the optimal actions throughout the 5000 runs is shown in Fig. 5. The oscillations inside the green dotted box show the probability update during the random actions' selection.

TABLE V

PROBABILITIES OF ALL THE POSSIBLE ACTIONS FOR SELECTING THE SECOND ACTIONS

Attack line index	Defense line index	Probability
26	21	0.941
20	1	0.001
26	10	0.001
6	46	0.001
3	40	0.001
20	45	0.001
...	...	...
3	10	0.000
20	46	0.002

TABLE VI

OPTIMAL ACTION SEQUENCES FOR THE PLAYERS WITH THEIR PROBABILITIES AND Q VALUES. THESE OUTCOMES ARE IN FAVOR OF THE DEFENDER, AND THE ATTACKER HAS HIGHER STRENGTH (PROBABILITY OF 0.8) THAN THE DEFENDER (PROBABILITY OF 0.2)

Time step	Attack line index	Defense line index	Probability	Q-value
1	5	40	0.92	0.66
2	43	46	0.95	0.73
3	37	10	0.94	0.81
4	33	29	0.97	0.90
5	20	21	0.98	1

Table V shows the probabilities of all the possible actions for the second action selection. These probabilities help the players to select their actions. Actions 26 and 21 have the highest probability of 0.94. In any case, if the actions with the highest probability are inaccessible or not available, the actions with the second-highest probability will be selected by the players. In this case, actions 20 and 46 will be selected as the action by the attacker and the defender, respectively.

Similarly, the optimal attack action sequences can be found for $\delta_i > \sigma_i$ and $\delta_i = \sigma_i$ along with their associated probabilities.

2) Case II ($U_A < U_D$): Similar to Case I, we conduct the game with different attacker's and defender's strengths. However, in this case, we define the conditions in a way that all the outcomes or policies go in favor of the defender. After conducting the game for the aforementioned scenarios, we found different attack and defense action policies with their associated probabilities.

$\delta_i > \sigma_i$: In this section, we find the optimal action sequences in favor of the defender. Here, the defender has higher strength than the attacker. Table VI shows the optimal actions in favor of the defender where the defender has the higher strength (probability of 0.8) than the attacker (probability of 0.2). The first column is providing the time steps of the attack and the defense actions. The second and third columns are providing the actions for the attacker and the defender (associated with that time step). The fourth column is providing the probabilities of these actions to be selected among all the possible actions. The fifth column is showing the Q values associated with those actions.

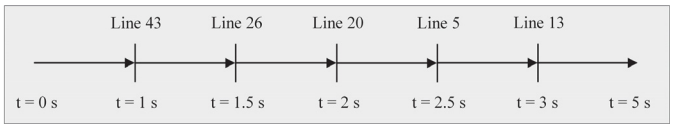


Fig. 6. Simulation is conducted for 5 s. Starting from $t = 0$, the first attack on line 43 is triggered at 1 s. Next, lines 26, 20, 5, and 13 are triggered at 1.5, 2, 2.5, and 3 s.

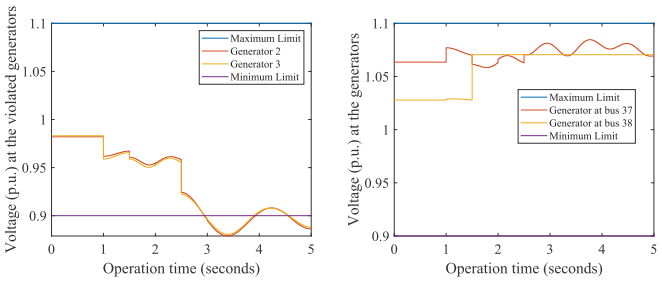


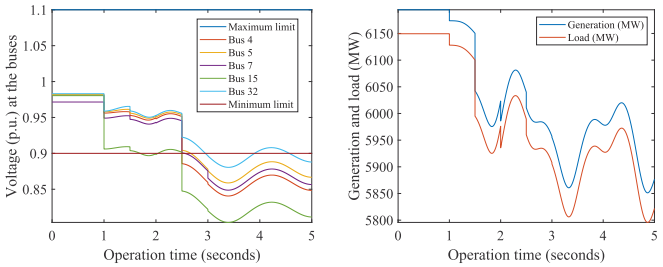
Fig. 7. (a) Per unit voltages (magnitudes) at the violated generators (generators 2 and 3) during the attack. (b) Per unit voltages (magnitudes) at the vulnerable generators during the attack.

Similarly, we can find the optimal action sequences in favor of the defender with $\delta_i < \sigma_i$ and $\delta_i = \sigma_i$ along with their associated probabilities. These optimal action sequences will be the defender's best action choices when the defending strength is less than the attacking strengths and both the attacking and defending strengths are equal, respectively.

V. ANALYSIS OF IMPACT ON THE POWER SYSTEM

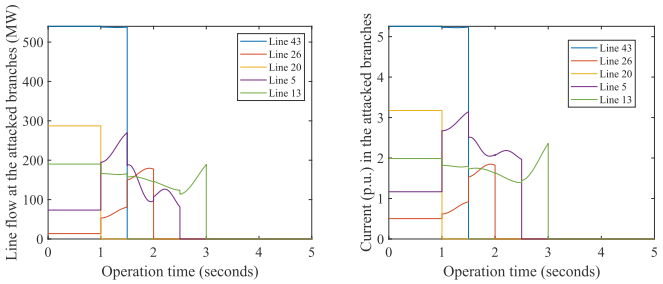
The impact analysis on the power system validates the damage caused by the learned attack actions. Upon validation, the power system operators can take additional measures (such as forced islanding, utilizing microgrid, and distributed energy resources) to protect the vulnerable components and continue providing quality electrical energy. In addition, the impact analysis also reveals some critical components (such as bus, generators, and transmission lines) that are vulnerable to failure due to the attacks. Power system operators should pay more attention to these vulnerable components to reduce the damages. In this section, we further analyze the impacts of attacks through a power system simulator. The impacts are analyzed for the first case and the first condition, where $U_A > U_D$ and $\delta_i < \sigma_i$. The attack in the power system can be illustrated in a timescale in Fig. 6.

For simulation purpose, we assume that the attacks are conducted with a time gap of 0.5 s. The simulation starts with the normal operating condition, where no transmission line is attacked at $t = 0$ s. After initiating all the transmission line switching attacks, the simulation ends at $t = 5$ s. To assess the disturbances in the system caused by the attacks, we consider the voltage violation of the system elements. There are some defined limits for the voltage violation used in the existing research works [60]–[63]. In our case, we consider the limit of voltage violation is from 0.9 to 1.1 V·p.u. The per unit voltages at the violated generators 2 and 3 are shown in Fig. 7(a).



(a) (b)

Fig. 8. (a) Some violated bus voltages (p.u.) during the attack. (b) Generation and load losses for the whole system during the attack.



(a) (b)

Fig. 9. (a) Power flow (MW) at the target transmission lines. (b) Current flow at the target transmission lines.

From Fig. 7(a), we can see that these generator voltages drop below the lower limit (0.9). Thus, these generators violate the voltage (p.u.) limits. Fig. 7(b) shows the generator voltages (p.u.) of the generator at bus 37 and bus 38. These bus voltages are close to the upper limit of the generator voltages (1.1 p.u.). These generators are vulnerable to violation because any minor changes in the load or other minor disturbances can shoot these generator voltages above the higher limit.

Fig. 8(a) shows the voltages (p.u.) of buses 4, 5, 7, 15, and 32. We can see that these voltages drop below the minimum limit after the attack at 2.5 s. Fig. 8(b) shows the generation and load losses during the attack.

Fig. 9(a) shows the power flow (MW) through the target transmission lines. The line flows drop to zero when they are attacked. Fig. 9(b) shows the current flow (A) through the target transmission lines. The current flow drops to zero when they are attacked.

VI. Conclusion

This article proposes a novel and effective solution for the two-person zero-sum repeated game between the adversaries in the CPPS security based on the RL algorithm. The learned attacker's and defender's optimal strategies give significant information about the critical transmission lines of a CPPS. Case I finds the optimal action sequences for both the players in favor of the attacker for different players' strengths. Case II provides the optimal action sequences for both the players in favor of the defender for different strengths. These case studies also provide alternative action choices with their probabilities for other possible actions. In addition, we illustrate the impact of the attack on the simulated physical system and identify the generators and buses that suffer from voltage violation from the attack. From these case studies, we suggest the optimal action choices to resolve the game in the attacker's favor or in the defender's favor regardless of their strengths. These action sets will help the authorities to defend the transmission lines more effectively and efficiently. Moreover, the defense actions are executed against the individual attack actions, which reduces the resource consumption of the defensive action schemes. In this article, we also identify the vulnerable elements of a CPPS in an adversarial environment by applying repeated game theory and Q-learning. The game formulation requires multiple agents, an interacting environment, and action evaluation and termination criteria. The RL algorithm requires information regarding the agent's resources, actions, and states of the environment. Given necessary information of the agents and the environment for gaming and learning, it is possible to extend the proposed approach for the general cyber–physical systems.

References

[1] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Syst. J.*, vol. 11, no. 3, pp. 1644–1652, Sep. 2017.

[2] D. S. Terzi, B. Arslan, and S. Sagiroglu, "Smart grid security evaluation with a big data use case," in *Proc. IEEE 12th Int. Conf. Compat., Power Electron. Power Eng. (CPE-POWERENG)*, Apr. 2018, pp. 1–6.

[3] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[4] S. Poudel, Z. Ni, and N. Malla, "Real-time cyber physical system testbed for power system security and control," *Int. J. Elect. Power Energy Syst.*, vol. 90, pp. 124–133, Sep. 2017.

[5] M. L. Tuballa and M. L. Abundo, "A review of the development of smart grid technologies," *Renew. Sustain. Energy Rev.*, vol. 59, pp. 710–725, Jun. 2016.

[6] Y. Yoldaş, A. Önen, S. M. Muyeen, A. V. Vasilakos, and İ. Alan, "Enhancing smart grid with microgrids: Challenges and opportunities," *Renew. Sustain. Energy Rev.*, vol. 72, pp. 205–214, May 2017.

[7] S. Paul, A. Parajuli, M. R. Barzegaran, and A. Rahman, "Cyber physical renewable energy microgrid: A novel approach to make the power system reliable, resilient and secure," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT-Asia)*, Nov. 2016, pp. 659–664.

[8] S. Poudel, Z. Ni, and W. Sun, "Electrical distance approach for searching vulnerable branches during contingencies," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3373–3382, Jul. 2018.

[9] S. Paul, M. R. Haq, A. Das, and Z. Ni, "A comparative study of smart grid security based on unsupervised learning and load ranking," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2019, pp. 310–315.

[10] X. Lu, X. Wang, D. Rimorov, H. Sheng, and G. Joós, "Synchrophasor-based state estimation for voltage stability monitoring in power systems," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2018, pp. 1–6.

[11] V. B. Krishna, C. A. Gunter, and W. H. Sanders, "Evaluating detectors on optimal attack vectors that enable electricity theft and DER fraud," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 4, pp. 790–805, Aug. 2018.

[12] S. Paul and Z. Ni, "Study of learning of power grid defense strategy in adversarial stage game," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2019, pp. 292–297.

[13] Z. Ding, Y. Xiang, and L. Wang, "Incorporating unidentifiable cyberattacks into power system reliability assessment," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2018, pp. 1–5.

[14] C.-W. Ten, K. Yamashita, Z. Yang, A. V. Vasilakos, and A. Ginter, "Impact assessment of hypothesized cyberattacks on interconnected bulk power systems," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4405–4425, Sep. 2018.

[15] S. Paul and Z. Ni, "A strategic analysis of attacker-defender repeated game in smart grid security," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2019, pp. 1–5.

[16] L.-Lu, H. J. Liu, and H. Zhu, "Distributed secondary control for isolated microgrids under malicious attacks," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2016, pp. 1–6.

[17] S. Paul and Z. Ni, "A study of linear programming and reinforcement learning for one-shot game in smart grid security," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[18] Z. Ni and S. Paul, "A multistage game in smart grid security: A reinforcement learning solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2684–2695, Sep. 2019.

[19] S. Paul and Z. Ni, "Vulnerability analysis for simultaneous attack in smart grid security," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Apr. 2017, pp. 1–5.

[20] Z. Ni, S. Paul, X. Zhong, and Q. Wei, "A reinforcement learning approach for sequential decision-making process of attacks in smart grid," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–8.

[21] K. Mahapatra and N. R. Chaudhuri, "Malicious corruption-resilient wide-area oscillation monitoring using online robust PCA," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2018, pp. 1–5.

[22] Y. Zhu, D. Zhao, and X. Li, "Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 714–725, Mar. 2017.

[23] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

[24] Q. Wei, D. Liu, Q. Lin, and R. Song, "Adaptive dynamic programming for discrete-time zero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 957–969, Apr. 2018.

[25] J. Li, H. Modares, T. Chai, F. L. Lewis, and L. Xie, "Off-policy reinforcement learning for synchronization in multiagent graphical games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2434–2445, Oct. 2017.

[26] M. Johnson, R. Kamalapurkar, S. Bhasin, and W. E. Dixon, "Approximate N-player nonzero-sum game solution for an uncertain continuous nonlinear system," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1645–1658, Aug. 2015.

[27] Y. Fu and T. Chai, "Online solution of two-player zero-sum games for continuous-time nonlinear systems with completely unknown dynamics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2577–2587, Dec. 2016.

[28] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[29] X. Zhong, H. He, D. Wang, and Z. Ni, "Model-free adaptive control for unknown nonlinear zero-sum differential game," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1633–1646, May 2018.

[30] F. Pourahmadi, M. Fotuhi-Firuzabad, and P. Dehghanian, "Application of game theory in reliability-centered maintenance of electric power systems," *IEEE Trans. Ind. Appl.*, vol. 53, no. 2, pp. 936–946, Mar./Apr. 2017.

[31] L. Wei, A. I. Sarwat, W. Saad, and S. Biswas, "Stochastic games for power grid protection against coordinated cyber-physical attacks," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 684–694, Mar. 2018.

[32] W. Liao, S. Salinas, M. Li, P. Li, and K. A. Loparo, "Cascading failure attacks in the power system: A stochastic game perspective," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2247–2259, Dec. 2017.

[33] Q. Wang, W. Tai, Y. Tang, M. Ni, and S. You, "A two-layer game theoretical attack-defense model for a false data injection attack against power systems," *Int. J. Electr. Power Energy Syst.*, vol. 104, pp. 169–177, Jan. 2019.

[34] K. Wang, M. Du, S. Maharjan, and Y. Sun, "Strategic honeypot game model for distributed denial of service attacks in the smart grid," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2474–2482, Feb. 2017.

[35] A. Farraj, E. Hammad, A. Al Daoud, and D. Kundur, "A game-theoretic analysis of cyber switching attacks and mitigation in smart grid systems," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1846–1855, Jul. 2016.

[36] Y. Zhu, J. Yan, Y. Tang, Y. Sun, and H. He, "The sequential attack against power grid networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 616–621.

[37] 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 Security*, vol. 12, no. 1, pp. 200–210, Jan. 2017.

[38] A. Ashok and M. Govindarasu, "Cyber-physical risk modeling and mitigation for the smart grid using a game-theoretic approach," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2015, pp. 1–5.

[39] T. AlSkaif, M. G. Zapata, B. Bellalta, and A. Nilsson, "A distributed power sharing framework among households in microgrids: A repeated game approach," *Computing*, vol. 99, no. 1, pp. 23–37, Jan. 2017.

[40] L. Song, Y. Xiao, and M. van der Schaar, "Demand side management in smart grids using a repeated game framework," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 7, pp. 1412–1424, Jul. 2014.

[41] D. B. Smith, M. Portmann, W. L. Tan, and W. Tushar, "Multi-source–destination distributed wireless networks: Pareto-efficient dynamic power control game with rapid convergence," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2744–2754, Jul. 2014.

[42] V. S. Varma, M. Mhiri, M. L. Treust, S. Lasaulce, and A. Samet, "On the benefits of repeated game models for green cross-layer power control in small cells," in *Proc. 1st Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, Jul. 2013, pp. 137–141.

[43] K. Wang, M. Du, D. Yang, C. Zhu, J. Shen, and Y. Zhang, "Game-theory-based active defense for intrusion detection in cyber-physical embedded systems," *ACM Trans. Embedded Comput. Syst.*, vol. 16, no. 1, p. 18, 2016.

[44] E. Yang and D. Gu, "Multiagent reinforcement learning for multi-robot systems: A survey," Univ. Essex, Colchester, U.K., Tech. Rep. N/A, 2004.

[45] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. 7th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA: Morgan Kaufmann, Jul. 1994, pp. 157–163.

[46] M. L. Littman, "Friend-or-foe q-learning in general-sum games," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA: Morgan Kaufmann, Jun. 2001, pp. 322–328.

[47] M. Bowling and M. Veloso, "An analysis of stochastic game theory for multiagent reinforcement learning," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-00-165, 2001.

[48] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, *Game Theory and Multi-agent Reinforcement Learning*. Berlin, Germany: Springer, 2012, pp. 441–470.

[49] J. Yan, Y. Tang, Y. Zhu, H. He, and Y. Sun, "Smart grid vulnerability under cascade-based sequential line-switching attacks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec.2015, pp. 1–7.

[50] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "A framework for cyber-topology attacks: Line-switching and new attack scenarios," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1704–1712, Mar. 2019.

[51] C. Y. T. Ma, D. K. Y. Yau, X. Lou, and N. S. V. Rao, "Markov Game analysis for attack-defense of power networks under possible misinformation," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1676–1686, May 2013.

[52] Y. Xiang and L. Wang, "A game-theoretic study of load redistribution attack and defense in power systems," *Electr. Power Syst. Res.*, vol. 151, pp. 12–25, Oct. 2017.

[53] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA: Morgan Kaufmann, Jul. 1998, pp. 242–250.

[54] K. Chatterjee, R. Majumdar, and M. Jurdziński, "On nash equilibria in stochastic games," in *Computer Science Logic*, J. Marcinkowski and A. Tarlecki, Eds. Berlin, Germany: Springer, 2004, pp. 26–40.

[55] M. Feng and H. Xu, "Deep reinforecement learning based optimal defense for cyber-physical system in presence of unknown cyber-attack," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–8.

[56] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Comput.*, vol. 6, pp. 1185–1201, Nov. 1994.

[57] A. Neyman and S. Sorin, *Stochastic Games and Applications*. New York, NY, USA: Kluwer, Jul. 1999.

[58] C. Szepesvári and M. L. Littman, "A unified analysis of value-function-based reinforcement-learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 2017–2060, 1999.

[59] D. P. Bertsekas and J. N. Tsitsiklis, "Convergence rate and termination of asynchronous iterative algorithms," in *Proc. 3rd Int. Conf. Supercomput. (ICS)*, New York, NY, USA, Jun. 1989, pp. 461–470.

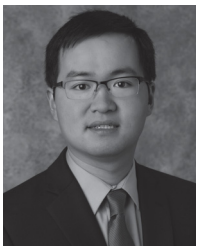
[60] S. R. Islam, D. Sutanto, and K. M. Muttaqi, "A decentralized multiagent-based voltage control for catastrophic disturbances in a power system," in *Proc. IEEE Ind. Appl. Soc. Annu. Meeting*, Oct. 2013, pp. 1–8.

[61] S. Satsangi, A. Saini, and A. Saraswat, "Clustering based voltage control areas for localized reactive power management in deregulated power system," *Int. J. Elect. Comput. Eng.*, vol. 6, no. 1, pp. 1348–1354, Jan. 2011.
[62] P. Song, Z. Xu, C. Luo, H. Cai, and Z. Xie, "Voltage sensitivity analysis based bus voltage regulation in transmission systems with UPFC series converter," in *Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2017, pp. 483–488.
[63] A. Onwuachumba, M. Musavi, and P. Lerley, "Identification of critical locations of power systems," in *Proc. 9th Annu. IEEE Green Technol. Conf. (GreenTech)*, Mar. 2017, pp. 290–296.

Shuva Paul (S'13–M'19) received the B.S. degree in electrical and electronics engineering and the M.S. degree from American International University-Bangladesh, Dhaka, Bangladesh, in 2013 and 2015, respectively, and the Ph.D. degree in electrical engineering and computer science from South Dakota State University, Brookings, SD, USA, in 2019.

His current research interests include computational intelligence, reinforcement learning, game theory, and smart grid security for power transmission and distribution systems, events and anomaly detection, and big data analytics.

Dr. Paul has been actively involved in numerous conferences, including the Session Chair of the IEEE EnergyTech in 2013 and the IEEE EIT (2019, USA). He also serves as a Reviewer for many reputed conferences, including the IEEE PES General Meeting (2019, USA, and 2020, Canada), SSCI (2017, Honolulu, HI, USA), IJCNN (2018, Brazil), PECI (2017, 2018, Chicago, USA), and ECCE (USA), and journals, including the IEEE TRANSACTIONS ON SMART GRID, *Neurocomputing*, the IEEE ACCESS, *The Journal of Engineering* (IET), and *Cyber-Physical Systems: Theory & Applications* (IET).



Zhen Ni (M'15) received the B.S. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in electrical, computer and biomedical engineering from The University of Rhode Island, Kingston, RI, USA, in 2015.

He is currently an Assistant Professor with the Department of Computer, Electrical Engineering, and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. His current research interests include computational intelligence, reinforcement learning, and smart grid applications. He was with the Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, USA, from 2015 to 2019.

Dr. Ni was a recipient of the INNS Aharon Katzir Young Investigator Award in 2019, the URI Excellence in Doctoral Research Award in 2016, and the Chinese Government Award for Outstanding Students Abroad in 2014. He will receive the prestigious IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award in 2020. He has been actively involved in numerous conference and workshop organization committees in the society, including the General Co-Chair of the IEEE CIS Winter School, Washington, DC, USA, in 2016. He was a Guest Editor of *Cyber-Physical Systems: Theory & Applications* (IET) from 2017 to 2018. He has been an Associate Editor of the *IEEE Computational Intelligence Magazine* since 2018 and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS since 2019.

Chaoxu Mu (M'15–SM'18) received the Ph.D. degree in control science and engineering from Southeast University, Nanjing, China, in 2012.

She was a Visiting Ph.D. Student with the Royal Melbourne Institute of Technology University, Melbourne, VIC, Australia, from 2010 to 2011, and Postdoctoral Fellow with the Department of Electrical, Computer and Biomedical Engineering, The University of Rhode Island, Kingston, RI, USA, from 2014 to 2016. She is currently a Professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. She has published over 100 journal and conference articles, coauthored two books, and authorized ten innovation patents. Her research interests include adaptive dynamics programming and approximate optimal control and learning-based control and algorithms as well as their applications.