A Game-Theoretical Approach to Cyber-Security of Critical Infrastructures Based on Multi-Agent Reinforcement Learning

Martina Panfili, Alessandro Giuseppi, Andrea Fiaschetti, Homoud B. Al-Jibreen, Antonio Pietrabissa, Francesco Delli Priscoli
Dept. of Computer, Control and Management Engineering University of Rome Sapienza
via Ariosto 25, 00185, Rome, Italy
{panfili, giuseppi, fiaschetti, aljibreen, pietrabissa,dellipriscoli}@diag.uniroma1.it

***Abstract*—This paper presents a control strategy for Cyber-Physical System defense developed in the framework of the European Project ATENA, that concerns Critical Infrastructure (CI) protection. The aim of the controller is to find the optimal security configuration, in terms of countermeasures to implement, in order to address the system vulnerabilities. The attack/defense problem is modeled as a multi-agent general sum game, where the aim of the defender is to prevent the most damage possible by finding an optimal trade-off between prevention actions and their costs. The problem is solved utilizing Reinforcement Learning and simulation results provide a proof of the proposed concept, showing how the defender of the protected CI is able to minimize the damage caused by his/her opponents by finding the Nash equilibrium of the game in the zero-sum variant, and, in a more general scenario, by driving the attacker in the position where the damage she/he can cause to the infrastructure is lower than the cost it has to sustain to enforce her/his attack strategy.**

Keywords — Stochastic Games, Reinforcement Learning, Vulnerability Management, Critical Infrastructure Protection, Composable Security.

I. INTRODUCTION

In recent years, the usage of game theory in several fields of security has increased. Game-theoretical approaches have been applied for studying cyber-physical security issues spacing from privacy preservation to Critical Infrastructure (CI) protection and industrial plant operation. The interested reader can find a relevant survey covering several recent works in [1]. The game-theoretical approach is suitable to model the interaction between agents, or players, that for security problems can be thought as attackers and defenders. In general, the attacker wishes to purposefully attack the target Cyber-Physical Systems (CPSs) maximizing the system corruption, while the system defender aims at minimizing such damage. Furthermore, the attacker's actions heavily depend on defender's strategy – modern attackers are able to exploit static security guidelines and common protocols. It is clear that, in this perspective, a two-player game, in which each player aims to maximize its own reward, can be employed to analyze the CPS security. In particular, in the project ATENA [2], we investigate the use of game theory to study the system decision-making strategy to determine optimal security system configurations. The concept of composable security was firstly introduced in the European projects pShield and nShield [3] and it is an aim of this work to expand it to the CPS and CI domain; in this respect, previous related work can be found, for instance, in [3]–[5]. The idea is to develop an "off-line" control which, by optimizing the current system configuration, is aimed at improving the security level of the protected system in the long term, assuring what may be defined as *structural security*, in contrast with the reaction to cyber-physical attacks that only guarantees an instantaneous response to threats.

In this section, we provide an overview of related works in the context of cyber-physical security. Zhu and Basar [6] have proposed a dynamic game-theoretical method to model the interaction between the cyber and physical systems, integrating robust system controllers of the physical space and resilient controllers of the cyber-space, modeled by a Markov process that depends on the action of defenders and attackers. In this work, the authors apply dynamic programming algorithms to achieve a saddle point equilibrium. A related work based on multi-objective reinforcement learning was proposed by Tozer et al. in [7]. Considering only cyber systems, Zhu and Basar in [8] introduce a proactive defense scheme, based on game theory, that dynamically modifies the attack surface of a cyber system to make it difficult for attackers to gather system information. Manadhata in [9] has proposed a game-theoretical approach that the system defender can use to optimally shift and reduce the system attack surface. For a dynamic cyber-security problem, Rasouli et al. in [10] have considered a min-max performance criterion and used dynamic programming to determine, within a restricted set of policies, an optimal policy for the defender. Ma et al. [11] have integrated cyber and physical space using

a payoff function; in this work, the authors have presented several game-theoretical formulations of attack and defense aspects of cyber-physical systems under different cost/reward functions. A similar approach is presented by He et al. [12]; in this work, the defender's strategy is based on Nash equilibrium and the authors analyze the sensitivities with respect to cyber and physical correlation coefficients, target revenues and costs. A modeling approach to evaluating the security of cyber-physical systems was presented in [13], in which a game-theoretical paradigm predicts the interactions between the attacker and the system. Differently from the aforementioned works, in [14] the authors have proposed a game and reliability framework involving a multi-objective approach and imperfect information so as to support decision-makers in choosing efficiently designed security systems. Other approaches, related to Multi-Agent Reinforcement Learning, have been proposed, for instance in [15], with respect to the problem of assuring, in real-time, to the users a satisfactory Quality of Experience.

In contrast with the reactive on-line game-theoretical approaches of the mentioned researches, this paper proposes a game-theoretical approach to find a proactive off-line strategy which has as the objective of maximizing the expected defense level against cyber-physical attacks, i.e., from a different viewpoint, of minimizing the expected damage that a smart attacker is able to produce to the infrastructure. The output of the proposed algorithm will be a set of security functionalities/improvements (e.g., placing a redundant power generator, setting up an encrypted communication channel or reinforcing a gate, …), hereafter referred to as *security countermeasures*, to be implemented during the system normal operation conditions to increase the system security to expected future attacks.

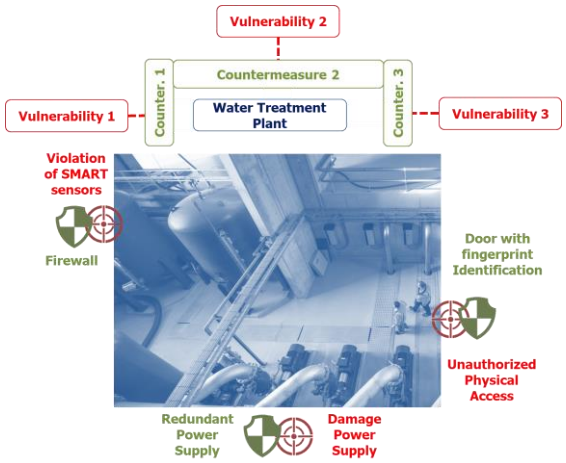

Figure 1 Example of CI subsystem and its protection strategies

Figure 1 reports a possible applicative scenario of a subsystem protected in the framework of project ATENA, and consist of a water storage station for water treatment. Among all the possible countermeasures available, the defender should select a combination that represents the optimal trade-off between implementation cost and damage prevention. Even in the very simplified example above, in which the countermeasures selected are reported in green and the potential attacks identified may exploit the vulnerabilities in reported in red, the reader may notice how a single countermeasure can affect multiple vulnerabilities (e.g. access control devices may increase the security related to the power supply). The work of this paper aims at the maximization of the overall protected system's defense level and the one of its interdependent CIs [16], for which a multi-agent approach is proposed.

II. Preliminaries on Stochastic Games

In this section, we provide a brief review of some basic concepts and definitions of game theory with particular attention to security games.

A generic strategic game G is generally defined by a tuple:

$$G = \{N, S_{i=1,\dots N}, f_{i=1,\dots N}\} \quad (1)$$

where N is set of players, $S_i, i = 1, \dots, N$, is the strategy set of each player i, generating the joint strategy set $S = S_1 \times S_2 \times \dots \times S_N$, and $f_i: S_i \to \mathbb{R}, i = 1, \dots, N$ is the payoff function of player i.

The outcome *pure* strategy s is the collection of the selected actions, one for each player: $s \coloneqq \{(s_i)_{i=1,\dots,N} | s_i \in S_i, i = 1, \dots N\} \subseteq S$. Strategies of this type are called *pure* strategies since each player i selects deterministically one action s_i from his/her strategy set $S_i, i = 1, \dots, N$. At a Nash equilibrium, no player has an incentive (i.e., payoff improvement) to take a different action with respect to the equilibrium strategy $s^* = (s_i^*)_{i=1,\dots,N}$: in this strategic game G, the strategy s^* is at Nash equilibrium if

$$f_i\left(s_i^*, s_{\setminus i}^*\right) \geq f_i\left(s_i, s_{\setminus i}^*\right), \forall s_i \in S_i, i = 1, \dots, N, \quad (2)$$

where $s_{\setminus i}^*$ is the profile of the actions taken by all the players other than player i.

In some problem formulations, the player is allowed to choose stochastically his/her actions. In this case, the player is using a mixed strategy. In particular, a mixed strategy π_i for player i is a probability distribution over the set of available actions S_i. In other words, for each player i, the mixed strategy consists in a $|S_i|$ dimensional vector:

$$\left(\pi_i\left(s_i^1\right), \dots, \pi_i\left(s_i^{|S_i|}\right)\right)$$
$$\text{with } \pi_i\left(s_i^k\right) \geq 0, \sum_{k=1,\dots,|S_i|} \pi_i\left(s_i^k\right) = 1, i = 1, \dots, N \quad (3)$$

Up to now, we have assumed that all players have perfect information of the game, i.e., the players' payoff functions and strategies are known. In this case, the game is called a complete information game. Conversely, if some players have private information, the game is called incomplete information game. In these games, at least one player has incomplete information regarding one or more game elements, such as payoff functions or other players' available actions and behavior.

In this paper, we are considering security games, a special class of games that study the interaction of malicious attackers and defenders. In the literature ([1], [6], [8], [11]) most of the security games are modeled as zero-sum games.

The game $G = \{N, S_{i=1,\dots N}, f_{i=1,\dots N}\}$ is a zero-sum game if

$$\sum_{i=1..N} f_{i(s_1,\dots,s_N)} = 0, \ \forall s \in S \quad (4)$$

In particular, a two-player zero-sum game is a game such that:

$$f_1(s) = -f_2(s), \forall s \in S \quad (5)$$

Finally, we briefly mention the class of stochastic games, which are the generalization of Markov decision processes (MDPs) to the multi-agent case. A stochastic game (SG) is a tuple $\{X, S_1, \dots, S_N, p, f_1, \dots, f_N\}$, where X is the discrete set of environment states, the S_i's are the agent strategies, $p: X \times S \times X \to [0,1]$ is the state transition probability function, such that $p(x, s, x')$ is the probability that the system reaches state x' when the agents use the strategy s in state $x \in X$, and $f_i: X \times S \times X \to \mathbb{R}, i = 1, \dots, N$, is the reward function of agent i, such that $f_i(x, s, x')$ is the reward incurred by agent i when the system reaches state x' from state x under strategy s. Note that, in the multi-agent case, the state transitions and the reward depend on the joint actions of all the agents. The various agent policies $\pi_i: X \times S_i \to [0,1]$ form together the joint policy π.

MDPs can be interpreted as single-agent SGs and can be solved by means of reinforcement learning algorithms [17] without the need of the knowledge of the state transition probabilities. A widely used algorithm is the Q-learning, which estimates the state-action value[1] functions with the following rule:

$$Q(x, s) \leftarrow (1 - \alpha) Q(x, s) + \alpha \left[r(x, s, x') + \gamma \max_{s' \in S} Q(x', s') \right] \quad (6)$$

Where the, potentially time-varying, parameters $0 < \alpha < 1, 0 < \gamma < 1$ are the learning rate and the discount factor. The learning rate models how much the current episode/reward should update the knowledge stored in $Q(x, s)$, while the discount factor captures the trade-off between immediate reward, and consequent greedy behavior, with long-term performances.

The action to be taken when the system is in state x is selected according to the ε-greedy rule:

$$\pi(x) = \begin{cases} \underset{s' \in S}{\operatorname{argmax}} Q(x, s') \text{ with probability } \varepsilon_t \\ \text{random action with probability } 1 - \varepsilon_t \end{cases} \quad (7)$$

where $\varepsilon_t \in (0,1)$ drives the trade-off between exploitation and exploration of the state-action space. Its dependency on simulation time t makes the policy (7) a so-called dynamic ε-greedy, as the controller may decrease the value of ε_t as the time increases, in order to exploit the acquired knowledge more, while encouraging exploration at the start of its training.

Similarly, SGs, i.e., multi-agent MDPs, can be solved by means of multi-agent Reinforcement Learning (MARL) algorithms. A multi-agent Q-learning approach for SGs can be defined as follows:

$$Q_i(x, s) \leftarrow (1 - \alpha) Q_i(x, s) + \alpha \left[f_i(x, s, x') + \right.$$
$$\left. + \gamma \, eval_i \left(\pi_i(x') Q_i(x', \pi_i(x')) \right) \right], i = 1, \dots N; \quad (8)$$

where $eval_i$ gives the expected return of agent i given the agent's strategy π_i, derived according to (7). Considering a generic mixed SG, where no constraints are imposed on the reward functions of the agents, the function π_i could lead to a particular type of equilibrium depending on the modelled problem, such has a Nash equilibrium [18], [19], a correlated equilibrium [20], a Stackelberg equilibrium [21], or a best-response strategy [22]. The reader may notice how in this formulation the state and the action x, s lack the subscript i. This is due to the fact that in the framework introduced they represent the aggregated state and policy for the overall system, while the matrix Q_i is not shared among the players

In [23], Littman presents a convergent algorithm, denoted with friend-or-foe Q-learning (FFQ), that is proved to converge in fully cooperative SGs – where all agents are collaborating to maximize their cumulative reward – and in fully competitive SGs – where agents can be divided into *friends* and *foes*. Furthermore, in a fully cooperative SG, if a centralized controller is available, the task is reduced to a Markov decision process (in this case the action space is the joint action space of the SG) and the goal can be achieved by learning the optimal joint-action values with the simple Q-learning rule (6).

III. Proposed Approach

In this section, we present a preliminary game theoretical approach to design the ATENA Composer Module, in charge of finding the optimal security configuration for off-line damage prevention strategies.

Critical Infrastructures can be modeled as Cyber-Physical Systems (CPS), as their control and operation heavily depend on connectivity functionalities, such as in SCADA systems [24], and in complex software/hardware components, as in Smart Grids. In particular, the network consisting in the interconnection of interdependent Critical Infrastructures is a complex physical-cyber-organizational system of systems that plays an extremely important role in modern society.

In the proposed formulation we consider the various interconnected CIs as a connected set of their most critical subsystems (e.g., water treatment plants for water distribution networks, gas turbines for power networks, …). We denote such subsystems as $sys_i \in SYS, i = 1, \dots, N_D$, where N_D is the number of subsystems protected and SYS is the set of all the subsystems present in the protected interconnected CIs. Furthermore, we consider a number N_D of defenders, implying that every critical subsystem has its own defender. Each subsystem is itself constituted by the interconnection of various elements that we will refer to as assets (e.g. water tank, electrical storage device, access controller, …).

[1] $Q^{\pi}(x, s)$ denotes the reward of the system when the action s is chosen in state x and following the strategy π thereafter.

Regarding the attackers, we can assume that they can perform an attack starting from their knowledge of the system vulnerabilities. For each considered subsystem sys_i, let $V_c(sys_i)$ and $V_p(sys_i)$ represent the sets of vulnerabilities in the cyber and in the physical domain, respectively. It follows that any attack that the attacker is able to launch is a combination of vulnerabilities contained in the set $V(sys_i) = V_c(sys_i) \cup V_p(sys_i)$. In the "prevent" perspective, in order to evaluate the worst cases of attacks, we can consider different potential adversaries characterized by peculiar features and resources (e.g., skilled hackers or organized armed criminal groups), in such a way that the available attacks (i.e., combination of exploited vulnerabilities) are unique to each individual attacker.

The proposed formulation is a state-less (i.e., $X = \emptyset$) general-sum game that can be defined by the following tuple:

$$\{P, S^D, S^A, r^D_{i=1,\dots N_D}, r^A_{j=1,\dots N_A}\} \tag{9}$$

In which:

- P is a finite set of N players;
- $S^D = S^D_1 \times \dots \times S^D_{N_D}$ is the defenders' joint strategies set, where S^D_i is the finite set of pure strategies of defender i, with $i = 1, \dots, N_D$;
- $S^A = S^A_1 \times \dots \times S^A_{N_A}$ is the attackers' joint strategies set, where S^A_j is the finite set of pure strategies of attacker j, with $j = 1, \dots, N_A$;
- $r^D_i: S^D \times S^A \to \mathbb{R}$, is the payoff function of defender i, with $i = 1, \dots, N_D$;
- $r^A_j: S^D \times S^A \to \mathbb{R}$, is the payoff function of attacker j, with $j = 1, \dots, N_A$.

A. Game Players

We consider $N = N_D + N_A$ players, including N_D defenders and N_A attackers. Hence, the finite set P of players is composed of the set D of defenders and the set A of attackers. The defender player $i \in D$ can define each other defender as a "Friend" and each attacker as a "Foe". The FFQ approach assumes that player i and her/his friends act to maximize the payoff of player i, and that the foes of i act together to minimize the payoff of player i (and, consequently, to maximize their own payoff). In other words, the attackers act together to maximize the damage caused to the interconnected CIs, while the defenders act together to maximize their payoff, which should represent the prevented damage.

B. Game Strategies

Let $C_i := \{c_{i,1}, c_{i,2}, \dots, c_{i,nc_i}\}$ be the set of nc_i available countermeasures for the assets protected by the defender player i. The defender i's strategy set is then

$$S^D_i = \{s^D_i = (c_{i,1}, c_{i,2}, \dots, c_{i,nc_i}) | c_{i,k} \in \{0,1\}, k = 1, \dots, nc_i\}$$

where $c_{i,k} = 1$ if the decision is to enable the countermeasure $c_{i,k}$, and $c_{i,k} = 0$ otherwise. The joint action chosen by all the defenders is denoted as $s^D = \{s^D_1, s^D_2, \dots, s^D_{N_D}\}$.

Similarly, we can define the attacker j's strategy set as

$$S^A_j = \{s^A_j = (v_{j,1}, \dots, v_{j,nv_i}) |\ v_{j,h} \in \{0,1\}, h = 1, \dots, nv_j\}$$

where nv_j is the number of vulnerabilities $v_{j,h}$ present in the system in its constituting assets that the attacker j can exploit, and the attackers' joint action is $s^A = \{s^A_1, s^A_2, \dots, s^A_{N_A}\}$.

C. Payoff functions

To model the attackers' and defenders' reward functions we should capture the impact that the various attacks have on the protected CIs. By definition, the countermeasures mitigate such impact, meaning that if a countermeasure $c_{i,k}$ is in place the damage that the attacker i can deal is reduced by a prevent factor $m(v_{j,h}, c_{i,k}) \in [0,1]$ that depends on the vulnerability considered and the given countermeasure. Furthermore, this impact shall depend on the criticality of the vulnerability and of the subsystem on which the vulnerability is present, capturing also its economic value.

In principle, the various countermeasures may or may not address different aspects of the same vulnerability (i.e. implementing two countermeasures that mitigate 50% of the attack impact does not imply that the vulnerability has been completely canceled out); for the sake of simplicity, we assume that if two countermeasures address the same vulnerability, the percentage of damage prevented depends on the vulnerability that has the highest prevent factor. The above discussion can be summarized in the following expression for the reward of attacker i:

$$r^A_j = \sum_{h=1,\dots,nv_j | v_{j,h}=1} \left(\Gamma(v_{j,h}) \left(1 - \max_{i=1,\dots,nc_i | c_{i,k}=1} m(v_{j,h}, c_{i,k}) \right) \right) \tag{10}$$

where $\Gamma(v_{j,h}) \in \mathbb{R}$ represents the criticality of the vulnerability $v_{j,h}$ taking also into account the value of the asset on which it is present that would be compromised by the attack. The term $\max_{i=1,\dots,nc_i | c_{i,k}=1} m(v_{j,h}, c_{i,k})$ is the maximum prevent factor among the countermeasures that the defenders chose to implement. This reward may also include a term representing the cost sustained by the attacker to exploit the selected vulnerabilities. Regarding the defenders, we want to capture the prevented damage, and this is archived by taking into account the following reward function:

$$r^D_i = -\sum_{j=1,\dots,N_A} \sum_{\substack{h | v_{j,h}=1, \\ v_{j,h} \in V(sys_i)}} \left(\Gamma(v_{j,h}) \left(1 - \max_{k=1,\dots,nc_i | c_{i,k}=1} m(v_{j,h}, c_{i,k}) \right) \right) \tag{11}$$

representing that each defender obtains a portion of its reward from each of the attackers that are expected to try to exploit one of the vulnerabilities in the protected system. The cost term that may be added in this case depends on the countermeasure costs. We note that, in security issues, the defense costs are usually much larger than the attack ones.

D. Friend or foe Q-learning

In this section we describe the method for learning the players' Q-function, using FFQ algorithm, in the proposed ATENA game-theoretical approach. The FFQ-learning

approach could be described by the following equations derived by eq. (8):

$$Q_i^D(s^D, s^A) \leftarrow (1-\alpha)Q_i^D(s^D, s^A) + \alpha\left[r_i^D(s_i^D, s^A) + \gamma FFQ_i^D\right], \quad i = 1, \dots N_D \tag{12}$$

$$Q_j^A(s^D, s^A) \leftarrow (1-\alpha)Q_j^A(s^D, s^A) + \alpha\left[r_j^A(s_j^A, s^D) + \gamma FFQ_j^A\right], \quad j = 1, \dots N_A \tag{13}$$

$$FFQ_i^D = \max_{s^D \in S^D} \min_{s^A \in S^A} Q_i^D(s^D, s^A) \tag{14}$$

$$FFQ_j^A = \max_{s^A \in S^A} \min_{s^D \in S^D} Q_j^A(s^D, s^A) \tag{15}$$

In this preliminary version, we have considered a simplified scenario consisting of two deterministic opponents, representing an ATENA operator defender and an external attacker.

The output of the proposed approach is the system optimal prevention strategy in terms of countermeasures activated or, in other words, the optimal security configuration. Then, this output can be fed into a decision support system to propose long-term security improvements.

IV. Simulation Results

The simulation scenario considers a subsystem of a critical infrastructure protected by an ATENA controller (defender) aimed at defending against a malicious agent (attacker). We can think of the following scenario as the fundamental building block of the approach presented, and, for the sake of clarity, we decided to model a scenario of the same nature of the one present in Figure 1. We assume that the attacker has knowledge on $nv = 6$ vulnerabilities in the system and it may choose to attack any combination of them. Equivalently, the defender may choose to implement any combination of the countermeasures at its disposal, whose number nc has been set to 6.

The aim of this work is to develop a domain independent defense strategy, so the actual nature of the protected system and its vulnerabilities are not of particular interest. In project ATENA the scenarios considered for the validation of the overall suite of tools consist of systems from the power, water and gas networks domain, and hence the subsystem we considered in this scenario could represent a portion of a water treatment plant, consisting of three assets: (i) a set of water tanks; (ii) a water heating system; (iii) a pump system. The relation of these assets with the vulnerabilities and countermeasures considered is reported in the table below.

Asset 1	Asset 2	Asset 3
Vulnerability 1,2,3	Vulnerability 4,5	Vulnerability 6
Countermeasure 1,2,3	Countermeasure 3,4,5	Countermeasure 1,2,6

Table 1: Mapping between asset, vulnerabilities, and countermeasures

For instance, in the scenario considered, by selecting the countermeasure 3 the defender is mitigating the damage on both assets 1 and 2, while attacking the vulnerabilities 1,2,3 will only deal damage to the first asset. We can think of countermeasure 3 as an access control system that partially addresses physical vulnerabilities linked to the water tanks and the heating system, while it doesn't do anything for the already physically secured heating system. Vulnerability 6, on the other hand, may represent a cyber-domain vulnerability on the remote control system for the heating system. The asset values were selected from a random uniform distribution of mean 5, while the criticality and prevent factors m for the various vulnerabilities and countermeasures were selected with random uniform distribution between 0 and 95%. The product of value and criticality was used to describe the function $\Gamma(v_{j,h})$ of equations (10) and (11), and due to the summations present in these formulas, as well as the relations in table 1, this simplified scenario is able to capture the interdependency between the assets considered, that can be translated into CI interdependencies in more complex scenarios.

For the First Simulation we decided to not consider the attack and the defense costs, and hence the game is reduced to a zero-sum one, i.e. the reward of the attacker is the opposite of the defender's one.

The algorithm parameters were set as follows: $\alpha = \frac{0.9}{1+\lfloor \frac{t}{100} \rfloor}$ where t is the iteration number, and the operator $\lfloor\ \rfloor$ represents the floor operator; ε_t of equation () was set at 0.9 and reduced by 25% after every 1000 iterations.

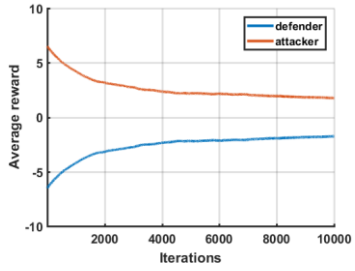

Figure 2 First simulation reward comparison

Figure 2 reports the rewards obtained by the two agents over the number of episodes, averaged every 500 iterations. It can be seen how convergence is obtained after few iterations, and how the defender strategy successfully manages to reduce the expected damage of the critical assets (note that, in this the simulation setup, the countermeasures only address a percentage of the vulnerabilities, hence the defender is not able to drive the attacker reward to zero).

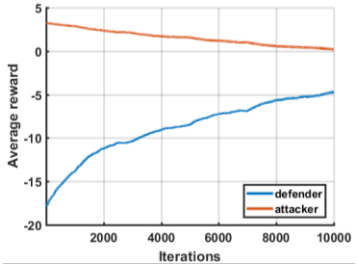

Figure 3 Second simulation reward comparison

In the Second Simulation, we considered a more general scenario, in which we took into account different defense and attack costs, respectively to set up the defense and attack strategies. We assumed that the vulnerability exploitation

costs are sampled from a unitary uniform distribution, whereas the countermeasure implementation costs were sampled from a five-time larger distribution to capture that, once a vulnerability is known, typically the cost for the attacker to exploit it is marginal compared to the cost of patching/resolving it.

Figure 3 shows how, in this case, both the rewards are lowered due to the additional costs. It is interesting to notice that the attacker's reward is driven to zero, meaning that the reward gained by the attacker is equalized by the cost to launch the attack. In this scenario the defender was hence able to put the attacker in a situation where the damage it can cause does not justify its cost, archiving the ideal configuration for preventive defense. It can be noticed how the defender agent, dealing with much greater costs, performs significantly worse than in the previous case, but still manages to find the optimal configuration in terms of price per performance.

ACKNOWLEDGMENT AND FUTURE WORKS

Future works will include the validation of the approach presented in a more complex and realistic scenario that involves systems representing interconnected Cis, as well as the integration of security standards metrics and practices in the formulation. The authors acknowledge the CRAT team involved in project ATENA as well as the other ATENA partners.

REFERENCES

[1] C. T. Do *et al.*, "Game Theory for Cyber Security and Privacy," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–37, May 2017.

[2] A. Consortium, "ATENA website," *2017*. [Online]. Available: https://www.atena-h2020.eu.

[3] A. Fiaschetti, V. Suraci, and F. Delli Priscoli, "The SHIELD framework: How to control Security, Privacy and Dependability in complex systems," in *2012 Complexity in Engineering (COMPENG). Proceedings*, 2012, pp. 1–4.

[4] S. Canale *et al.*, "Resilient planning of PowerLine Communications networks over Medium Voltage distribution grids," in *2012 20th Mediterranean Conference on Control & Automation (MED)*, 2012, pp. 710–715.

[5] P. Capodieci *et al.*, "Improving Resilience of Interdependent Critical Infrastructures via an On-Line Alerting System," in *2010 Complexity in Engineering*, 2010, pp. 88–90.

[6] Q. Zhu and T. Basar, "Robust and resilient control design for cyber-physical systems with an application to power systems," in *IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 4066–4071.

[7] B. Tozer, T. Mazzuchi, and S. Sarkani, "Optimizing Attack Surface and Configuration Diversity Using Multi-objective Reinforcement Learning," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 144–149.

[8] Q. Zhu and T. Başar, "Game-Theoretic Approach to Feedback-Driven Multi-stage Moving Target Defense," 2013, pp. 246–263.

[9] P. K. Manadhata, "Game Theoretic Approaches to Attack Surface Shifting," Springer, New York, NY, 2013, pp. 1–13.

[10] M. Rasouli, E. Miehling, and D. Teneketzis, "A Supervisory Control Approach to Dynamic Cyber-Security," Springer, Cham, 2014, pp. 99–117.

[11] 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 *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011, pp. 708–713.

[12] F. He, J. Zhuang, N. S. V. Rao, C. Y. T. Ma, and D. K. Y. Yau, "Game-theoretic resilience analysis of Cyber-Physical Systems," in *2013 IEEE 1st International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA)*, 2013, pp. 90–95.

[13] 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, Jun. 2017.

[14] I. D. Lins, L. C. Rêgo, M. das C. Moura, and E. L. Droguett, "Selection of security system design via games of imperfect information and multi-objective genetic algorithm," *Reliab. Eng. Syst. Saf.*, vol. 112, pp. 59–66, Apr. 2013.

[15] L. Ricciardi Celsi *et al.*, "A Q-Learning based approach to Quality of Experience control in cognitive Future Internet networks," in *2015 23rd Mediterranean Conference on Control and Automation (MED)*, 2015, pp. 1045–1052.

[16] A. Di Giorgio and F. Liberati, "Interdependency modeling and analysis of critical infrastructures based on Dynamic Bayesian Networks," in *2011 19th Mediterranean Conference on Control & Automation (MED)*, 2011, pp. 791–797.

[17] P. R. Montague, "Reinforcement Learning: An Introduction, by Sutton, R.S. and Barto, A.G.," *Trends Cogn. Sci.*, vol. 3, no. 9, p. 360, Sep. 1999.

[18] J. Hu, J. Hu, and M. P. Wellman, "Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm," 1998.

[19] J. Hu and M. P. Wellman, "Nash Q-Learning for General-Sum Stochastic Games," *J. Mach. Learn. Res.*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[20] A. Greenwald and K. Hall, "Correlated Q-Learning," in *ICML'03 Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 2003, pp. 242–249.

[21] V. Kononen, "Asymmetric multiagent reinforcement learning," in *IEEE/WIC International Conference on Intelligent Agent Technology, 2003. IAT 2003.*, pp. 336–342.

[22] M. Weinberg and J. S. Rosenschein, "Best-Response Multiagent Learning in Non-Stationary Environments," *Proc. Third Int. Jt. Conf. Auton. Agents Multiagent Syst. - Vol. 2*, pp. 506–513, 2004.

[23] M. M. L. Littman, "Friend-or-foe Q-learning in general-sum Games," in *ICML*, 2001, vol. 1, pp. 322–328.

[24] A. Mercurio, A. Di Giorgio, and P. Cioci, "Open-Source Implementation of Monitoring and Controlling Services for EMS/SCADA Systems by Means of Web Services— IEC 61850 and IEC 61970 Standards," *IEEE Trans. Power Deliv.*, vol. 24, no. 3, pp. 1148–1153, Jul. 2009.