Identification of Sensor Replay Attacks and Physical Faults for Cyber-Physical Systems

Kangkang Zhang, Christodoulos Keliris, Thomas Parisini, *Fellow, IEEE* and Marios M. Polycarpou, *Fellow, IEEE*,

***Abstract*— This letter proposes a threat discrimination methodology for distinguishing between sensor replay attacks and sensor bias faults, based on the specially designed watermark integrated with adaptive estimation. For each threat type, a watermark is designed based on the changes that the threat imposes on the system. Threat discrimination conditions are rigorously investigated to characterize quantitatively the class of attacks and faults that can be discriminated by the proposed scheme. A numerical simulation is presented to illustrate the effectiveness of our approach.**

***Index Terms*— Threat discrimination, replay attack, sensor bias fault, Cyber-physical systems**

I. Introduction

Cyber-physical systems (CPS) are widely utilized in modern automation processes. Unfortunately, vulnerabilities to malicious cyber threats increase greatly due to the complex integration of computing, communication and control in CPS [1]–[3]. Developing malicious cyber attack detection and identification techniques is crucially important.

Replay attacks are commonly used in practical systems due to the simplicity in implementation and some key stealthiness properties. For example, the Stuxnet attack on the Iranian nuclear facilities was a type of replay attacks. The attacker steals access to the communication links and records data from the normal operation and then replays it to the supervisory system [4]. Hence, replay attacks possess high stealthiness as a result of the used malicious attack data taken from the normal system operation. Moreover, replay attacks can hide other types of non-stealthy cyber attacks. In more detail, the non-stealthy attacks occurring during the replaying procedure of a replay attack can remain concealed from typical anomaly detectors due to the cover provided by the replay attack. In addition, identification of replay attacks and sensor bias faults is more challenges since they both occur in the sensor-to-controller channels of a CPS, which motives this paper to consider replay attacks and sensor bias faults.

In the past decade, detection methodologies for integrity attacks [5]–[7] based on dynamic models has been rigorously investigated, and are divided into two categories [3]: a) active detection approaches, such as the moving target method [8] and the watermark approach [5], [9], and b) extension of anomaly diagnosis approaches, such as the ones in [10]–[12]. A key issue not yet adequately addressed is the threat discrimination problem between attacks and other anomalies such as physical faults. Typical anomaly detectors may be able to detect and locate anomalies in CPS, but are not capable to distinguish between different types of the occurring anomalies. Threat discrimination fills this gap, aiming to identify the threat type, namely to determine whether a cyber attack or another anomaly (e.g., physical faults) is occurring. Threat discrimination is important for practical CPS, since it helps the operators to make correct decisions and take suitable and different remediation actions for attacks and faults, respectively. Naturally, targeted cyber attack events are more dangerous, which may require immediate response actions. Hence, the accommodation strategies against cyber attacks and physical faults are usually different. Ignoring a "small" fault in a sensor might not affect the system safe operation. However, an attack on a sensor might require to shut down the whole process to avoid catastrophic consequences. Physical maintenance, such as replacing communication cables, can be effective in preventing physical faults, but cannot remediate the issues caused by cyber attacks. Updating communication protocols and firewalls are the general prevention approaches against cyber attacks. More approaches for preventing and mitigating cyber attacks can be found in [3]. Moreover, the threat discrimination problem between replay attacks and physical sensor bias faults remains an open problem. The stealthiness characteristics of replay attacks prevents typical anomaly detectors from distinguishing between replay attacks and sensor faults. On the other hand, typical fault isolation schemes mainly focus on gaining information about the location of the faults whereas the threat type discrimination problem is usually overseen.

This letter proposes an approach for discriminating between the occurrence of replay attacks and sensor faults. Specifically, a linear parameterization form of the replay attacks is proposed for the first time, contributing to attack parameter estimation. The designed adaptive observer provides a procedure for estimating the outputs of the system in the nominal scenario, which is novel in the attack case and in the fault case. Also, for the residual and threshold generation, two signal processors (integrated with the watermarks) are introduced and are designed for the replay attack case and the sensor fault case distinctively in order to design suitable residual generator and threshold signals. Finally, the discrimination ability is also rigorously investigated to characterize quantitatively the class of attacks and faults that can be identified by the proposed methodology.

II. Problem Formulation

In this paper, we consider a type of typical CPS depicted in Fig. 1, which consists of a linear time-invariant physical plant $\mathcal{P}$, a sensor data communication network $\mathcal{N}_s$, an output-feedback controller $\mathcal{C}$ and an anomaly detector $\mathcal{D}$. The closed-loop CPS in the nominal case (no attacks and faults), is described by

$$\mathcal{W}_n : \left\{ \begin{array}{rcl} \dot{x}_n & = & Ax_n + BK\tilde{y}_n, \\ \tilde{y}_n & = & y_n = Cx_n, \end{array} \right. \tag{1}$$

where $x_n \in \mathbb{R}^{n_p}$ is the state, and $y_n \in \mathbb{R}^{n_y}$ is the output. The variable $\tilde{y}_n \in \mathbb{R}^{n_y}$ indicates the output of $\mathcal{N}_s$. Moreover,

This work has been supported by the European Union's Horizon 2020 Research and Innovation Program (grant no. 739551 (KIOS CoE)), the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN) (grant no. 2017YKXYXJ), the National Natural Science Foundation of China (grant no. 61903188), and the Natural Science Foundation of Jiangsu Province (grant no. BK20190403).

K. Zhang, C. Keliris and M. Polycarpou are with the KIOS Research and Innovation Center of Excellence and the Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus (e-mail: zhang.kangkang@ucy.ac.cy; keliris.chris@gmail.com; mpolycar@ucy.ac.cy).

T. Parisini is with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K., with the Dept. of Engineering and Architecture, University of Trieste, Trieste, 34127, Italy, and with the KIOS Research and Innovation Center of Excellence, Cyprus (e-mail: t.parisini@gmail.com).

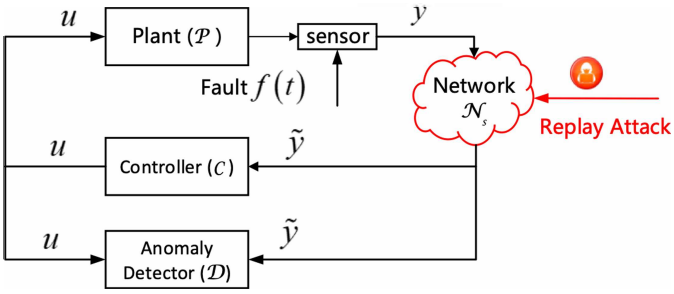

Fig. 1. Diagram of CPS under sensor replay attacks and physical faults.

$A \in \mathbb{R}^{n_p \times n_p}$, $B \in \mathbb{R}^{n_p \times n_u}$, $K \in \mathbb{R}^{n_u \times n_y}$ and $C \in \mathbb{R}^{n_y \times n_p}$ are known matrices, $A+BKC$ is a Hurwitz matrix and the pair (A, C) is observable. In this paper, we consider two types of threats: 1) a replay attack in $\mathcal{N}_s$; and 2) constant sensor bias faults. In addition, we consider that an anomaly is detected at a time T_d by the anomaly detector $\mathcal{D}$ in Fig. 1.

1) Replay Attack Scenarios: In general, replay attacks are equipped with recording and replaying functionalities [5]. For the considered replay attack in this letter, the adversary first records y_n communicated through the network $\mathcal{N}_s$ starting at a time $T_a - T$ and for a recording time T. Then, the replaying procedure starts at T_a and ends at $T_a + T$. Therefore, the replaying attack signal during $[T_a, T_a+T)$ is the output of $\mathcal{W}_n$ during the time interval $[T_a-T, T_a)$. Thus, the virtual attack model [5] is

$$\mathcal{W}'_n : \begin{cases} \dot{x}'_n &= Ax'_n + BK\tilde{y}'_n, \\ \tilde{y}'_n &= y'_n(t) = Cx'_n, \end{cases} \tag{2}$$

where $x'_n(t) \triangleq x_n(t-T)$, $\tilde{y}'_n(t) \triangleq \tilde{y}_n(t-T)$ and $y'_n(t) \triangleq y_n(t-T)$. Let x_a, y_a and $\tilde{y}_a$ denote the state, output and the received output of $\mathcal{W}_n$ in the replaying procedure of the replay attack. Then, we have $\tilde{y}_a = \tilde{y}'_n = y'_n$, and by defining a virtual attack signal $a(t) \triangleq y'_n(t) - Cx_a(t)$, the system $\mathcal{W}_n$ in the replaying procedure of a replay attack is described by

$$\mathcal{W}_a : \begin{cases} \dot{x}_a &= Ax_a + BK\tilde{y}_a, \\ \tilde{y}_a &= Cx_a + a(t). \end{cases} \tag{3}$$

In terms of x_a and x'_n, we can derive from (2) and (3) that $a(t) = Ce^{A(t-T_d)}(x'_n(T_d) - x_a(T_d))$, which indicates that $a(t)$ can be linearly parameterized as

$$a(t) = F^a(t)\theta^a, \ \forall\, t \in [T_a, T_a + T)\,, \tag{4}$$

where $F^a(t) \triangleq Ce^{A(t-T_d)}$ (known) and $\theta^a \triangleq x'_n(T_d) - x_a(T_d)$ is the unknown attack parameter vector and is supposed to satisfy the following assumption.

Assumption 1. The attack parameter θ^a is constant and bounded by a scalar $\sigma_a > 0$ known by the defender, i.e.,

$$\theta^a \in \Theta^a \triangleq \{\theta \in \mathbb{R}^{n_p} |\ |\theta| \le \sigma_a\}. \tag{5}$$

2) Sensor Bias Fault Scenarios: In this work, we consider that one/ multiple sensor bias faults occur at time T_f. In the presence of the sensor bias faults, $\mathcal{W}_n$ in (1) is described for $t \ge T_f$ by:

$$\mathcal{W}_f : \begin{cases} \dot{x}_f &= Ax_f + BK\tilde{y}_f, \\ \tilde{y}_f &= y_f = Cx_f + F^f(t)\theta^f, \end{cases} \tag{6}$$

where x_f is the state of the plant under the fault, $F^f(t)$ is a known $n_y \times n_y$ matrix, and $\theta^f \in \mathbb{R}^{n_y}$ represents the unknown constant fault vector. In this paper, we consider the special case of constant bias faults, in which case, $F^f(t) = I_{n_y \times n_y}$.

Remark 1. The discrimination scheme can also be applied to the case of time-varying sensor faults that can be linearly parameterized exactly as $F^f(t)\theta^f$, given that $F^f(t)$ is known by the defender and is also sufficiently different from $F^a(t)$. The reason is that the discrimination scheme relies on adaptive approximation methods for estimating the unknown parameters θ^a and θ^f, and hence, the matrix F^a and F^f should be sufficiently different to allow the creation of a sufficient mismatch for threat discrimination. ∇

Assumption 2. The fault parameter θ^f is constant and bounded by a scalar $\sigma_f > 0$ known by the defender, i.e.,

$$\theta^f \in \Theta^f \triangleq \{\theta \in \mathbb{R}^{n_y} |\ |\theta| \le \sigma_f\}. \tag{7}$$

The bounds of θ^a and θ^f are restricted in Assumptions 1 and 2 respectively, which may be conservatively estimated in the following and do not need to be very precise. To guarantee the stealthiness of the replay attack, θ^a must be bounded. Physical sensor faults are always bounded in practice (see e.g., [13], [14]) and hence, θ^f is bounded as well. Moreover, Θ^a and Θ^f are used as the regions of the projection operators in the adaptive estimators, guaranteeing the boundedness of the generated estimates. Therefore, only possibly large bounds of θ^a and θ^f are needed by the defender. Such bounds are also required in many fault diagnosis literature such as [13]–[15]. Regarding the approach to obtain the bounds of θ^a and θ^f, since $|\theta^a| = |x'_n(T_d) - x_a(T_d)|$, σ^a in (5) typically may be obtained based on a priori knowledge of the physical bounds of the system states. Also, σ^f in (7) may be obtained by exploiting a priori knowledge of the sensor bias deviation based on the technical characteristics of the sensors.

Regarding the threat scenarios considered in this paper, we have the following assumption.

Assumption 3. It is assumed that only one type threat occurs during a threat event, namely, 1) the replay attack scenario and 2) the sensor bias fault scenario.

Assumption 3 is made for focusing the presentation to the scope of this letter. Note that replay attacks typically last for a limited time duration, i.e., $t \in [T_a - T, T_a + T]$, and hence the case that the sensor fault(s) appear after the replay attack, i.e., $T_f \ge T_a + T$, is also covered implicitly by Assumption 3. In addition, the developed threat discrimination scheme can also handle the case that the sensor fault(s) appear before the replay attack, i.e., $T_f \le T_a - T$. As for the case that a bias fault occurs during a replay attack, i.e., $T_f \in [T_a - T, T_a + T]$, the probability of this case is very low in practice since a replay attack can only endure for a relatively short time. In such a case, both $\mathcal{W}_a$ in (3) and $\mathcal{W}_f$ in (6) cannot exactly describe the system and the developed discrimination methodology cannot be used. Therefore, new techniques are needed for identifying this threat case.

3) Objective: We suppose that a threat has been detected by the anomaly detector $\mathcal{D}$ in Fig. 1 at time T_d using additive watermarks [5] in control inputs or multiplicative watermarks [9] in sensor measurements, where $T_d \ge T_a$ and $T_d \ge T_f$. However, the type of such a threat cannot be identified by $\mathcal{D}$. The *objective* of this paper is to design an methodology to identify which type of threat has occurred. Note that at the initiation time instant T_d, the applied watermarks for the detection purpose are removed and hence do not affect the threat discrimination schemes. The exclusion-logic-based approach is used to isolate different faults in previous works, such as [14]. However, it cannot handle the stealthy replay attacks since $\tilde{y}_a(t)$ is close to $\tilde{y}_n(t)$. In this paper, watermarks are introduced and integrated with the exclusion-logic-based approach to deal with this problem.

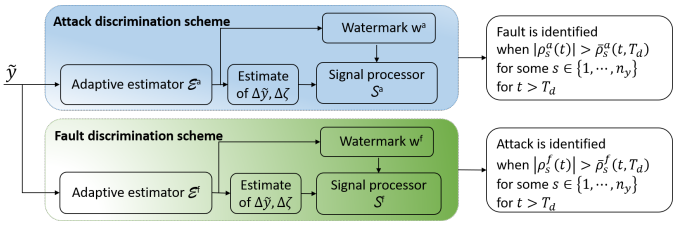

Fig. 2. Schematic diagram of the proposed threat discrimination approach.

III. Threat Discrimination

In this section, we propose an approach to identify replay attacks and sensor faults, and also design and analyze estimators and signal processors with watermarks. We start by presenting an observer used for analytical purpose:

$$\mathcal{O}_n: \ \dot{\zeta}_n(t) = A_\zeta \zeta_n(t) + L_\zeta \tilde{y}_n(t), \tag{8}$$

where $\zeta_n \in \mathbb{R}^{n_p}$ is the state of the observer in the nominal case, $A_\zeta \triangleq A - L_0 C$ with L_0 being designed such that A_ζ is a Hurwitz matrix, and $L_\zeta = L_0 + BK$. In the threat case, $\mathcal{O}_n$ is described by $\mathcal{O}: \ \dot{\zeta} = A_\zeta \zeta + L_\zeta \tilde{y}$ where ζ is the state in the threat case. The changes of ζ and $\tilde{y}$ due to the threat are defined respectively as $\Delta\zeta \triangleq \zeta - \zeta_n$ and $\Delta\tilde{y} \triangleq \tilde{y} - \tilde{y}_n$. From $\mathcal{O}$ and $\mathcal{O}_n$, we obtain

$$\Delta\zeta(t) = e^{A_\zeta (t-T_d)} \Delta\zeta(T_d) + \int_{T_d}^{t} e^{A_\zeta (t-\tau)} L_\zeta \Delta\tilde{y}(\tau) d\tau. \tag{9}$$

The threat discrimination methodology is depicted in Fig. 2, which integrates adaptive estimators and signal processors possessing watermarks. Two estimators corresponding to the replay attack case and the sensor fault case are designed to estimate $\Delta\tilde{y}$ and $\Delta\zeta$. Signal processors with watermarks are developed based on the estimate of $\Delta\zeta$ and produce corresponding adaptive thresholds. The exclusion-logic-based approach is then applied to make the identification decision.

A. Threat Discrimination Estimator Design

Two distinct adaptive estimators, activated at T_d and corresponding to the attack threat case and the fault case respectively, are designed for estimating x_n. To this end, $\mathcal{W}_a$ in (3) and $\mathcal{W}_f$ in (6) are presented in a unified form as follows:

$$\mathcal{W}: \ \begin{cases} \dot{x} &= Ax + BK\tilde{y}, \\ \tilde{y} &= Cx + F(t)\theta, \end{cases} \tag{10}$$

where $F\theta \in \{F^a\theta^a, F^f\theta^f\}$. In addition, in order to present the approach to estimate x_n, the following lemma is needed:

Lemma 1. *Consider an auxiliary system defined as follows:*

$$\dot{\Omega} = A_0\Omega + BKF(t), \ \Omega(T_d) = 0, \tag{11}$$

where $F \in \{F^a, F^f\}$, $\Omega \in \mathbb{R}^{n_p \times n_p}$ when $F = F^a$ or $\Omega \in \mathbb{R}^{n_p \times n_y}$ when $F = F^f$, and $A_0 \triangleq A + BKC$ is a Hurwitz matrix. Consider also a state variable z and its corresponding output $\tilde{y}_{nz}$ defined as

$$z \triangleq x - \Omega\theta, \ \tilde{y}_{nz} \triangleq Cz, \tag{12}$$

where x and $\theta \in \{\theta^a, \theta^f\}$ are given in (10). Then,

$$x_n(t) = z(t) + e_{nz}(t), \ \forall\ t \geq T_d, \tag{13}$$

where x_n is given in (1) and $e_{nz}(t) \triangleq e^{A_0(t-T_d)}\Delta x(T_d)$.

Proof. From (11), (12) and $\dot{z} = \dot{x} - \dot{\Omega}\theta$, we have $\dot{z} = A_0 z$. From (1) and $e_{nz} = x_n - z$, we can derive $\dot{e}_{nz} = A_0 e_{nz}$. It indicates that e_{nz} converges to zero exponentially due to the Hurwitz matrix A_0, and thus, z is an estimate of x_n. It follows from (11) and (12) that $z(T_d) = x(T_d)$ and $e_{nz}(T_d) = \Delta x(T_d)$. Hence, the result (13) follows. □

According to Lemma 1, z is an estimate[1] of x_n since e_{nz} goes to zero exponentially. Hence, the estimation procedure for x_n is achieved through estimating z in (12) in the sequel. We start by proposing an adaptive estimator for $\mathcal{W}$ corresponding to the threat $i \in \{a, f\}$ as follows:

$$\mathcal{E}^i: \begin{cases} \dot{\hat{x}}^i &= A\hat{x}^i + BK\tilde{y} - L(\tilde{y} - \hat{\tilde{y}}^i) + \Omega^i\dot{\hat{\theta}}^i, \\ \dot{\Omega}^i &= A_0\Omega^i + LF^i, \ \Omega^i(T_d) = 0, \\ \hat{\tilde{y}}^i &= C\hat{x}^i + F^i(t)\hat{\theta}^i, \\ \dot{\hat{\theta}}^i &= \mathcal{P}_{\Theta^i}\{\gamma^i(C\Omega^i + F^i(t))^T(\tilde{y} - \hat{\tilde{y}}^i)\}, \end{cases} \tag{14}$$

where $\hat{x}^i$, $\hat{\tilde{y}}^i$ and $\hat{\theta}^i$ are the estimates of x, $\tilde{y}$ and θ in the i-th threat case, respectively, and $\hat{x}^i(T_d) = 0$. The gain $L \triangleq BK$ such that $A + LC = A_0$ and Ω^i has the form of Ω defined in (11). Moreover, the projection operator $\mathcal{P}_{\Theta^i}$ restricts $\hat{\theta}^i$ to the convex region Θ^i (to guarantee the stability of the estimation error system for the adaptive estimator (14) in the presence of modeling uncertainties) [16]. Since Θ^i is a hypersphere of radius σ^i, the mathematical representation of $\mathcal{P}_{\Theta^i}$ is same as the one in [13]. The $\hat{\theta}^i(T_d)$ is chosen such that $\hat{\theta}^i(T_d) \in \Theta^i$, and $\gamma^i > 0$ is the learning rate.

Remark 2. The estimation vector $\hat{\theta}^i$ provides useful information for threat discrimination. For example, considering the estimator $\mathcal{E}^i$, the uniform bound of $\theta - \hat{\theta}^i$ in the i-th threat case is smaller than in the j-th threat case due to the correct matching of $\mathcal{E}^i$ to the occurred i-th threat case, which helps in discriminating between the i-th and the j-th threat cases. However, for both sensor faults and replay attacks, it cannot be guaranteed that $\hat{\theta}^i$ will converge to the true value θ in the absence of a restrictively persistent excitation condition. Note that we do not assume or require persistency of excitation. ∇

Remark 3. Compared with typical adaptive observers, the form of the adaptive observer $\mathcal{E}^i$ in (14) is able to provide estimates of x_n and $\tilde{y}^n$. The estimate of $\tilde{y}^n$, detailed discussed in Theorem 1, can be used to construct control signals, allowing to mitigate the effects of the attacks and faults, which will be dealt with in future work. ∇

We now investigate the stability and the learning capability of $\mathcal{E}^i$. Based on (12), and using $\hat{x}^i$ and $\hat{\theta}^i$, we define the following variables:

$$\hat{z}^i \triangleq \hat{x}^i - \Omega^i\hat{\theta}^i, \ \hat{\tilde{y}}^i_n \triangleq C\hat{z}^i, \tag{15}$$

where $\hat{z}^i$ and $\hat{\tilde{y}}^i_n$ are estimates of z and $\tilde{y}_{nz}$ in (12) in the i-th threat case, respectively. Moreover, according to Lemma 1, both $\hat{z}^i$ and z are estimates of x_n, and $\tilde{y}_{nz}$ and $\hat{\tilde{y}}^i_n$ are estimates of $\tilde{y}_n$. Also, we define the estimation errors:

$$\begin{aligned} &e^i_x \triangleq x - \hat{x}^i, \ e^i_y \triangleq \tilde{y} - \hat{\tilde{y}}^i = Ce^i_x + F\theta - F^i\hat{\theta}^i, \\ &\bar{e}^i_x \triangleq z - \hat{z}^i, \ \bar{e}^i_y \triangleq \tilde{y}_{nz} - \hat{\tilde{y}}^i_n = C\bar{e}^i_x = Ce^i_x + C\Omega^i\hat{\theta}^i - C\Omega\theta, \\ &\tilde{\theta}^i \triangleq \theta^i - \hat{\theta}^i, \ \forall\ i \in \{a, \ f\}. \end{aligned}$$

Since θ is a constant vector, it follows from (12) that $\dot{z} = \dot{x} - \dot{\Omega}\theta$. Also, it follows from (15) that $\dot{\hat{z}}^i = \dot{\hat{x}}^i - \dot{\Omega}^i\hat{\theta}^i - \Omega^i\dot{\hat{\theta}}^i$. Thus, based on (10) and (14), we can obtain

$$\dot{\bar{e}}^i_x = A(x - \hat{x}^i) + L(\tilde{y} - \hat{\tilde{y}}^i) - \dot{\Omega}\theta + \dot{\Omega}^i\hat{\theta}^i,$$

and based on $\dot{\Omega}$ in (11) and $\dot{\Omega}^i$ in (14), we have $\dot{\bar{e}}^i_x = A_0\bar{e}^i_x$. In addition, we define the following mismatch function between the j-

[1]For a signal vector $x(t) \in \mathbb{R}^n$, $\hat{x}(t) \in \mathbb{R}^n$ is considered to be an estimate of $x(t)$ if $\lim_{t\to\infty}(x(t) - \hat{x}(t)) = 0$.

th threat and the i-th estimator:

$$d^{ij} \triangleq C\Omega^j\theta^j + F^j\theta^j - C\Omega^i\hat{\theta}^i - F^i\hat{\theta}^i,\ \forall\ i,j = \{a,\ f\}. \quad (16)$$

It follows that $\bar{e}_y^i$ can be written as

$$\bar{e}_y^i = e_y^i - F\theta + F^i\hat{\theta}^i + C\Omega^i\hat{\theta}^i - C\Omega\theta = e_y^i - d^{ij}.$$

Therefore, the estimation error system can be obtained as

$$\dot{\bar{e}}_x^i = A_0\bar{e}_x^i,\ \bar{e}_y^i = C\bar{e}_x^i = e_y^i - d^{ij}, \quad (17)$$

$$\dot{\hat{\theta}}^i = \mathcal{P}_{\Theta^i}\left\{\gamma(C\Omega^i + F^i(t)))^T\right\}e_y^i. \quad (18)$$

The stability and learning properties of the estimator $\mathcal{E}^i$ are described in the following theorem.

Theorem 1. *Consider the system $\mathcal{W}$ in (10) with the pair (A, C) being observable, the replay attack and the sensor bias fault satisfying Assumptions 1 and 2, respectively. Moreover, suppose that Assumption 3 holds, and that a threat is detected at time T_d ($T_d > T_a$ or $T_d > T_f$). Then, both estimators $\mathcal{E}^i$ in (14), $i \in \{a, f\}$, guarantee that the errors e_x^i, e_y^i and $\tilde{\theta}^i$ are uniformly bounded. Moreover, in the occurring i-th threat, the i-th estimator satisfies:*

$$\lim_{t\to\infty}(x_n(t) - \hat{z}^i(t)) = 0,\ \lim_{t\to\infty}(\tilde{y}_n(t) - \hat{\tilde{y}}_n^i(t)) = 0, \quad (19)$$

where $\hat{z}^i$ and $\hat{\tilde{y}}_n^i$ are defined in (15) and x_n and $\tilde{y}_n$ are given in (1). Also, in the i-th threat case, there exist $\eta_0 > 0$ and a bounded function η^i such that e_y^i satisfies

$$\int_{T_d}^t |e_y^i(\tau)|^2\ d\tau \le \eta_0 + 2\int_{T_d}^t |\eta^i(\tau)|^2\ d\tau,\ \forall\ t \ge T_d. \quad (20)$$

Proof. Since A_0 is a Hurwitz matrix, $\bar{e}_x^i$ in (17) converges to zero asymptotically. Thus, $\bar{e}_x^i \in L_\infty$ and $\bar{e}_y^i(t) \in L_\infty$. Due to the parameter projection, $\hat{\theta}^i, \tilde{\theta}^i \in L_\infty$. In addition, (14) indicates that $\Omega^i \in L_\infty$. From $e_x^i = \bar{e}_x^i + (\Omega\theta - \Omega^i\theta^i)$ and $e_y^i = \bar{e}_y^i - d^{ij}$ and based on Assumptions 1 and 2, we can conclude that $e_x^i = x - \hat{x}^i \in L_\infty$ and $e_y^i = \tilde{y} - \hat{\tilde{y}}^i \in L_\infty$.

Regarding $x_n - \hat{z}^i$ and $\tilde{y}_n - \hat{\tilde{y}}_n^i$, we have $x_n - \hat{z}^i = x_n - z + z - \hat{z}^i$. According to Lemma 1, $\lim_{t\to\infty}(x_n - z) = 0$. Since A_0 is a Hurwitz matrix, it follows from (17) that $\lim_{t\to\infty}(z - \hat{z}^i) = \lim_{t\to\infty}\bar{e}_x^i = 0$. Thus, we have $\lim_{t\to\infty}(x_n - \hat{z}^i) = 0$ and $\lim_{t\to\infty}(\tilde{y}_n - \hat{\tilde{y}}_n^i) = \lim_{t\to\infty} C(x_n - \hat{z}^i) = 0$. Hence, (19) follows.

The third part of the theorem concerns the learning capability of the estimator $\mathcal{E}^i$ in the i-th threat case (the estimator matches the threat case). Let $\bar{e}_x^i \triangleq \xi_1^i + \xi_2^i$ for $t > T_d$. Then, it follows from (17) that $\dot{\xi}_1^i = A_0\xi_1^i,\ \xi_1^i(T_d) = 0$, and $\dot{\xi}_2^i = A_0\xi_2^i,\ \xi_2^i(T_d) = \bar{e}_x^i(T_d)$. Thus, e_y^i can be written as

$$e_y^i = C(\xi_1^i + \xi_2^i) + (C\Omega^i + F^i(t))\tilde{\theta}^i.$$

Considering a Lyapunov function candidate $V = \frac{1}{2\gamma}(\tilde{\theta}^i)^T\tilde{\theta}^i + \int_t^\infty |C\xi_2^i|^2 d\tau$, its time derivative along (18) and the trajectory of $\xi_2^i(t)$ is

$$\begin{aligned}\dot{V} &= -\frac{1}{\gamma}(\tilde{\theta}^i)^T\dot{\hat{\theta}}^i + |C\xi_2^i|^2 \\ &= \frac{1}{\gamma}(\tilde{\theta}^i)^T\mathcal{P}_{\Theta^i}\{\gamma(C\Omega^i + F^i)\}^T e_y^i(t) + |C\xi_2^i|^2.\end{aligned}$$

By following the logic in [16] to deal with the projection operator $\mathcal{P}_{\Theta^i}$, we obtain $\frac{1}{\gamma}(\tilde{\theta}^i)^T\mathcal{P}_{\Theta^i}\{\gamma(C\Omega^i + F^i)\}^T e_y^i \le (\tilde{\theta}^i)^T\{(C\Omega^i + F^i)\}^T e_y^i$. By using e_y^i obtained previously and completing the squares, $\dot{V} \le -|e_y^i|^2/2 + |C\xi_1^i|^2$. Thus, by letting $\eta^i(t) \triangleq |C\xi_1^i(t)|$, we can deduce that

$$V(t) - V(T_d) \le -\int_{T_d}^t |e_y^i|^2/2 d\tau + \int_{T_d}^t |\eta^i|^2\ d\tau,\ \forall\ t \ge T_d,$$

where $\eta^i \triangleq |C\xi_1^i|$. Due to the boundedness of $\bar{e}_x^i$ and $\tilde{\theta}^i$, ξ_1^i and η^i are also bounded. Therefore, by letting $\eta_0 \triangleq 2V(T_d)$, the inequality (20) follows. □

Remark 4. The estimator $\mathcal{E}^i$ guarantees that e_x^i, e_y^i and $\tilde{\theta}^i$ are uniformly bounded in both threat cases (attack case and fault case), which indicates that the unmatched threat case j does not cause divergence of the estimation errors of the estimator $\mathcal{E}^i$ ($i \neq j$) and hence stability is guaranteed. It should be noted that for the estimator $\mathcal{E}^i$, the uniform bounds of e_x^i, e_y^i and $\tilde{\theta}^i$ are smaller in the i-th threat case than in the j-th threat case due to the correct matching of the structure of $\mathcal{E}^i$ to the occurred threat case. ▽

B. Signal Processors with Watermarks

Signal processors and watermarks are proposed in this section. We start by constructing the estimate of $\Delta\tilde{y} \triangleq \tilde{y} - \tilde{y}_n$ by using the estimates obtained from $\mathcal{E}^i$ in (14). Since in the i-th threat case, $\hat{\tilde{y}}_n^i$ in (14) is an estimate of $\tilde{y}_n$ (see (19) in Theorem 1), based on the definition of $\Delta\tilde{y}$ and by using $\hat{\tilde{y}}_n^i$, an estimate of $\Delta\tilde{y}$ is proposed as

$$\Delta\hat{\tilde{y}}^i \triangleq \tilde{y} - \hat{\tilde{y}}_n^i = \Delta\tilde{y} + Ce_{nz} + \bar{e}_y^i, \quad (21)$$

where $\tilde{y}_n = \tilde{y}_{nz} + Ce_{nz}$ is used. It can be easily verified that $\lim_{t\to\infty}(\Delta\hat{\tilde{y}}^i - \Delta\tilde{y}) = 0$ since e_{nz} and $\bar{e}_y^i$ converge to zero as t goes infinity. Next, the following signal processors are used for threat discrimination:

$$\mathcal{S}^i:\ \begin{cases} \dot{x}_w^i = A_\zeta x_w^i + L_\zeta w^i(t), & (22a) \\ \rho^i(t) = C\Delta\hat{\zeta}^i(t) - Cx_w^i, & (22b)\end{cases}$$

where $x_w^i \in \mathbb{R}^{n_p}$ is a vector signal used for compensation purposes, $\rho^i \in \mathbb{R}^{n_y}$ is the output vector, and $\Delta\hat{\zeta}^i$ is constructed based on $\Delta\zeta$ in (9) ($\Delta\zeta(T_d)$ is ignored for simplicity) as follows:

$$\Delta\hat{\zeta}^i(t) = \int_{T_d}^t e^{A_\zeta(t-\tau)}L_\zeta\Delta\hat{\tilde{y}}^i(\tau)d\tau. \quad (23)$$

The compensation system (22a) starts at time T_d, and $x_w^i(T_d) = 0$. The task of x_w^i is to generate distinguishable outputs ρ^i in the fault case and the attack case, respectively. More specifically, w^i is designed such that x_w^i is able to completely compensate for $\Delta\tilde{y}$ in the i-th threat case, but cannot compensate for $\Delta\tilde{y}$ in the j-th threat case, $j \neq i$. In order to achieve the above task, the input signal w^i, referred to as "watermark", is designed as

$$w^i \triangleq (C\Omega^i + F^i)\hat{\theta}^i,\ i \in \{a, f\}. \quad (24)$$

Thus, from $\Delta\tilde{y} = \tilde{y}_j - \tilde{y}_n$, $j \in \{a, f\}$, we obtain

$$w^i = \Delta\tilde{y} + Ce_{nz} - d^{ij},\ i,j \in \{a, f\}, \quad (25)$$

where e_{nz} is given after (11), and d^{ij} is defined in (16) and represents the compensation error for $\Delta\tilde{y}$. By solving the differential equation (22a) and using (23), ρ^i in (22b) is obtained as $\rho^i(t) = Cg(t, \Delta\hat{\tilde{y}} - w^i)$ where

$$g(t, \Delta\hat{\tilde{y}} - w^i) \triangleq \int_{T_d}^t e^{A_\zeta(t-\tau)}L_\zeta(\Delta\hat{\tilde{y}} - w^i)d\tau. \quad (26)$$

Before presenting the main theorem, a fact is given [13]: for a Hurwitz matrix A and a fixed time t_0, there exist two scalars $k > 0$ and $\lambda > 0$ such that

$$|e^{A(t-t_0)}| \le \epsilon(k, \lambda, t, t_0) \triangleq ke^{-\lambda(t-t_0)}. \quad (27)$$

Theorem 2. *Consider the system $\mathcal{W}_n$ in (1) with the pair (A, C) being observable, the replay attack and the sensor bias fault satisfying Assumptions 1 and 2, respectively. Also, consider the output ρ^i of the signal processor $\mathcal{S}^i$ (22) with the watermark $w^i(t)$ in (24) where $i \in \{a, f\}$. Moreover, suppose that Assumption 3 holds.*
(i) *If the i-th threat case has occurred, then the s-th element of $\rho^i(t)$ with $s \in \{1, \cdots, n_y\}$, satisfies $|\rho_s^i(t)| \leq \bar{\rho}_s^i(t, T_d)$ for $t \geq T_d$, where*

$$\bar{\rho}_s^i(t, T_d) \triangleq |C_s| \cdot |L_\zeta| \cdot \int_{T_d}^{t} \epsilon(k_\zeta, \lambda_\zeta, t, \tau) \left(|C| \epsilon_x^i(\tau) + |C\Omega^i + F^i| \delta^i(\tau) \right) d\tau. \quad (28)$$

In (28), C_s indicates the s-th row of C, $\epsilon_x^i(t) \triangleq |x(T_d)| \epsilon(k_0, \lambda_0, t, T_d)$ and $\delta^i(t) \geq |\theta^i - \hat{\theta}^i(t)|$. Moreover, the function ϵ along with the pairs (k_0, λ_0) and (k_ζ, λ_ζ) are specified in (27) with respect to A_0 and A_ζ, respectively.
(ii) *If the i-th threat has occurred, and there exists at least one $s \in \{1, \cdots, n_y\}$ and a time instant $t^i \geq T_d$ such that*

$$|C_s g(t^i, d^{ji})| \geq \bar{\rho}_s^j(t^i, T_d) + |C_s| \cdot |L_\zeta C| \int_{T_d}^{t^i} \epsilon(k_\zeta, \lambda_\zeta, t^i, \tau) \epsilon_x^j(\tau) d\tau, \ \forall\ j \neq i. \quad (29)$$

Then, the i-th threat case can be discriminated.

Proof. Some bounds that will be used in the sequel are first presented. At the initial time instant T_d, we have $\left|\bar{e}_x^i(T_d)\right| = |z(T_d) - \hat{z}(T_d)| = |x(T_d) - \hat{x}(T_d)| = |x(T_d)|$. According to (27), for $\bar{e}_x^i = e^{A_0(t-T_d)} x(T_d)$, we have

$$\left|\bar{e}_x^i(t)\right| \leq |x(T_d)|\, \epsilon(k_0, \lambda_0, t, T_d) = \epsilon_x^i(t). \quad (30)$$

Furthermore, for $\bar{e}_y^i(t) = C\bar{e}_x^i(t)$, by using (17) we have

$$\left|\bar{e}_y^i(t)\right| = \left|C\bar{e}_x^i(t)\right| \leq |C| \epsilon_x^i(t). \quad (31)$$

(i) In the i-th threat case, $i \in \{a, f\}$, it follows from (21) and (25) that

$$\Delta\hat{\tilde{y}} - w^i = \bar{e}_y^i + d^{ii},$$

where $d^{ii} = (C\Omega_i + F^i)\tilde{\theta}^i$. Then, $\rho^i = Cg(t, \bar{e}_y^i + d^{ii})$. By using (31) and $|d^{ii}| \leq |C\Omega^i + F^i| \delta^i$, the s-th element of ρ^i satisfies

$$|\rho_s^i| \leq |C_s| \cdot |L_\zeta| \cdot \int_{T_d}^{t} \epsilon(k_\zeta, \lambda_\zeta, t, \tau) \left(|C| \epsilon_x^i + \left|C\Omega^i + F^i\right| \delta^i \right) d\tau.$$

Thus, $\bar{\rho}_s^i(t, T_d)$ in (28) is obtained.
(ii) Considering the i-th threat case and the j-th threat discrimination scheme, it follows from (21) and (25) that

$$\Delta\hat{\tilde{y}} - w^j = \bar{e}_y^j + d^{ji}.$$

Then, $\rho^j = Cg(t, \bar{e}_y^j + d^{ji})$. In order to identify the i-th threat, the j-th threat must be excluded and hence, the inequality $|\rho_s^j(t)| > \bar{\rho}_s^j(t, T_d)$ must hold at some time $t^i > T_d$ and for some $s \in \{1, \cdots, n_y\}$. By using the inverse triangle inequality, we have

$$|\rho_s^j(t^i)| \geq |C_s g(t^i, d^{ji})| - |C_s g(t^i, \bar{e}_y^j)|,$$

where from (31) and $\rho^i(t) = Cg(t, \Delta\hat{\tilde{y}} - w^i)$, we have

$$|C_s g(t^i, \bar{e}_y^j)| \leq |C_s| \cdot |L_\zeta C| \cdot \int_{T_d}^{t^i} \epsilon(k_\zeta, \lambda_\zeta, t^i, \tau) \epsilon_x^j d\tau.$$

Thus, the sufficient condition (29) is obtained. □

According to Theorem 2, the *threat discrimination logic* is given based on the exclusion-logic as follows: if there is a time t^f such that $|\rho^a(t^f)| > \bar{\rho}^a(t^f, T_d)$, then the detected threat is identified as sensor bias fault(s), and if there is a time t^a such that $|\rho^f(t^a)| > \bar{\rho}^f(t^a, T_d)$, then the detected threat is identified as a replay attack. In the presence of a false alarm at time instant T_d, the residuals $\rho_s^a(t)$ and $\rho_s^f(t)$ remain below their corresponding thresholds $\bar{\rho}_s^a(t, T_d)$ and $\bar{\rho}_s^f(t, T_d)$ for all $s \in \{1, \cdots, n_y\}$. Hence, no decision regarding the threat type can be made based on the threat discrimination logic. Algorithm 1 is given in the sequel to summarize the designed details of the threat discrimination methodology.

Algorithm 1 Threat Discrimination Algorithm

procedure OBSERVER $\mathcal{O}_n(L_0)$ ▷ in (8)
  $A_\zeta = A - L_0 C$ is a Hurwitz matrix;
end procedure
procedure ADAPTIVE ESTIMATOR $\mathcal{E}^i(L)$ ▷ in (14)
  $L = BK$;
end procedure
procedure SIGNAL PROCESSOR $\mathcal{S}^i\ (w^i)$ ▷ in (22)
  Watermark $w^i(t) = (C\Omega^i(t) + F^i(t))\hat{\theta}^i(t)$; ▷ in (24)
  Residual $\rho^i(t) = Cg(t, \Delta\hat{\tilde{y}} - w^i)$; ▷ in (26)
  Threshold $\bar{\rho}_s^i(t, T_d)$; ▷ in (28)
end procedure
Decision logic:
if $|\rho^a(t^f)| > \bar{\rho}^a(t^f, T_d)$ **then** sensor fault is identified
else if $|\rho^f(t^a)| > \bar{\rho}^f(t^a, T_d)$ **then** replay attack is identified
end if

Remark 5. Optimizing the gain matrix L_0 of the observer $\mathcal{O}_n$ in (8) provides a potential way to improve the threat discrimination ability. By optimizing the matrix L_0, the threshold $\bar{\rho}_s^i$ in (28) in Theorem 2 can be minimized whereas the residual $g^i(t^i, d^{ji})$ in (29) can be maximized. Hence, in this way, the discrimination ability of the developed discrimination methodology can be improved. ∇

Remark 6. The condition (29) in Theorem 2 indicates that the mismatch term d^{ij} in (16) should be sufficiently large to allow the discrimination between the replay attack and the sensor fault. Based on the definition of d^{ij} in (16), F^f and F^a need to be "sufficiently different" to guarantee the requirement for d^{ij}. The case of constant sensor bias faults, leading to $F^f = I_{n_y \times n_y}$, is an example of such a "sufficient difference". As indicated in *Remark 1*, the scheme can also be applied to time-varying faults $F^f(t)\theta^f$ given that some additional conditions are satisfied. ∇

Remark 7. This paper utilizes the difference between the distribution matrices F^a of the replay attack and F^f of the sensor bias faults to discriminate between the replay attack scenario and the sensor bias fault scenario. The specially designed adaptive estimators $\mathcal{E}^i$ and the signal processors $\mathcal{S}^i$ with the watermarks w^i ($i \in \{a, f\}$) are integrated to formulate the threat discrimination approach. Compared with the fault isolation literature [13], [14], [17] and the replay attack detection literature [5], [7], [9], this paper derives for the first time the linear parameterization form of replay attacks by introducing the virtual attack signal. In addition, by using the adaptive estimators and the watermarks, this paper is able to utilize the difference between the distribution matrices F^a and F^f to discriminate the threats. ∇

IV. Simulation Results

In this section, an illustrative simulation example is presented. The matrices of the plant (1) are given as follows:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, B = \begin{bmatrix} -1.67 & 0 & 0 \\ 0 & -1.93 & 0 \\ 0 & 0 & 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Moreover, the gain matrices K and L_0 are:

$$K = \begin{bmatrix} -0.18 & 0.60 \\ -0.02 & -0.67 \\ 0.03 & -0.42 \end{bmatrix}, \; L_0 = \begin{bmatrix} -0.3 & -1.0 \\ -0.04 & 1.29 \\ 0.03 & -0.42 \end{bmatrix}.$$

Regarding Assumptions 1 and 2, σ_a in (5) and σ_f in (7) are given by $\sigma_a = 10$ and $\sigma_f = 20$. For the simulation purpose, the threat information is given as follows: a) the attacker starts recording the data at $T_a - T = 0.5$s and then the attacker starts replaying the data at $T_a = 50$s. Thus, both the recording and the replaying procedures last for $T = 49.5$s; b) the sensor bias faults occur at $T_f = 49$s, and the constant faults are given by $\theta^f = [12, 13]^T$. According to Assumption 3, these two threat scenarios are considered separately.

We consider that a threat is detected at $T_d = 50.2$s. The learning rate is set to $\gamma = 0.5$ and the initial conditions are set as $\hat{\theta}^a(T_d) = [0, 0, 0]^T$ and $\hat{\theta}^f(T_d) = [0, 0]^T$. Moreover, the parameters for obtaining the threshold (28) are given as follows: $\lambda = \lambda_\zeta = 0.3$, $k_0 = k_\zeta = 3.66$, $|x(T_d)| \leq 10$ and $\delta^a(t) = \delta^f(t) = 10$. In each threat case, the identification results are respectively shown in Figs. 3 and 4.

1) Attack Case: In this simulation, only the replay attack is performed. It can be seen from Fig. 3 that $\hat{\rho}_1^a(t)$ and $\hat{\rho}_2^a(t)$ generated by the attack discrimination scheme are lower than their thresholds $\hat{\bar{\rho}}_1^a(t, T_d)$ and $\hat{\bar{\rho}}_2^a(t, T_d)$ respectively. The residual $\hat{\rho}_1^f(t)$ generated by the fault discrimination scheme exceeds its threshold $\hat{\bar{\rho}}_1^f(t, T_d)$ at $t = t^a$, and hence, the fault type is excluded, indicating that the threat is an attack.

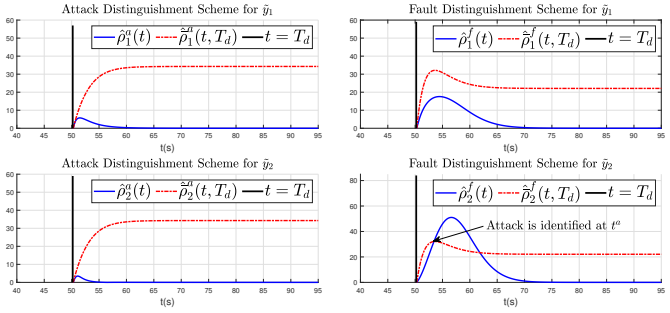

Fig. 3. Time responses of residuals and thresholds in the attack case.

2) Fault Case: In this simulation, the sensor bias fault is performed. Fig. 4 shows that $\hat{\rho}_1^f(t)$ and $\hat{\rho}_2^f(t)$ generated by the fault discrimination scheme are lower than their thresholds $\hat{\bar{\rho}}_1^f(t, T_d)$ and $\hat{\bar{\rho}}_2^f(t, T_d)$ respectively, whereas, the residual $\hat{\rho}_1^a(t)$ generated by the attack discrimination scheme exceeds its threshold $\hat{\bar{\rho}}_1^a(t, T_d)$ at $t = t^f$ and hence, the attack case is excluded, indicating that the threat is a fault.

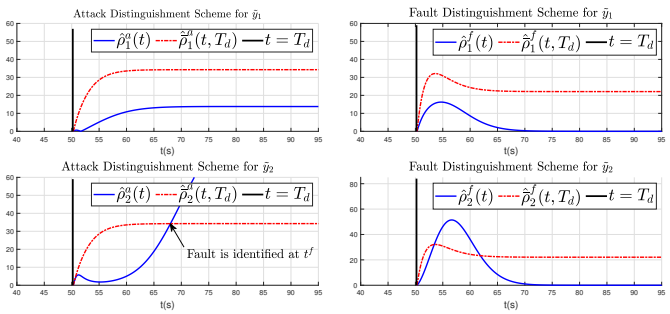

Fig. 4. Time responses of residuals and thresholds in the fault case.

V. Conclusion

A threat discrimination methodology for identifying the occurring threat type between sensor replay attacks and sensor bias faults has been proposed in this letter. Adaptive estimators and signal processors with watermarks have been integrated to formulate the threat discrimination framework. Threat discrimination conditions are rigorously investigated to characterize quantitatively the class of attacks and faults that can be identified by the proposed scheme.

Future work will be devoted to deal with the threat discrimination problems under disturbances. Also, we will focus on developing a unified threat discrimination framework, allowing to discriminate between general cyber attacks and physical faults.

References

[1] A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

[2] F. Pasqualetti, F. Dörfler, and F. Bullo, "Control-theoretic methods for cyber-physical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 110–127, 2015.

[3] S. Dibaji, M. Pirani, D. Flamholz, A. Annaswamy, K. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

[4] R. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 82–92, 2015.

[5] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2009, pp. 911–918.

[6] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[7] R. Ferrari and A. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-Papers OnLine*, vol. 50, no. 1, pp. 7363–7368, 2017.

[8] S. Weerakkody, O. Ozel, Y. Mo, and B. Sinopoli, "Resilient control in cyber-physical systems: countering uncertainty, constraints, and adversarial behavior," *Foundations and Trends® in Systems and Control*, vol. 7, no. 1-2, pp. 339–499, 2019.

[9] R. Ferrari and A. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2558 – 2573, 2021.

[10] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[11] T. Keijzer and R. Ferrari, "A sliding mode observer approach for attack detection and estimation in autonomous vehicle platoons using event triggered communication," in *58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 5742–5747.

[12] A. Gallo, M. Turan, F. Boem, T. Parisini, and G. Ferrari, "A distributed cyber-attack detection scheme with application to DC microgrids," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3800–3815, 2020.

[13] X. Zhang, T. Parisini, and M. Polycarpou, "Sensor bias fault isolation in a class of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 370–376, 2005.

[14] C. Keliris, M. Polycarpou, and T. Parisini, "An integrated learning and filtering approach for fault diagnosis of a class of nonlinear dynamical systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 4, pp. 988–1004, 2017.

[15] C. Edwards, S. K. Spurgeon, and R. J. Patton, "Sliding mode observers for fault detection and isolation," *Automatica*, vol. 36, no. 4, pp. 541–553, 2000.

[16] J. Farrell and M. Polycarpou, *Adaptive approximation based control: unifying neural, fuzzy and traditional adaptive approximation approaches*. John Wiley & Sons, 2006.

[17] K. Zhang, B. Jiang, X. Yan, and Z. Mao, "Incipient voltage sensor fault isolation for rectifier in railway electrical traction systems," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6763–6774, 2017.