Resiliency of Perception-Based Controllers Against Attacks

Amir Khazraei AMIR.KHAZRAEI@DUKE.EDU

Henry Pfister HENRY.PFISTER@DUKE.EDU

Miroslav Pajic MIROSLAV.PAJIC@DUKE.EDU

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

This work focuses on resiliency of learning-enabled perception-based controllers for nonlinear dynamical systems. We consider systems equipped with an *end-to-end* controller, mapping the perception (e.g., camera images) and sensor measurements to control inputs, as well as a statistical or learning-based anomaly detector (AD). We define a general notion of attack stealthiness and find conditions for which there exists a sequence of stealthy attacks on perception and sensor measurements that forces the system into unsafe operation without being detected, for *any* employed AD. Specifically, we show that systems with unstable physical plants and exponentially stable closed-loop dynamics are vulnerable to such stealthy attacks. Finally, we use our results on a case-study.

Keywords: resiliency of learning-enabled controlers, perception-based control, anomaly detection.

1. Introduction

Due to the recent advances in deep learning and perception, the next generation of control systems is incorporating perception modules to extract information from the environment for control and decision making. This includes end-to-end control systems that directly incorporate camera images, LiDAR 3D point clouds, and other sensor information to compute control inputs at runtime (e.g., Pan et al. (2017); Rausch et al. (2017); Jaritz et al. (2018); Polvara et al. (2018); Codevilla et al. (2018)), as well as controllers that first extract state information from the images followed by classic feedback controllers (e.g., Dean and Recht (2020); Dean et al. (2020b,a)). Yet, despite the tremendous promise, resiliency of perception-based controllers to well-documented adversarial threats has not be well addressed, limiting the use of these learning-enabled controllers in real-world applications.

In particular, the main focus of adversarial machine learning has been on vulnerability of deep neural networks (DNNs) to small input perturbation, effectively focusing on robustness analysis of DNNs; e.g., targeting DNNs classification or control performance when bounded noise is added to the images in camera-based control systems. On the other hand, an attacker capable of compromising system perception/sensing would not limit their actions to bounded measurement perturbation. Moreover, little consideration has been given on potential impact of stealthy (i.e., undetectable) attacks, which are especially dangerous in the control context.

Consequently, in this work, we focus on resiliency analysis of perception-based control systems under attack. Specifically, we consider the impact stealthy false-data injection attacks can have on these learning-enabled control systems. Our notion of attack stealthiness is closely related to the one by Bai et al. (2017) for non-perception systems, where an attack is considered stealthy if and only if it is undetected by an optimal detector. However, unlike Bai et al. (2017), we do not restrict our analysis (and stealthiness requirement) only to steady-state behaviors. For different threat models

(mainly the level of information the attacker has about the system), we derive conditions that there exists a stealthy and effective attack sequence forcing the system far from the operating point. In particular, we assume the attacker knows the open-loop plant dynamics, and differentiate the cases where they have (or do not have) access to an estimation of the plant's state during the attack. We show that under these conditions, the probability of attack remaining stealthy can be chosen arbitrarily close to one, if the attacker's state estimation error can be arbitrarily close to zero. When the attacker does not have access to an estimate of the state, the stealthiness level of the attack depends on the system's performance in attack-free operation. Specifically, we show that unlike for linear time-invariant (LTI) systems where stealthy and effective attacks are independent of the control design, for nonlinear systems the level of stealthiness is closely related to the level of closed-loop system stability – i.e., if the closed-loop systems is more stable, the attack can have stronger stealthiness guarantees. On the other hand, the attack impact (i.e., control degradation) fully depends on the level of open-loop system instability (e.g., the size of unstable eigenvalues for LTI systems).

Related Work. The initial work by Szegedy et al. (2013) on adversarial example generation showed that DNNs are vulnerable to small input perturbations. Afterward, the majority of works have applied this idea to adversarial attacks on physical world such as malicious stickers on traffic signs to fool the detectors and/or classifiers (e.g., Eykholt et al. (2018); Song et al. (2018); Papernot et al. (2017); Sun et al. (2020)). However, all these methods only consider classification tasks in a static manner; i.e., without consideration of the longitudinal (i.e., over time) system behaviours.

Recent works by Boloor et al. (2020, 2019); Sato et al. (2020); Jia et al. (2020); Yoon et al. (2021); Hallyburton et al. (2022) have studied the vulnerability of perception-based autonomous vehicles in a longitudinal way. For instance, Boloor et al. (2019, 2020) consider autonomous vehicles with end-to-end DNN controllers that directly map perceptual inputs into the vehicle steering angle, and target the systems by painting black lines on the road. However, all these works consider specific applications and address attack impact in an ad-hoc manner, limiting the use of their results in other systems/domains. Further, they lack any consideration of attack stealthiness, as injecting e.g., adversarial patches that only maximize the disruptive impact on the control can be detected by most ADs. For instance, Cai and Koutsoukos (2020) introduce an AD that easily detects the adversarial attacks from Boloor et al. (2020). On the other hand, in this work, we focus on nonlinear system dynamics, define general notions of attack stealthiness, and introduce sufficient conditions for a perception-based control system to *not* be resilient against perception and sensing attacks.

Finally, for *non-perception* control systems, stealthy attacks have been well-defined in e.g., Mo and Sinopoli (2009, 2010); Teixeira et al. (2012); Smith (2015); Pajic et al. (2017); Bai et al. (2017); Jovanov and Pajic (2019); Sui et al. (2020); Khazraei and Pajic (2020, 2021); Khazraei et al. (2022). However, all these work also only focus on LTI systems and linear controllers, as well as on specific AD design (e.g., χ^2 detector). Recently Khazraei et al. (2022) have introduced a learning-based attack design for systems with nonlinear dynamics; yet, their work only considers stealthiness with respect to the χ^2-based AD, and does not consider perception-based controllers.

Notation. $\mathbb{P}$ denotes the probability for a random variable. For a square matrix A, $\lambda_{max}(A)$ is the maximum eigenvalue. For a vector $x \in \mathbb{R}^n$, $||x||_p$ denotes the p-norm of x; when p is not specified, the 2-norm is implied. For a vector sequence, $x_0 : x_t$ denotes the set $\{x_0, x_1, ..., x_t\}$. A function $f : \mathbb{R}^n \to \mathbb{R}^p$ is Lipschitz with constant L if for any $x, y \in \mathbb{R}^n$ it holds that $||f(x) - f(y)|| \leq L||x - y||$. For a set X, ∂X and X^o define the boundary and the interior of the set, respectively. B_r denotes a closed ball with radius r; i.e., $B_r = \{x \in \mathbb{R}^n \mid \|x\| \leq r\}$, whereas $\mathbf{1}_A$ is the indicator

function on a set A. For a function f, we denote $f' = \frac{\partial f}{\partial x}$ as the partial derivative of f with respect to x and $\nabla f_i(x)$ is the gradient of the function f_i (i-th element of the function f). Finally, if $\mathbf{P}$ and $\mathbf{Q}$ are probability distributions relative to the same Lebesgue measure with densities $\mathbf{p}$ and $\mathbf{q}$, respectively, then the total variation between them is defined as $\text{TV}(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \int |\mathbf{p}(x) - \mathbf{q}(x)| dx$. The Kullback–Leibler (KL) divergence between $\mathbf{P}$ and $\mathbf{Q}$ is $KL(\mathbf{P}, \mathbf{Q}) = \int \mathbf{p}(x) \log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} dx$.

2. Problem Description: System and Attack Models

In this section, we introduce the system and attack model. Specifically, we consider the setup from Fig. 1 where each of the components is modeled as follows.

2.1. Plant and Perception Model

We assume the plant can be modeled with a nonlinear dynamics in the standard state-space form

$$x_{t+1} = f(x_t) + Bu_t + w_t, \quad y_t^s = C_s x_t + v_t^s, \quad z_t = G(x_t). \tag{1}$$

Here, $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, $w_t \in \mathbb{R}^n$, $z_t \in \mathbb{R}^l$, $y_t^s \in \mathbb{R}^s$ and $v_t \in \mathbb{R}^s$ denote the state, input, system disturbance, observations from perception-based sensors, (non-perception) sensor measurements, and sensor noise, at time t, respectively. We assume f is globally Lipschitz with constant L, and without loss of generality that $f(0) = 0$[1]. The perception-based sensing is modeled by an unknown generative model G, which is nonlinear and potentially high-dimensional. Finally, the process and measurement noise vectors w and v^s are independent and identically distributed (iid) Gaussian processes with $w \sim \mathcal{N}(0, \Sigma_w)$ and $v^s \sim \mathcal{N}(0, \Sigma_{v^s})$.

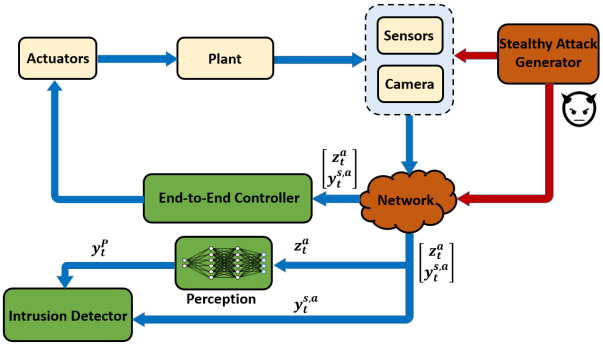

Figure 1: The considered system architecture.

For example, consider a camera-based lane keeping. Here, the observations z_t are the captured images; the map G generates the images based on the vehicle's position.

2.2. Control Unit

The control unit, shown in Fig. 1, consists of perception, controller and anomaly detector units.

Perception. We assume that there exists a perception map P that imperfectly estimates the partial state information; i.e.,

$$y_t^P = P(z_t) = C_P x_t + v^P(x_t), \tag{2}$$

where P is a deep neural network (DNN) trained using any supervised learning method on a data set $\mathcal{X} = \{(z_i, x_i)\}_{i=1}^N$ collected densely around the operating point x_o of the system, as in Dijk and Croon (2019); Lambert et al. (2018). In addition, $v^P \in \mathbb{R}^p$ is the perception map error that may depend on the state of the system – e.g., smaller around points from the training data set. To capture perception guarantees, we employ the model for robust perception-based control by Dean and Recht (2020). Specifically, if the model is trained effectively, the perception error v^P around the operating point x_o should be bounded, i.e., the following assumption from Dean et al. (2020a) holds.

1. Otherwise, we can define a new coordinate system (i.e., transform coordinates) to shift the equilibrium point to zero.

Assumption 1 *There exists a safe set $\mathcal{S}$ around the operating point such that for all $x \in \mathcal{S}$, it holds that $\|P(z) - C_P x\| \leq \gamma_e$, where $z = G(x)$ – i.e., for all $x \in \mathcal{S}$, $\|v^P(x)\| \leq \gamma_e$. Without loss of generality, in this work we consider the origin as the operating point – i.e., $x_o = 0$.*

Controller. The system (1) is controlled by a (general) nonlinear controller $u_t = \pi(z_t, y_t^s) = \Pi(x_t, v_t^s)$. Hence, for $h(x_t, v_t^s) = f(x_t) + B\pi(z_t, y_t^s) = f(x_t) + B\Pi(x_t, v_t^s)$, the evolution of the closed-loop system can be captured as

$$x_{t+1} = h(x_t, v_t^s) + w_t. \tag{3}$$

In the general form, the controller can employ ***any*** end-end control policy that uses the perception and sensor measurements. When the system is noiseless, the the state dynamics can be captured as

$$x_{t+1} = h(x_t, 0). \tag{4}$$

We now introduce definitions that describe the property of the controlled system.

Definition 1 *The origin of the system (4) is exponentially stable on a set $\mathcal{D} \subseteq \mathbb{R}^n$ if for any $x_0 \in \mathcal{D}$, there exist $0 < \alpha < 1$ and $M > 0$, such that $\|x_t\| \leq M\alpha^t \|x_0\|$, for all $t \geq 0$.*

Lemma 2 *Kushner (2014) For the system of (4), if there exists a function $V : \mathbb{R}^n \to \mathbb{R}$ such that for any $x_t \in \mathcal{D} \subseteq \mathbb{R}^n$ it holds*

$$c_1\|x_t\|^2 \leq V(x_t) \leq c_2\|x_t\|^2, \quad V(x_{t+1}) - V(x_t) \leq -c_3\|x_t\|^2, \quad \|\frac{\partial V(x)}{\partial x}\| \leq c_4\|x\|, \tag{5}$$

for some positive c_1, c_2, c_3 and c_4, then the origin is exponentially stable.

Assumption 2 *We assume that for the closed-loop system (4) is exponentially stable on a set $\mathcal{D} = B_d$. Using the converse Lyapunov theorem (see Khalil (2002)), there exists a Lyapunov function that satisfies the inequalities in (5) with constants c_1, c_2, c_3 and c_4 on a set $\mathcal{D} = B_d$.*

Remark 3 *The assumptions made for closed-loop system are critical for system guarantees without the attack; i.e., if the system does not satisfy the stability property in attack-free condition then a possible strategy for the attacker would be to wait until the system fails by itself. We refer the reader to the recent work e.g., by Dean and Recht (2020); Dean et al. (2020a) on design of such controllers.*

Definition 4 *The class of functions $\mathcal{U}_\rho$ contains all functions f such that the dynamics $x_{t+1} = f(x_t) + d_t$, where d_t satisfies $\|d_t\| \leq \rho$, becomes arbitrarily large for some nonzero initial state x_0. Also, for a function f from $\mathcal{U}_\rho$ and initial condition x_0, we define $T_f(\alpha, x_0) = \min\{t \mid \|x_t\| \geq \alpha\}$.*

In other words, $T_f(\alpha, x_0)$[2] is the minimum time needed for an unstable dynamics f, subject to the initial condition x_0, to leave a bounded ball with radius α subject to the initial condition x_0.

Anomaly Detector. The system is equipped with an anomaly detector (AD) designed to detect the presence of any anomalous behaviours. We use $y_t = \begin{bmatrix} y_t^P \\ y_t^s \end{bmatrix}$ and $y_t^a = \begin{bmatrix} y_t^{P,a} \\ y_t^{s,a} \end{bmatrix}$ to capture sensor and perception-based (from (2)) values without and under attack, respectively (the full attack model is introduced below). Then, the standard binary hypothesis testing problem is considered:

H_0: normal condition (the AD receives $y_0 : y_t$);

2. To simplify our notation, and since we consider specific f from the plant dynamics (1), we drop the the subscript f.

H_1: abnormal behaviour (the AD receives $y_0^a : y_t^a$).

Effectively, the AD uses both the extracted state information from the perception map (i.e., (2)) and sensor measurements. Given a random sequence $Y = (y_0 : y_t)$, it either comes from the distribution $\mathbf{P}$ (null hypothesis H_0) or from a distribution $\mathbf{Q}$ (the alternative hypothesis H_1). For a given AD specified by function $D : Y \to \{0, 1\}$, two types of error may occur. Error type (I), also referred as *false alarm*, occurs if $D(Y) = 1$ when $Y \sim \mathbf{P}$, whereas type (II) error (*miss detection*) occurs if $D(Y) = 0$ when $Y \sim \mathbf{Q}$. Hence, the sum of conditional error probabilities of AD D for a given random sequence Y is

$$p_t^e = \mathbb{P}(D(Y) = 0|Y \sim \mathbf{Q}) + \mathbb{P}(D(Y) = 1|Y \sim \mathbf{P}). \tag{6}$$

We define the attack to be stealthy if there exists no detector that can do better than ignoring the received measurements and making a random guess to decide between the two hypotheses. Since in random guess the output of the detector is independent of whether $Y \sim \mathbf{P}$ or $Y \sim \mathbf{Q}$, it holds that $p_t^e = \mathbb{P}(D(Y) = 0) + \mathbb{P}(D(Y) = 1) = 1$. This is also equivalent to the case when there exists no detector that satisfies $p_t^e < 1$ or $\mathbb{P}(D(Y) = 1|Y \sim \mathbf{Q}) \geq \mathbb{P}(D(Y) = 1|Y \sim \mathbf{P})$ which means the probability of true detection be greater than false alarm.

2.3. Attack Model

We assume that the attacker has the ability to compromise camera images as well as (potentially) the sensor measurements y_t^s (see Fig. 1); e.g., this can be achieved by directly compromising the sensors or using a network-based attack. Moreover, the attack starts at $t = 0$, and we use the superscript a to differentiate all signals of the attacked system, for all $t \geq 0$; the attack sequence is $\{z_t^a, y_t^{s,a}\}_{t\geq 0}$, where e.g., the value of the observation delivered to the perception unit at time t is denoted by z_t^a. Therefore, in the presence of an attack, the system dynamics can be captured as

$$\begin{aligned} x_{t+1}^a &= f(x_t^a) + Bu_t^a + w_t^a, \\ u_t^a &= \pi(z_t^a, y_t^{s,a}). \end{aligned} \tag{7}$$

In this work, we assume the attacker has full knowledge of the system, its dynamics and employed architecture. Further, the attacker has the required computation power to calculate suitable attack signals to inject, planning ahead as needed. We formally define the attack stealthiness as follows.

Remark 5 *In our notation, $x_0^a : x_t^a$ denotes a state trajectory of the system under attack (for an attack starting at $t = 0$), while $x_0 : x_t$ denotes the state trajectory of the attack-free system; we refer to such state trajectory as the* attack-free *trajectory. Thus, when comparing the attack-free trajectory and the system trajectory under attack (i.e., from (7)), we assume $w_t^a = w_t$ and $v_t^{s,a} = v_t^s$.*

Definition 6 *Consider the system from (1). An attack sequence is **strictly stealthy** if there exists no detector such that the sum of conditional error probabilities p_t^e satisfies $p_t^e < 1$, for any $t \geq 0$. An attack is ϵ-**stealthy** if for a given $\epsilon > 0$, there exists no detector such that $p_t^e < 1 - \epsilon$, for any $t \geq 0$.*

Theorem 7 *An attack sequence is strictly stealthy if and only if $KL\big(\mathbf{Q}(y_0^a : y_t^a)||\mathbf{P}(y_0 : y_t)\big) = 0$ for all $t \geq 0$, where KL represents the Kullback–Leibler divergence operator. An attack sequence is ϵ-stealthy if the corresponding observation sequence $y_0 : y_t$ satisfies*

$$KL\big(\mathbf{Q}(y_0^a : y_t^a)||\mathbf{P}(y_0 : y_t)\big) \leq \log(\frac{1}{1-\epsilon^2}). \tag{8}$$

Proof First, we prove the strictly stealthy case. Using Neyman-Pearson Lemma for any existing detector D, it follows that

$$p_t^e \geq \int \min\{\mathbf{P}(y), \mathbf{Q}(y)\}dy, \tag{9}$$

where the equality holds for the Likelihood Ratio function as $D^* = \mathbf{1}_{\mathbf{P}\geq\mathbf{Q}}$ (Lemma 3 in Krishnamurthy (2017)). Since $1 - \int \min\{\mathbf{P}(y), \mathbf{Q}(y)\}dy = \frac{1}{2}\int |\mathbf{P}(x) - \mathbf{Q}(x)|dx$, from Polyanskiy and Wu (2014), from the definition of total variation distance between $\mathbf{P}$ and $\mathbf{Q}$, it follows that

$$p_t^e \geq 1 - \text{TV}(\mathbf{P}, \mathbf{Q}). \tag{10}$$

Now, it holds that $\text{TV}(\mathbf{P}, \mathbf{Q}) \leq \sqrt{1 - e^{-KL(\mathbf{Q}||\mathbf{P})}}$ (Eq. (14.11) in Lattimore and Szepesvári (2020)). Thus, if we have $KL\big(\mathbf{P}(y_0^a : y_t^a)||\mathbf{Q}(y_0 : y_t)\big) = 0$, then $p_t^e \geq 1$ for any detector D. On the other hand, if for all detectors we have $p_t^e \geq 1$, then the equality holds for $\text{TV}(\mathbf{P}, \mathbf{Q}) = 0$, which is equivalent to $\mathbf{P} = \mathbf{Q}$ and therefore, $KL\big(\mathbf{P}(y_0^a : y_t^a)||\mathbf{Q}(y_0 : y_t)\big) = 0$.

For the ϵ-stealthy case, we combine (10) with the inequality $\text{TV}(\mathbf{P}, \mathbf{Q}) \leq \sqrt{1 - e^{-KL(\mathbf{Q}||\mathbf{P})}}$ and the ϵ-stealthy condition (8), to show $p_t^e \geq 1 - \text{TV}(\mathbf{P}, \mathbf{Q}) \geq 1 - \sqrt{1 - e^{-KL(\mathbf{Q}||\mathbf{P})}} \geq 1 - \epsilon$. ■

Attack Goal is to *maximize* the degradation of control performance. Specifically, as we consider the origin as the operating point, the attack objective is to maximize the (norm of) states x_t. Moreover, the attacker wants *to remain stealthy – i.e., undetected by the AD*, as formalized below.

Definition 8 *Attack sequence, denoted as* $\{z_0^a, y_0^{s,a}\}, \{z_1^a, y_1^{s,a}\}, \ldots$ *is an* (ϵ, α)*-successful attack if there exists* $t' \geq 0$ *such that* $\|x_{t'}\| \geq \alpha$ *and the attack is* ϵ*-stealthy for all* $t \geq 0$*. When such a sequence exists for a system, the system is called* (ϵ, α)*-attackable. Finally, when the system is* (ϵ, α)*-attackable for arbitrarily large* α*, the system is referred to as perfectly attackable.*

Our goal is the derive methods to capture the impact of stealthy attacks; specifically, in the next section we derive conditions for existence of a *stealthy* yet *effective* attack sequence $\{z_0^a, y_0^{s,a}\}$, $\{z_1^a, y_1^{s,a}\}, \ldots$ resulting in $\|x_t\| \geq \alpha$ for some $t \geq 0$ – i.e., we find conditions for a system to be (ϵ, α)-attackable. For an attack to be stealthy, we focus on the ϵ-stealthy notion; i.e., that the best AD could only improve the probability detection by ϵ compared to a random-guess baseline detector.

3. Conditions for (ϵ, α)-Attackable Systems

To provide sufficient conditions for a system to be (ϵ, α)-attackable, in this section, we introduce two methodologies for design of attack sequences on perception and (classical) sensing data. The difference in these strategies is the level of knowledge the attacker has about the system; we show that the stronger attack impact can be achieved with the attacker having full knowledge of the system. Specifically, we start with the attack strategy where the attacker has access to the current plant state; in such case, we show that the stealthiness condition is less restrictive, simplifying design of ϵ-stealthy attacks. For the second attack strategy, we show that the attacker can launch the attack sequence with only knowing the function f (i.e., plant model); however, achieving ϵ-stealthy attack in this case is harder as more restrictive conditions are imposed on the attacker.

3.1. Attack Strategy *I*: Using Estimate of the Plant State

Consider the attack sequence where z_t^a and $y_t^{s,a}$ injected at time t, for all $t \geq 0$, satisfy

$$z_t^a = G(x_t^a - s_t), \;\; y_t^{s,a} = C_s(x_t^a - s_t) + v_t^{s,a}, \tag{11}$$

with $s_{t+1} = f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t)$, and for a nonzero s_0; here $\hat{x}_t^a$ is a state estimation (in the presence of attacks), and thus $\zeta_t = \hat{x}_t^a - x_t^a$ is the corresponding state estimation error. Note that the attacker can obtain $\hat{x}_t^a$ by e.g., running a local estimator. We assume that the estimation error is bounded by b_ζ – i.e., $\|\zeta_t\| \leq b_\zeta$, for all $t \geq 0$. On the other hand, the above attack design may not require access to the true plant state x_t^a, since only the 'shifted' (i.e., $x_t^a - s_t$) outputs of the real sensing/perception are injected. For instance, in the lane centring control (i.e., keeping the vehicle between the lanes), $G(x_t - s_t)$ only shifts the actual image s_t to the right or left depending on the coordinate definition.

The idea behind the above attacks is to have the system believe that its (plant) state is equal to the state $e_t \triangleq x_t^a - s_t$; thus, referred to as the *fake state*. Note that effectively both z_t^a and $y_t^{s,a}$ used by an AD are directly function of the fake state e_t. Thus, if the distribution of $e_0 : e_t$ is close to $x_0 : x_t$ (i.e., attack-free trajectory), then the attacker will be successful in designing stealthy attacks.

Definition 9 *For an attack-free state trajectory $x_0 : x_t$, and for any $T \geq 0$ and $b_x > 0$, $\delta(T, b_x, b_v)$ is the probability that the system state and physical sensor noise v^s remain in the ball with radius b_x and b_v, respectively, during $0 \leq t \leq T$ – i.e.,*

$$\delta(T, b_x, b_v) \triangleq \mathbb{P}\big(\sup_{0\leq t\leq T} \|x_t\| \leq b_x, \sup_{0\leq t\leq T} \|v_t^s\| \leq b_v\big).$$

The next result captures conditions under which a perception-based control system is not resilient to attacks, in the sense that it is (ϵ,α)-attackable.

Theorem 10 *Consider the system (1) with closed-loop control as in Assumption 2. Assume that the functions f, f' and Π' (i.e., derivatives of f and Π) are Lipschitz, with constants L_f, L'_f and L'_Π, respectively, and let us define $L_1 = L'_f(b_x + 2b_\zeta + d)$, $L_2 = \min\{2L_f, L'_f(\alpha + b_x + b_\zeta)\}$ and $L_3 = L'_\Pi(b_x + d + b_v)$. Moreover, assume that b_x has the maximum value such that the inequalities $L_1 + L_3\|B\| < \frac{c_3}{c_4}$ and $L_2 b_\zeta < \frac{c_3 - (L_1 + L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r$, for some $0 < \theta < 1$, are satisfied. Then, the system (1) is (ϵ,α)-attackable with probability $\delta(T(\alpha + b + b_x, s_0), b_x, b_v)$ for some $\epsilon > 0$, if $f \in \mathcal{U}_\rho$ with $\rho = 2L_f(b_x + b + b_\zeta)$ and $b = \frac{c_4}{c_3 - (L_1 + L_3\|B\|)c_4}\sqrt{\frac{c_2}{c_1}}\frac{L_2 b_\zeta}{\theta}$.*

Proof Due to space constraint the proof is provided in Khazraei et al. (2021). ■

From (5), c_3 can be viewed as a measure of the closed-loop system stability (larger c_3 means the system is 'more' stable); on the other hand, from Theorem 10, closed-loop perception-based systems with larger c_3 are more vulnerable to stealthy attacks as the conditions of the theorem are easier to satisfy. However, if the plant's dynamics is very unstable, $T(\alpha + b_x + b, s_0)$ is smaller for a fixed α and s_0. Thus, the probability of attack success $\delta(T(\alpha + b_x + b, s_0), b_x, b_v)$ is larger for a fixed b_x. Moreover, in the extreme case when $b_\zeta = 0$ (i.e., the attacker can exactly estimate the plant state), the condition $L_2 b_\zeta < \frac{c_3 - (L_1 + L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r$ will be relaxed and the other condition $L_1 + L_3\|B\| < \frac{c_3}{c_4}$ becomes less restrictive as L_1 becomes smaller. Therefore, in this case, if the attacker initiate the attack with arbitrarily small s_0, then ϵ *can be arbitrarily close to zero and the attack will be very close to being strictly stealthy*. Hence, the following result holds.

Corollary 11 *Let $b_\zeta = 0$, $L_1 + L_3\|B\| < \frac{c_3}{c_4}$ with $L_1 = L'_f(b_x + d)$, and $L_3 = L'_\Pi(b_x + d)$. Then, if $f \in \mathcal{U}_\rho$ with $\rho = 2L_f(b_x + b)$, the system from (1) is (ϵ,α)-attackable with probability $\delta(T(\alpha + b_x + b, s_0), b_x, b_v)$, where $\epsilon = \sqrt{1 - e^{-b_\epsilon}}$ for*

$$b_\epsilon = \Big(\lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1}) \times \min\{T(\alpha + b_x + b, s_0), \sqrt{\frac{c_2}{c_1}}\frac{e^{-\beta}}{1 - e^{-\beta}}, \}\Big)\|s_0\|^2.$$

Finally, the above results depend on determining ρ such that $f \in \mathcal{U}_\rho$. Hence, the following result provides a sufficient condition for $f \in \mathcal{U}_\rho$.

Proposition 12 *Let $V : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function satisfying $V(0) = 0$, and define $U_{r_1} = \{x \in B_{r_1} \mid V(x) > 0\}$. Assume that $\|\frac{\partial V(x)}{\partial x}\| \leq \beta(\|x\|)$, and for any $x \in U_{r_1}$ it holds that $V(f(x)) - V(x) \geq \alpha(\|x\|)$, where $\alpha(\|x\|)$ and $\beta(\|x\|)$ are in class of $\mathcal{K}$ functions (see Khalil (2002)). Further, assume that r_1 can be chosen arbitrarily large. Now, if $\lim_{\|x\|\to\infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} \to \infty$, then $f \in \mathcal{U}_\rho$ for any $\rho > 0$. However, if $\lim_{\|x\|\to\infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} = \gamma$, then $f \in \mathcal{U}_\rho$ for any $\rho < \gamma$.*

Proof Due to space constraint the proof is provided in Khazraei et al. (2021). ■

Note that our results only focus on the existence of perception measurements $G(x_t - s_t)$, obtained by shifting the current perception scene by s_t, that results in (ϵ, α)-successful attack, and not how to compute it. Further, to derive attack sequence using Attack Strategy *I*, the attacker needs the estimation of the plant states. Thus, in the next subsection, we introduce Attack Strategy *II* that relaxes this assumption, with the attacker only needing to have knowledge about the plant's (open-loop) dynamics f and the computation power to calculate $s_{t+1} = f(s_t)$ ahead of time.

3.2. Attack Strategy *II*: Using Plant Dynamics

Similarly to Attack Strategy *I*, consider the attack sequence where z_t^a and $y_t^{s,a}$, for all $t \geq 0$, satisfy

$$z_t^a = G(x_t^a - s_t), \quad y_t^{s,a} = C_s(x_t^a - s_t) + v_t^{s,a}, \quad s_{t+1} = f(s_t), \tag{12}$$

for some nonzero s_0. Here the attacker does not need an estimate of the plant's state; they simply follow the plant's dynamics $s_{t+1} = f(s_t)$ to find the desired 'perturbations' of the compromised measurements. Now, we define the state $e_t \triangleq x_t^a - s_t$ as the *fake state*, and the attacker's goal is to make the system believe that the plant state is equal to e_t.

Theorem 13 *Consider the system (1) with closed-loop control as in Assumption 2. Assume that both functions f' and Π' are Lipschitz, with constants L'_f and L'_Π, respectively. Moreover, assume that b_x has the maximum value such that the inequalities $L_1 + L_3\|B\| < \frac{c_3}{c_4}$ and $L_2 b_x < \frac{c_3 - (L_1 + L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r$, with $0 < \theta < 1$ are satisfied, where $L_2 = L'_f(\alpha + b_x)$, $L_1 = L'_f(\alpha + d)$ and $L_3 = L'_\Pi(b_x + d + b_v)$. Then, the system (1) is (ϵ,α)-attackable with probability $\delta(T(\alpha + b_x + b, s_0), b_x, b_v)$ and $b = \frac{c_4}{c_3 - (L_1 + L_3\|B\|)c_4}\sqrt{\frac{c_2}{c_1}}\frac{L_2 b_x}{\theta}$, for some $\epsilon > 0$, if $f \in \mathcal{U}_0$.*

Unlike in Theorem 10, L_1 and L_2 in Theorem 13 increase as α increases. Therefore, unless $L'_f = 0$, one cannot claim that the attack can be ϵ-stealthy for arbitrarily large α as the inequality $L_1 + L_3\|B\| \leq \frac{c_3}{c_4}$ may not be satisfied. However, in an extreme case where the system is linear, it holds that $L'_f = 0$ and $L_1 = L_2 = 0$. Also, for linear time-invariant (LTI) systems, Attack Strategies *I* and *II* become identical as $s_{t+1} = A(\hat{x}_t^a) - A(\hat{x}_t^a - s_t) = As_t$. Hence, the following holds.

Corollary 14 *Consider an LTI perception-based control system with $f(x_t) = Ax_t$. If $L_3\|B\| < \frac{c_3}{c_4}$ with $L_3 = L'_\Pi(b_x + d + b_v)$, and the matrix A is unstable, then the system is (ϵ,α)-attackable with probability $\delta(T(\alpha + b_x, s_0), b_x, b_v)$, for arbitrarily large α and $\epsilon = \sqrt{1 - e^{-b_\epsilon}}$, where*

$$b_\epsilon = \Big(\lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1}) \times \min\Big\{\max\{T(\alpha + b_x, s_0), t_1\}, \sqrt{\frac{c_2}{c_1}\frac{e^{-\beta}}{1 - e^{-\beta}}}\Big\}\Big)\|s_0\|^2.$$

Both Theorems 10 and 13 assume an *end-to-end* controller that directly maps the perception and sensor measurements to the control input. However, there are controllers that first extract the state information using the perception module P and then use a feedback controller to find the control input (e.g., Dean et al. (2020a,b); Dean and Recht (2020)). For instance, for a linear feedback controller denoted by $u_t = \Pi(y_t) = Ky_t$, we have that $L'_\Pi = 0$. Thus, we obtain the following.

Corollary 15 *Consider an LTI perception-based control system with $f(x_t) = Ax_t$ and a linear feedback controller. If the matrix A is unstable, the system is (ϵ,α)-attackable with probability one for arbitrarily large α and $\epsilon = \sqrt{1 - e^{-b_\epsilon}}$, where $b_\epsilon = \Big(\lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T \Sigma_v^{-1} C_s + \Sigma_w^{-1}) \times \sqrt{\frac{c_2}{c_1}} \frac{e^{-\beta}}{1-e^{-\beta}}\Big) \|s_0\|^2$ and $e^{-\beta}$ is the largest eigenvalue of the closed-loop system.*

Finally, note that for an LTI system, the attacker will be impactful if and only if s_0 is not orthogonal to all unstable eigenvectors of the matrix A. Therefore, by choosing such s_0 that is also arbitrarily close to zero the attacker can be ϵ-stealthy with ϵ being near zero.

4. Simulation Results

We illustrate and evaluate our vulnerability analysis of perception-based controller systems on a case-study. Specifically, we consider a fixed-base inverted pendulum equipped with an end-to-end controller and a perception module that estimates the pendulum angle from camera images. By using $x_1 = \theta$ and $x_2 = \dot{\theta}$, the inverted pendulum dynamics can be modeled in the state-space form

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{g}{r} \sin x_1 - \frac{b}{mr^2} x_2 + \frac{L}{mr^2}; \end{aligned} \tag{13}$$

here, θ is the angle of pendulum rod from the vertical axis measured clockwise, b is the Viscous friction coefficient, r is the radius of inertia of the pendulum about the fixed point, m is the mass of the pendulum, g is the acceleration due to gravity, and L is the external torque that is applied at the fixed base Formal'skii (2006). Finally, we assumed $g = 9.8$, $m = .2Kg$, $b = .1$, $r = .3m$ and discretized the model with $T_s = 10\ ms$.

Using Lyapunov's indirect method one can show that the origin of the above system is unstable because the linearized model has an unstable eigenvalue. However, the direct Lyapunov method can help us to find the whole unstable region $-\pi < \theta < \pi$ (see Khalil (2002)). We used a data set $\mathcal{S}$ with 500 sample of pictures of the fixed-base inverted pendulum with different angles in $(-\pi, \pi)$ to train a DNN P (perception module) to estimate the angle. Fig. 2 shows the predicted values of the trained DNN with data set $\mathcal{S}$ versus the actual pendulum pod angle. The angular velocity is also measured directly by the sensor. We trained a deep reinforcement learning-based controller directly mapping the image pixels and angular velocity values to the input control. To detect the presence of attack, we designed a standard χ^2 Kalman filter anomaly detector that receives the data from the perception module as well as the sensed angular velocity, and outputs the residue/anomaly alarm.

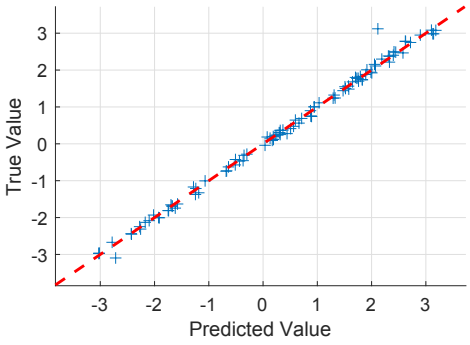

Figure 2: Perception map (P)-predicted value vs true values of pendulum angle θ.

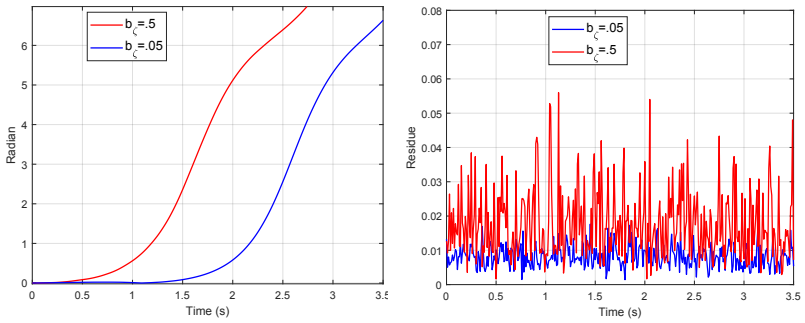

Figure 3: (a) Evolution of the angle's (θ) absolute value over time for different levels of b_ζ. (b) The norm of the residue over time when the attack starts at time $t = 0$.

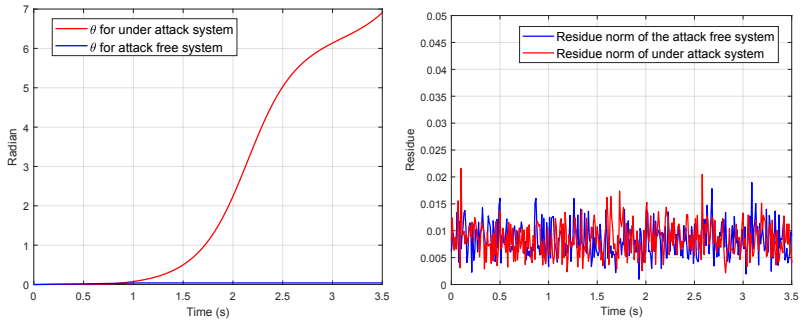

Figure 4: (a) Angle's (θ) absolute value over time for for Attack Strategy *II* (red) and normal condition (blue) (b) The residue norm over time for both under attack and attack-free systems.

We first choose $s_0 = \begin{bmatrix} .001 & .001 \end{bmatrix}^T$ and used Attack Strategy *I* to design attacks. To derive the current adversarial image at each time step the attacker receives the actual image and compromise it by deviating the pendulum rod by s_t degrees. This compromised image is used by the perception module to evaluate the system state. Again, the attacker does not need to have access to the perception map P; the knowledge about the dynamics f and estimate of the current plant state $\hat{x}_t^a$ is sufficient to craft the perturbed images. Fig. 3(a) shows the actual pendulum pod angle for different estimation uncertainty level b_ζ (by the attacker) when the attack starts at $t = 0$. In both cases, the attacker can drive the pendulum pod into an unsafe region. Fig. 3(b) shows the residue signal over time; the attack stealthiness level decreases as b_ζ increases, consistent with our results in Sec. 3.

Fig. 4(b) presents the residue of the system in normal (i.e., attack-free) condition as well as under Attack Strategy *II*. The residue level of both Attack Strategy *I* with $b_\zeta = .05$ and Attack Strategy *II* are the same as for the attack-free system. The red and blue line in Fig. 4(a) also show the pendulum pod angle trajectory for Attack Strategy *II* and in the normal condition, respectively.

5. Conclusion

In this work, we considered the problem of resiliency under sensing and perception attacks for perception-based control systems, focusing on nonlinear dynamical plants. We assumed that the noiseless closed-loop system equipped with an end-to-end controller and anomaly detector, is exponentially stable on a set around the equilibrium point. We introduced the notion of ϵ-stealthiness as a measure of difficulty in attack detection from the set of perception measurements and sensor values. Further, we derived sufficient conditions for an effective yet ϵ-stealthy attack sequence to exists. Here, control performance degradation was considered as moving the system state outside of the safe region defined by a bounded ball with radius α, resulting in an (ϵ, α)-successful attack. Finally, we illustrated our results on fixed-base inverted pendulum case-study.

Acknowledgments

This work is sponsored in part by the ONR under the agreement N00014-20-1-2745, AFOSR under the award number FA9550-19-1-0169, by the NSF under the CNS-1652544 award as well as the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant CNS-2112562.

References

Cheng-Zong Bai, Fabio Pasqualetti, and Vijay Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, 82:251–260, 2017.

Adith Boloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. In *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, pages 1–7. IEEE, 2019.

Adith Boloor, Karthik Garimella, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Attacking vision-based perception in end-to-end autonomous driving models. *Journal of Systems Architecture*, 110:101766, 2020.

Feiyang Cai and Xenofon Koutsoukos. Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 174–183. IEEE, 2020.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4693–4700. IEEE, 2018.

Sarah Dean and Benjamin Recht. Certainty equivalent perception-based control. *arXiv preprint arXiv:2008.12332*, 2020.

Sarah Dean, Nikolai Matni, Benjamin Recht, and Vickie Ye. Robust guarantees for perception-based control. In *Learning for Dynamics and Control*, pages 350–360. PMLR, 2020a.

Sarah Dean, Andrew J Taylor, Ryan K Cosner, Benjamin Recht, and Aaron D Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. *arXiv preprint arXiv:2010.16001*, 2020b.

Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2183–2191, 2019.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

AM Formal'skii. An inverted pendulum on a fixed and a moving base. *Journal of applied mathematics and mechanics*, 70(1):56–64, 2006.

R Spencer Hallyburton, Yupei Liu, Yulong Cao, Z. Morley Mao, and Miroslav Pajic. Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles. In *31st USENIX Security Symposium (USENIX SECURITY)*, 2022.

Maximilian Jaritz, Raoul De Charette, Marin Toromanoff, Etienne Perot, and Fawzi Nashashibi. End-to-end race driving with deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2070–2075. IEEE, 2018.

Yunhan Jia, Yantao Lu, Junjie Shen, Qi A Chen, Zhenyu Zhong, and Tao Wei. Fooling detection alone is not enough: First adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR)*, 2020.

Ilija Jovanov and Miroslav Pajic. Relaxing integrity requirements for attack-resilient cyber-physical systems. *IEEE Transactions on Automatic Control*, 64(12):4843–4858, Dec 2019. ISSN 2334-3303. doi: 10.1109/TAC.2019.2898510.

Hassan K Khalil. Nonlinear systems. 2002.

Amir Khazraei and Miroslav Pajic. Perfect attackability of linear dynamical systems with bounded noise. In *2020 American Control Conference (ACC)*, pages 749–754. IEEE, 2020.

Amir Khazraei and Miroslav Pajic. Attack-resilient state estimation with intermittent data authentication. *Automatica*, page 110035, 2021.

Amir Khazraei, Henry Pfister, and Miroslav Pajic. Resiliency of Perception-Based Controllers Against Attacks, Technical Report. 2021. available at https://cpsl.pratt.duke.edu/files/docs/khazraei_ld4c22_full.pdf.

Amir Khazraei, R. Spencer Hallyburton, Qitong Gao, Yu Wang, and Miroslav Pajic. Learning-based vulnerability analysis of cyber-physical systems. In *13th EEE/ACM International Conference on Cyber-Physical Systems (ICCPS)*, 2022.

Akshay Krishnamurthy. Lecture 21: Minimax theory. 2017.

Harold J Kushner. A partial history of the early development of continuous-time nonlinear stochastic systems theory. *Automatica*, 50(2):303–334, 2014.

Alexander Lambert, Amirreza Shaban, Amit Raj, Zhen Liu, and Byron Boots. Deep forward and inverse perceptual models for tracking and prediction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 675–682. IEEE, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Yilin Mo and Bruno Sinopoli. Secure control against replay attacks. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 911–918. IEEE, 2009.

Yilin Mo and Bruno Sinopoli. False data injection attacks in control systems. In *First workshop on Secure Control Systems*, pages 1–6, 2010.

Miroslav Pajic, Insup Lee, and George J. Pappas. Attack-resilient state estimation for noisy dynamical systems. *IEEE Transactions on Control of Network Systems*, 4(1):82–92, March 2017.

Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Riccardo Polvara, Massimiliano Patacchiola, Sanjay Sharma, Jian Wan, Andrew Manning, Robert Sutton, and Angelo Cangelosi. Toward end-to-end control for uav autonomous landing via deep reinforcement learning. In *2018 International conference on unmanned aircraft systems (ICUAS)*, pages 115–123. IEEE, 2018.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Viktor Rausch, Andreas Hansen, Eugen Solowjow, Chang Liu, Edwin Kreuzer, and J Karl Hedrick. Learning a deep neural net policy for end-to-end control of autonomous vehicles. In *2017 American Control Conference (ACC)*, pages 4914–4919. IEEE, 2017.

Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jack Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack. *arXiv preprint arXiv:2009.06701*, 2020.

Roy S Smith. Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, 35(1):82–92, 2015.

Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

Tianju Sui, Yilin Mo, Damián Marelli, Ximing Sun, and Minyue Fu. The vulnerability of cyber-physical system under stealthy attacks. *IEEE Transactions on Automatic Control*, 66(2):637–650, 2020.

Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 877–894, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

André Teixeira, Iman Shames, Henrik Sandberg, and Karl H Johansson. Revealing stealthy attacks in control systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1806–1813. IEEE, 2012.

Hyung-Jin Yoon, Hamid Jafarnejad Sani, and Petros Voulgaris. Learning image attacks toward vision guided autonomous vehicles. *arXiv preprint arXiv:2105.03834*, 2021.