Security of Perception-Based Control Modeling and Fundamental Limits

Miroslav Pajic

CPSL@Duke
Department of Electrical and Computer Engineering
Department of Computer Science
Department of Mechanical Engineering and Material Science
Duke University
Security-Aware Design of CPS with Varying Levels of Autonomy

Control Stack

Mission Planner

Tactical Planner

Low-level Control

Vehicle

Control view

Long-horizon views

Short-horizon views

Continuous/discrete control with constraints

Modeling view

Adding Resiliency

[CCS23*, USENIX Sec’22, TAC23*, TAC22*, CDC21, ICRA21a, ICRA21b, ICRA20, ICRA19, CAV’19a, THMS19]

[Aut22*, TII21, TASE22, CDC19a, CDC19b, IoTDI19]

[CDC22a, CDC22b, L4DC22, ICCPS22, TCPS20, ACC20, AUT22, AUT21, AUT18, TECS17, RTSS17, TCNS17a, TCNS17b, CSM17, CDC18, CDC17,…]

Our Goal: Add resiliency to controls across different/all levels of the autonomy stack
Low-Level Control in the Presence of Attacks

\[
x_{k+1} = f(x_k, u_k) + w_k \\
y_k = g(x_k) + e_k + v_k
\]

\[
e_k^a = x_k^a - \hat{x}_k^a
\]
Can Attacker Reach Any State?

Theorem 1 [1,2,3,4,5]:
A system presented above is perfectly attackable if and only if it is unstable, and at least one eigenvector \( \mathbf{v} \) corresponding to an unstable mode satisfies \( \text{supp}(C\mathbf{v}) \subseteq \mathcal{K} \) and \( \mathbf{v} \) is a reachable state of the dynamic system.

Physics-based detectors cannot always protect us from an intelligent attacker

\[
\begin{align*}
\mathbf{x}_{k+1} &= f(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{w}_k \\
\mathbf{y}_k &= C\mathbf{x}_k + \mathbf{a}_k + \mathbf{v}_k \\
\text{supp} (\mathbf{a}_k) &= \mathcal{K} \\
\mathbf{a}_{k,i} &= 0, \forall i \in \mathcal{K}^c
\end{align*}
\]
What happens when we include perception?


• How vulnerable is autonomy to perception attacks?
  – Physical dynamics – time-series analysis!

• Considering various threat models
  – E.g., effective and stealthy attacks
    \[ y_t^{s,a} = y_t^s + a_t \]

• \( H_0 \): Normal condition (the ID receives \( Y = y_0: y_t \) with distribution \( P \))
• \( H_1 \): Abnormal behavior (the ID receives \( Y^a = y_0^a: y_t^a \) with distribution \( Q \))

• Stealthy against which anomaly detector?

(Standard) Assumptions for Perception-Based Control

**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 $x \in \mathcal{S}$, $\|v^P(x)\| < \gamma_e$. Without loss of generality, in this work we consider the origin as the operating point—i.e., $x_o = 0$.

**Assumption 2:** We assume that the closed-loop noise-free system is exponentially stable on a set $\mathcal{D} = B_d$. Using the converse Lyapunov theorem, there exists a Lyapunov function that satisfies the following inequalities hold with constants $c_1, c_2, c_3$ and $c_4$ on a set $\mathcal{D} = B_d$.

\[
c_1 \|x_t\|^2 \leq V(x_t) \leq c_2 \|x_t\|^2 \quad V(x_{t+1}) - V(x_t) \leq -c_3 \|x_t\|^2
\]

\[
\left\| \frac{\partial V}{\partial x} \right\| \leq c_4 \|x\|
\]
Formal Notion of Stealthiness

- $H_0$: Normal condition (the ID receives $Y = y_0: y_t$ with distribution $P$)
- $H_1$: Abnormal behavior (the ID receives $Y^a = y^a_0: y^a_t$ with distribution $Q$)

Definition: An attack sequence is

- **strictly stealthy** if there exists no detector that satisfies $p^{FA}_t < p^{TD}_t$, for any $t \geq 0$,
- **$\epsilon$-stealthy** if for a given $\epsilon > 0$, there exists no detector such that $p^{FA}_t < p^{TD}_t - \epsilon$ for any $t \geq 0$.

Theorem: An attack sequence is

- **strictly stealthy** if and only if $KL(Q(Y^a_0: Y^a_t) || P(Y_0: Y_t)) = 0$ for any $t \geq 0$,
- **$\epsilon$-stealthy** if it satisfies $KL(Q(Y^a_0: Y^a_t) || P(Y_0: Y_t)) \leq \log\left(\frac{1}{1-\epsilon^2}\right)$ for any $t \geq 0$. 

$p^{FA} = P(\mathcal{D}(\bar{Y}) = 1 | \bar{Y} \sim P)$

$p^{TD} = P(\mathcal{D}(\bar{Y}) = 1 | \bar{Y} \sim Q)$
Definition 2: Attack sequence \( \{ z_0^a, y_0^{s,a}, z_1^a, y_1^{s,a}, \ldots \} \) is an \((\epsilon, \alpha)\)-successful attack if there exists \( t' \geq 0 \) such that \( \| x_{t'}^a \| \geq \alpha \) and the attack is \( \epsilon \)-stealthy for all \( t \geq 0 \).

When such a sequence exists for a system, the system is called \((\epsilon, \alpha)\)-attackable.
Attack Strategies

Idea: \textit{Fake state} \( e = x_t^a - S_t \)

- Attack Strategy I: Using estimate of the plant state

\[
\begin{align*}
    z_t^a &= G(x_t^a - s_t) \\
    s_{t+1} &= f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t) \\
    y_t^{s,a} &= C_s(x_t^a - s_t) + v_t^s \\
    \zeta &= x_t^a - \hat{x}_t^a, \quad \|\zeta\| \leq b_\zeta
\end{align*}
\]

- Attack Strategy II: Does not need the estimate of the plant state

\[
\begin{align*}
    z_t^a &= G(x_t^a - s_t) \\
    s_{t+1} &= f(s_t) \\
    y_t^{s,a} &= C_s(x_t^a - s_t) + v_t^s
\end{align*}
\]
Attack Strategy I: Using Estimate of the Plant State

\[ z_t^a = G(x_t^a - s_t) \]

\[ y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s \]

Attack dynamics: \[ s_{t+1} = f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t) \]

Idea:
Fake state \( e = x_t^a - s_t \)

Assumption: \( \zeta = x_t^a - \hat{x}_t^a, \|\zeta\| \leq b_\zeta \)

**Theorem:** Assume that the functions \( f, f' \) and \( \Pi' \) (i.e., derivatives of \( f \) and \( \Pi \)) are Lipschitz with constants \( L_f, L'_f \) and \( L'_\Pi \), respectively, and let us define

\[ L_1 = L'_f(b_x + 2b_\zeta + d), \quad L_2 = \min\{2L_f, L'_f(\alpha + b_x + b_\zeta)\} \quad \text{and} \quad L_3 = L'_\Pi(b_x + d + b_v). \]

Moreover, assume that \( b_x \) has the maximum value such that the inequalities

\[ L_1 + L_3 \|B\| < \frac{c_3}{c_4} \quad \text{and} \quad L_2 b_\zeta < \frac{c_3 - (L_1 + L_3 \|B\|)c_4}{c_4} \sqrt{\frac{c_1}{c_2}} \theta r \]

for some \( 0 < \theta < 1 \), are satisfied.

Then, the system is \((\epsilon, \alpha)\)-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 \( \rho = 2L_f(b + b_x + b_\zeta) \) and \( b = \frac{c_4}{c_3 - (L_1 + L_3 \|B\|)c_4} \sqrt{\frac{c_2}{c_1}} \frac{L_2 b_\zeta}{\theta} \).
Theorem: Assume that the functions $f'$ and $\Pi'$ (i.e., derivatives of $f$ and $\Pi$) are Lipschitz, with constants $L_f, L'_f$ and $L'_\Pi$, respectively, and let us define $L_1 = L'_f(\alpha + d)$, $L_2 = L'_f(\alpha + b_x)$ 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 $L_2 b_x < \frac{c_3(1+L_1\|B\|)c_4}{c_4} \sqrt{\frac{c_1}{c_2}} \theta r$ for some $0 < \theta < 1$, are satisfied.

Then, the system is $(\epsilon, \alpha)$-attackable with probability $\delta(T(\alpha + b + b_x, s_0), b_x, b_v)$ for some $\epsilon > 0$, if $f \in \mathcal{U}_0$ and $b = \frac{c_4}{c_3(1+L_1\|B\|)c_4} \sqrt{\frac{c_2}{c_1}} L_2 b_x \theta$. 

Attack Strategy II

Attack injection

\[ z_t^a = G(x_t^a - s_t) \]

\[ y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s \]

Attack dynamics: $s_{t+1} = f(s_t)$
Attack on LTI Systems

\[ z_t^a = G(x_t^a - s_t) \]
\[ y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s \]
\[ s_{t+1} = f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t) = A\hat{x}_t^a - A(\hat{x}_t^a - s_t) = As_t = f(s_t) \]

**Corollary 1:** Consider an LTI perception-based control system with \( f(x_t) = Ax_t \). If \( L_3 \| B \| < \frac{c_3}{c_4} \) with \( L_3 = L_1^\Pi(b_x + d + b_v) \) and the matrix \( A \) is unstable, the system is \((\epsilon, \alpha)\)-attackable with probability \( \delta(T(\alpha + b_x, s_0), b_x, b_v) \) for arbitrarily large \( \alpha \) and \( \epsilon = \sqrt{1 - e^{-b_\epsilon}} \), where

\[ b_\epsilon = \left( \lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1})\min \left\{ T(\alpha + b_x, s_0), \frac{c_2}{c_1} \frac{e^{-\beta}}{1 - e^{-\beta}} \right\} \right) \|s_0\| \]
and \( e^{-\beta} \) is the largest eigenvalue of the closed-loop system.

**Corollary 2:** 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 \((\epsilon, \alpha)\)-attackable with probability one for arbitrarily large \( \alpha \) and \( \epsilon = \sqrt{1 - e^{-b_\epsilon}} \), where \( b_\epsilon = \left( \lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1}) \sqrt{\frac{c_2}{c_1} \frac{e^{-\beta}}{1 - e^{-\beta}}} \right) \|s_0\| \) and \( e^{-\beta} \) is the largest eigenvalue of the closed-loop system.
Case Study: Inverted Pendulum

The norm of the residue for attack strategy I

The norm of the residue for attack strategy II
Perception-Based Attacks on UAVs
Stealthy & Effective Attacks
Connected and Automated Vehicles

Perception based on deep neural networks vulnerable to attack

Point cloud (LiDAR) data & algorithms are under-analyzed in the security community

Must consider multi-module, longitudinal AV, unlike previous studies

Must consider multi-sensor fusion, unlike previous studies of LiDAR-only

Saturation Attacks

Clear-Box Adversarial Perturbation

Clear-Box Adversarial Injections

Structured Injection and Spoofing
LiDAR Spoofing Threat Model

**Threat Model**

**Attack Model**
Road-side attack laser, photodiode

**Attacker Knowledge**
Line-of-sight to victim to receive and transmit signal

**Attacker Capability**
Up to 200 spoof points

**Attack Designs**

**Naïve Attack**
Spoofing in front-near position of victim without contextual information

**Frustum Attack**
Spoofing relative to a "target car" -- in front or behind, relative to victim

Naïve spoofing attack at 8m

From Cao et al. 2019
Attacking Camera-LiDAR Perception

• How feasible are such attacks?
  – Physical dynamics – time-series analysis!

• Beyond Naïve Attack: Novel Frustum Attack

\[
z_t^a = G(x_t^a - s_t)
\]
\[
y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s
\]

Three candidate realizations of the frustum attack.
Additional configurations shown later

- Stable spoof points placed in frustum
- Target car in front of victim
- Spoofer set behind target car

Rapid Prototyping/Evaluation is Critical!

AVstack: A Reconfigurable Platform for Autonomous Vehicles

Perception - leveraging MMDet
- 8 LiDAR object detection
- 13 Monocular 2D object detection
- 5 Monocular 3D object detection
- 13 Monocular Instance Segmentation

Localization
- Traditional Kalman Filtering
- Multi-Sensor Kalman Filtering (MSE, e.g., Baidu Apollo)

Tracking
- Integrating Open-source filterPy implementations
- Modular multi-target tracking

Prediction
- Kinematic State Estimation & prediction
- Neural-Network Prediction Agents

Planning - leveraging PythonRobotics
- Custom planning elements, levels 2-5
- A*, D*, RRT*, planning integration

Control
- PID
- MPC

General, Module-Based AV Stack

Perception → Localization → Tracking → Prediction → Control

[ICCPS’23*]
Frustum Attack is Widely Successful

Compromise Fusion (and LiDAR-only)
- Frustum attack demonstrated to compromise BOTH LiDAR-only AND camera-LiDAR fusion
- Frustum attack shown indefensible by state-of-the-art defenses (CARLO, SVF, ShadowCatcher, LIFE)

Extensive Evaluations
- We perform the most extensive evaluation of attacks on perception to-date with 8 algorithms and 4 defenses (7 and 3 for large-scale evaluation)
- > 75 million attack traces evaluated --> number of spoof points, distance of spoof point placement, each object, each frame of data

Frustum attack widely successful with 60 spoof points
Frustum attack successful even with just 2 spoof points!
Attacking Camera-LiDAR Perception: Longitudinal Attacks

Evaluation of Multi-Frame Tracking

- 1-Sigma projected track bounds on [0, 1.2] seconds later
- Track over 9 subsequent injections
- Initial detection of vehicle
- Track over 9 subsequent injections
- 1-Sigma projected track bounds on [0, 2] seconds later
- Final detection of vehicle

Evaluation on industry-grade Avs: Baidu’s Apollo + SVL

Beginning of scene, before spoof

Target car moves as scene evolves...

After frustum spoof

What happens if we consider cyber threat model?
LiDAR security analysis in longitudinal AV situations

Robust Longitudinal Evaluations

Complete AV Algorithms for Evaluations – Same API as for simulators

In-Place Metrics for Evaluation
Do things change when we add radar?

Preliminary Results

Proof of concept attack successfully implemented starting at 80 meters with a velocity of 15m/s towards the victim

\[
z_t^a = G(x_t^a - s_t) \\
y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s
\]
Conclusions

• Security-aware modeling and analysis framework for perception-based control systems

• We cannot solve problems by focusing only on controls/learning

• Critical need to look at realistic attack vectors to provide system assurance

• We do need to reach out of the security community
  • Longitudinal (time-series) analysis
Thank you