# Research Statement: End-to-End Security for Cyber-Physical Systems

Earlence Fernandes University of Wisconsin–Madison

**Vision.** Computers have evolved to control many physical systems around us — electronic appliances, homes, buildings, cars, factories. Different from traditional computing issues, lapses in the security of these cyber-physical systems (CPSs) can lead to direct negative impacts in our physical world. My goal is to secure these systems using an end-to-end perspective, a design principle first discussed by Saltzer, Reed and Clark. My current and future focus is on two critical areas within CPS — (1) hybrid human-machine systems that rely on a combination of human decision making and machine learning for automation; (2) software that stitches together diverse physical devices and makes seamless automation possible.

## 1 Major Research Initiatives

### 1.1 Securing Hybrid Human-Machine Systems

Hybrid systems that combine human and machine strengths are an emerging and critical type of CPS. An example is semi-autonomous vehicles on the road today that provide safety features like forward collision warning and lane keep assist. My goal is to understand the security issues in these hybrid systems and then build secure versions using an end-to-end perspective. I am concretely exploring this problem in the context of driver-assistance systems that are increasingly becoming available on vehicles today. Specifically, my effort is around two thrusts.

#### 1.1.1 Establish and Quantify Risk

As a community, we currently do not understand the scope of possible attacks on human-machine hybrid systems and we lack a way to test for that risk. Before we can begin to create defenses, we must understand the scope and realism of the threat. My work is contributing some of the first studies in this space through the lens of driver-assistance systems. Quantifying the risk in these systems requires understanding the vulnerability of machine learning, traditional control systems and computational models of human behavior. My efforts are structured along several thrusts.

**Sequential Attacks.** Using the example of Forward Collision Warning (FCW), my recent work shows how an attacker can plan a sequence of actions to cause unsafe situations. Specifically, FCW determines whether a collision is about to occur and then warns the driver in time to avoid that collision. It relies on a combination of machine learning (perception), traditional control (vehicle state estimation) and human behavior (reaction time, distractedness, braking habits). My team's recent work developed the notion of a planning-based attack framework [5] that takes into account all of the above components including a simplified model of human decision-making in near-crash scenarios.

**Computational Models of Human Decisions.** Based on these results, several questions arise, that I am planning to investigate. First, how can one automatically derive a computational model of human decision making in near-crash scenarios? A related question is once that model is known, can an attacker exploit human behavior to further increase the efficacy of their attacks? For example, this type of *trust hacking* might occur when a human is far too dependent on the machine part of the hybrid system.

**Time-varying Adversarial Examples.** Second, the attack on the ML components of these hybrid systems requires a notion of time — a concept that does not exist with current adversarial example attacks on computer vision. For example, my past work showed how attackers can create robust physical attacks to perform one-shot attacks on computer vision (e.g., stickers on a Stop sign) [4]. However, in a sequential attack on driver assistance, the perturbation changes over time. Motivated by this, my group is doing early work in dynamic physical adversarial examples [6]. Specifically, we show how an attacker can manipulate the light falling on objects to cause adversarial ML attacks. This type of attack is dynamic because that attacker can simply switch the light patterns at runtime, thus endowing it with a notion of time.

#### 1.1.2 Create End-to-End Security

Based on the lessons from the prior thrust, I will focus on adapting ideas from system security to establish end-to-end guarantees on the behavior of autonomous vehicles. A few potential research directions include: (1) Using the attack algorithms from the prior thrust to build a security testing framework that will help designers determine what kinds of worst-case attacks their systems are vulnerable to; (2) Evaluating whether current techniques in robust training of ML models make achieving these attacks harder by requiring the attacker to expend more resources; (3) Investigating how techniques from robust control will help defend against attacks in an end-to-end manner.

### 1.2 Security and Privacy Foundations for User-Centric Physical Automation

In line with my vision of securing commodity CPSs, my second line of work focuses on physical automation platforms — software that stitches together heterogeneous smart devices and provides a common environment where end-users can write and run control applications. They are a key enabler for the vision of smart homes, buildings and cities.

Unfortunately, their benefits currently come at the price of large-scale security and privacy issues where attackers can steal sensitive data and manipulate physical devices (e.g., an attacker can compromise IFTTT, a popular rule-based physical automation system, and gain instant privileged access to the data and devices of its 20 million users).

The main challenge that I address is providing security and privacy properties despite an adversary who controls the automation platform. Securing computer systems in the face of system compromise is a long-standing challenge in our community that often involves invasive design changes or heavyweight machinery. My key insight is that the unique structure of physical automation platforms enables us to create end-to-end security mechanisms while striking a balance between usability, performance and security.

Specifically, I am introducing hardware-software design approaches with the key guarantee that even if attackers have unfettered access to the platform, the user-specified automations will run correctly without the attacker being able to learn of the user's sensitive data or being able to manipulate the control logic and physical devices. To meet the high bar for security, I introduce lightweight techniques for secure rule execution by leveraging the following properties of physical automation: (1) Structure of user-created rules; (2) Physics of automation.

In terms of ongoing work, we have a paper accepted to IEEE Security and Privacy 2021 that shows how such platforms can compute on encrypted data using garbled circuits [3]. It leverages the insight that automation programs compute on a limited set of operations, permitting us to build specialized and efficient garbled circuit protocols. Furthermore, the structure of these platforms enable us to use efficient semi-honest implementations of garbled circuits despite the presence of a malicious evaluator. Put together, this is an example of my work achieving an end-to-end security guarantee. We are currently evaluating a different trade-off point in the design space of minimizing the privileges of physical automation platforms [2]. This work explores language-based data minimization [1] as a technique to limit the privacy damage a compromised automation platform can cause while being usable and compatible with existing commercial automation platforms.

## 2 Support from this Fellowship

Funds from the MSR award will help me continue work on the two major research initiatives by hiring PhD students and a possible postdoctoral researcher. Concretely, for my work on evaluating the security of hybrid human-machine systems, a postdoc with expertise in human behavior modeling would complement my expertise in security and controls. A concrete first step there is to analyze NTSB crash data to investigate how decision models can be automatically derived and then integrated with my existing framework. For my work on secure physical automation platforms, the next step is to generalize the framework for practical data minimization with the goal of applying it to other systems that expose REST APIs for device control and data communication.

## 3 Fostering Diversity

My research career and development has greatly benefited from people who are part of under-represented groups and I am paying that forward. Currently, I am a mentor in the Wisconsin Science and Computing Emerging Research Stars (WISCERS) program.[1] Our goal is to encourage undergrad participation in research including those from under-represented minorities. Part of the funding from the MSR fellowship will help me hire WISCERS students for summer research and beyond. In the past, I've participated in similar programs (e.g., Washington State Academic Red Shirts). I also promote diversity in my research group where my PhD students come from diverse backgrounds and gender-identity. The funding will further help me maintain that healthy balance.

## References


[1] Thibaud Antignac, David Sands, and Gerardo Schneider. Data Minimisation: A Language-Based Approach. In Sabrina De Capitani di Vimercati and Fabio Martinelli, editors, *32th IFIP International Conference on ICT Systems Security and Privacy Protection (SEC)*, volume AICT-502 of *ICT Systems Security and Privacy Protection*, pages 442–456, Rome, Italy, May 2017. Springer International Publishing. Part 7: Software Security and Privacy.

[2] Y. Chen, M. Alhanahnah, A. Sabelfeld, R. Chatterjee, and E. Fernandes. Practical Data Access Minimization in Trigger-Action Systems. In *In review at USENIX Security Symposium*, 2021.

[3] Y. Chen, A. Chowdhury, R. Wang, A. Sabelfeld, R. Chatterjee, and E. Fernandes. Data Privacy in Trigger-Action Systems. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy (S&P)*, 2021.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Y. Ma, J. Sharp, R. Wang, E. Fernandes, and X. Zhu. Sequential Attacks on Kalman Filter-based Forward Collision Warning Systems. In *35th AAAI Conference on Artificial Intelligence (AAAI)*, February 2021.

[6] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and E. Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Computer Vision and Pattern Recognition 2021*.


[1] https://wiscers.cs.wisc.edu/