Research Statement

Qingzhao Zhang, University of Michigan

Modern cyber-physical systems (CPSs) develop advanced autonomy, presenting both great opportunities and intricate security/safety challenges. A prime example is autonomous driving systems, which are revolutionizing transportation but also raising public concerns about safety and reliability. To this end:

> My research aims to ensure the **security**, **safety**, and **reliability** of modern cyber-physical systems, given the challenge of formulating intricate **physical constraints** across **multiple system layers**.

Unique challenges in this research arise from the complex physical constraints shaping the behavior of CPS, such as resource limitations, timing requirements, and the laws of physics. Previous security studies often fail to adequately justify the system's compliance to above physical constraints, making the threat model unrealistic. Instead, my research emphasizes the investigation of not only realistic attack vectors that capture end-to-end reproducible attacks but also assessment of their impacts in the physical world. Taking autonomous driving systems as an example, attack vectors must be physically realizable, meaning they can be executed using the limited physical resources available on the vehicles and must adhere to strict timing constraints, such as the decision-making deadlines inherent to autonomous driving. Furthermore, the impacts of these attacks must be observable in the physical environment, ensuring that the resulting vehicle behaviors align with physical dynamics and lead to safety-critical consequences, such as collisions. Only by defining realistic attack vectors and attack goals incorporating sophisticated modeling of the physical constraints, the CPS security analysis ensures its validity in real-world environments.

To address the challenges of formulating and enforcing the aforementioned physical constraints in CPS security or safety analysis, my approach uniquely combines formal mathematical methods and empirical analysis of real-world experiments to model CPS behavior in the real physical environment, and integrates the modeling with techniques of software analysis, adversarial machine learning, and network systems design. My research discovers and analyzes previously unexplored security and safety threats, such as safety-critical autonomous driving software bugs, AI model vulnerabilities leading to vehicle collisions, and inadequate data integrity in real-time vehicular communication. It conveys my key insight that CPS security measures should ensure realizability by analyzing the security threats in the real physical environment.

My research is multifaceted and interdisciplinary, which involves three key components spanning software, AI and network systems domains: (1) *correctness of control software*, (2) *robustness of AI components*, and (3) *security and reliability of multi-agent collaboration*. The remainder of the research statement will summarize major contributions for each component. I then outline future research directions, which not only expand my research methodology to broader CPS applications, but also develop in-depth security solutions such as certification of system behaviors and hardware-related security measures.

Broader interests in large language model (LLM) systems and security: My other projects disclose vulnerabilities of LLMs [10], use LLMs as security tools [6], and improve the efficiency of LLM serving [4, 3].

1 Correctness of Control Software

The correctness of CPS software directly influences the safety and reliability of the entire system. The correctness of such software additionally involves its compliance to safety policies in the physical world, unlike conventional software engineering in only digital spaces. Ensuring such safety-related correctness is challenging due to the representation gap of the code implemented by programming languages and real-world rules defined by natural language. To address the challenge, my research explores semi-automatic software analysis pipelines leveraging formal methods to reason about physical-world impact of software logic.

Formal reasoning of CPS safety. AVChecker [7] is the **first** analysis of driving rule compliance of autonomous driving software, using a framework integrating formal verification techniques and static program analysis. Traditional approaches to AV safety testing, largely based on dynamic, black-box testing methods, fall short in achieving high coverage on testing scenarios. AVChecker overcomes these limitations by employing formal methods to create a detailed representation of driving scenarios, including both map layouts and dynamic behaviors of moving objects. The formal representation seamlessly works with program analysis on the source code to detect inconsistencies between code logic and expected driving behavior. AVChecker uncovers 19 rule violations that could lead to serious safety risks in open-source software Baidu Apollo and Autoware.

Another work of mine [12] leverages a similar approach and proposes a defensive component for industrial control systems, which uses formal methods to validate the compliance of the configuration of industrial control systems to safety specifications. The proposed solution can detect and automatically repair improper system configurations at runtime.

Automated instrumentation of safety policies. We design AVMaestro [13], a novel policy enforcement framework designed to enhance the safety and security of full-stack self-driving systems. AVMaestro stands out for its comprehensive and automated approach, featuring a code instrumentation module that injects safety policies defined by specifications to the AV software with a minimal manual effort.

2 Robustness of AI Components

Building robust AI models for CPS functionalities is a critical challenge. Taking autonomous driving as an example, AI components dominate the perception of the surrounding environments, ranging from object detection to motion prediction. The criticality of these models lies in their need to adapt to diverse real-world road conditions, and make split-second decisions to ensure the safety of vehicle behavior in the physical world. My research emphasizes the importance of incorporating physical constraints in AI robustness analysis, targeting realistic, physically realizable threats that could ultimately lead to unsafe system behaviors.

Adversarial attacks against trajectory prediction. This pivotal research [8] is the first study on the adversarial robustness of trajectory prediction models, a crucial element for the safety and navigation of Autonomous Vehicles (AVs). I developed the **first** adversarial attack on trajectory prediction, wherein attackers strategically control a vehicle near the victim AV while following a meticulously crafted, malicious trajectory. This adversarial trajectory is optimized to maximize the error margin in the victim AV's trajectory prediction algorithms thus could induce erroneous driving decisions. Through rigorous experiments across three different models and datasets, we demonstrated that such adversarial interventions could escalate prediction errors by over 150%, leading to potential safety hazards.

Robust perception pipeline. We design CALICO [5] and Cocoon [2], perception pipelines designed for robustness against adversarial attacks and challenging scenarios. Both methods fuse camera and LiDAR data, where cameras provide detailed texture and color information, while LiDAR delivers precise depth and spatial insights. CALICO employs contrastive learning during pre-train stage, enabling the model to learn comprehensive and complementary representations across modalities, thereby reducing the effectiveness of adversarial attacks targeting a single sensor. Cocoon introduces a conformal analysis module that dynamically adjusts the weights of each sensor during sensor fusion based on their assessed reliability. For instance, in low-light conditions, Cocoon prioritizes LiDAR data, compensating for the decreased performance of cameras in such challenging scenarios.

3 Security and reliability of multi-agent collaboration

Security and reliability issues in CPS multi-agent collaboration are of paramount importance. For instance, the collaborative perception of connected and autonomous vehicles (CAVs) relies on the exchange of sensory data among multiple vehicles to enhance perception capabilities and is inherently vulnerable to a range of security or reliability threats, including malicious data manipulation and real-world data corruption. Therefore, my research aims to build a foundation for the safe and secure deployment of such an emerging application in a realistic setting, considering timing constraints of real-time perception processes and limited computing resources on vehicles. The systematic evaluation is achieved by both simulation on a hybrid of simulators and road tests in a real-world testbed (i.e., Mcity [1]).

Message fabrication in collaborative perception. In this pioneering research [9], we are the first to systematically explore the feasibility of real-time data fabrication attacks within collaborative perception systems. We illustrate how attackers can significantly distort the perception results of CAVs by crafting and delivering malicious data, leading to unwarranted hard braking or increased collision risks. In response to these vulnerabilities, we have developed a novel anomaly detection approach, specifically tailored to identify and counteract such malicious data fabrications by cross-validating the knowledge of different CAVs. Our experimental findings reveal a remarkably high success rate of these attacks (i.e., $> 90\%$) and a high detection rate of the anomaly detection algorithm (91.5%).

Robust data sharing protocol in collaborative perception. Our system RAO [11] is a sophisticated and efficient cooperative perception system addressing several pivotal challenges. First, RAO deploys a novel efficient protocol letting CAVs periodically broadcast their data needs and selectively share the most needed data, thereby achieving a good trade-off between bandwidth overhead and perception accuracy. More importantly, RAO addresses the data asynchrony problem: data items shared by different CAVs are in different timestamps because of the asynchronous nature of data sharing, thus merging them directly often results in inaccurate perception results. RAO innovatively leverages prediction algorithms to compensate for the objects' motion in the time gaps, significantly enhancing the system's overall perception accuracy. Notably, RAO maintains minimal latency and low data transmission overheads.

4 Future Research Directions

In the next 5-10 years, I plan to expand the breadth of my research by applying insights gained from existing research to other cyber-physical system (CPS) applications or components, and develop the depth of my research by developing certified protection that integrates the technology across multiple layers and fields. I am also building broad collaborations with experts in robotics, AI, and hardware security to advance a comprehensive, cross-layer system security analysis, focusing on but not limited to CPS. I will emphasize several research directions as follows:

- **Moving to other CPS applications**. My existing research focuses on autonomous driving and industrial systems, especially on perception and planning tasks. My in-progress research effort is extending the obtained insights to other CPS applications, e.g., drone platforms, localization and mapping tasks, reputation management, etc. For example, drone platforms follow a similar perception-planning-control workflow as autonomous driving systems, making them vulnerable to sensor security issues and adversarial attacks. However, drones have a more complex control model and stricter computational constraints, adding challenges to designing effective real-time security measures.
- **Secure use of emerging technologies in CPS**. Emerging technologies are enhancing CPS design, such as large language models (LLMs) and vision-language models (VLMs) in the context of autonomous driving. This introduces new research questions regarding the adversarial robustness of these models. A future research direction is thoroughly understanding this emerging security threat and securing CPS against these vulnerabilities. My recent research [10] took the initial step analyzing a denial-of-service vulnerability of LLM systems that could prove critical for real-time systems.
- **Certification of system behaviors**. A key direction for my future research is advancing formal verification methods for CPS. While my prior work has applied formal methods to rule-based software logic [7, 12], there remains a gap in certifying AI-based controllers and even entire systems. AI certification can provide hard guarantees on specific behaviors of AI models, offering stronger security properties. Given the challenges of scalability in traditional formal verification, my research will focus on making these methods more practical for real-world CPS applications. Building on my experience in verification and AI robustness, I will continue collaborating with domain experts to drive this effort forward.
- **Hardware-related security measures**. The critical role of hardware in CPSs introduces unique attack models. Existing hardware attacks, such as physical sensor attacks, are often studied in controlled lab environments. Beyond that, I plan to assess the real impact of these threats on end-to-end applications in a dynamic realistic world. This approach will yield findings with more substantial and practical impacts. My experience in real-device security analysis and collaborations with hardware security experts provide a strong foundation for this research direction.

References

[1] Mcity - University of Michigan. `https://mcity.umich.edu/`, 2023.

[2] Cho, M., Cao, Y., Sun, J., Zhang, Q., Pavone, M., Park, J. J., Yang, H., and Mao, Z. M. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. *arXiv preprint arXiv:2410.12592* (2024).

[3] Jin, S., Liu, X., Zhang, Q., and Mao, Z. M. Compute or load kv cache? why not both? *arXiv preprint arXiv:2410.03065* (2024).

[4] Jin, S., Wu, Y., Zheng, H., Zhang, Q., Lentz, M., Mao, Z. M., Prakash, A., Qian, F., and Zhuo, D. Adaptive skeleton graph decoding. *arXiv preprint arXiv:2402.12280* (2024).

[5] Sun, J., Zheng, H., Zhang, Q., Prakash, A., Mao, Z. M., and Xiao, C. Calico: Self-supervised camera-lidar contrastive pre-training for bev perception. *arXiv preprint arXiv:2306.00349* (2023).

[6] Wu, F., Zhang, Q., Bajaj, A. P., Bao, T., Zhang, N., Wang, R., Xiao, C., et al. Exploring the limits of chatgpt in software security applications. *arXiv preprint arXiv:2312.05275* (2023).

[7] Zhang, Q., Hong, D. K., Zhang, Z., Chen, Q. A., Mahlke, S., and Mao, Z. M. A systematic framework to identify violations of scenario-dependent driving rules in autonomous vehicle software. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 5*, 2 (2021), 1–25.

[8] Zhang, Q., Hu, S., Sun, J., Chen, Q. A., and Mao, Z. M. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 15159–15168.

[9] Zhang, Q., Jin, S., Sun, J., Zhang, X., Zhu, R., Chen, Q. A., and Mao, Z. M. On data fabrication in collaborative vehicular perception: Attacks and countermeasures, 2023.

[10] Zhang, Q., Xiong, Z., and Mao, Z. M. Safeguard is a double-edged sword: Denial-of-service attack on large language models. *arXiv preprint arXiv:2410.02916* (2024).

[11] Zhang, Q., Zhang, X., Zhu, R., Bai, F., Naserian, M., and Mao, Z. M. Robust real-time multi-vehicle collaboration on asynchronous sensors. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (2023), pp. 1–15.

[12] Zhang, Q., Zhu, X., Zhang, M., and Mao, Z. M. Automated runtime mitigation for misconfiguration vulnerabilities in industrial control systems. In *Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses* (2022), pp. 333–349.

[13] Zhang, Z., Singapuram, S. S. V., Zhang, Q., Hong, D. K., Nguyen, B., Mao, Z. M., Mahlke, S., and Chen, Q. A. Avmaestro: A centralized policy enforcement framework for safe autonomous-driving environments. In *2022 IEEE Intelligent Vehicles Symposium (IV)* (2022), IEEE, pp. 1333–1339.