Research Statement

Qingzhao Zhang, University of Michigan

Modern cyber-physical systems (CPSs) are becoming a sophisticated integration of control software, artificial intelligence (AI)-based algorithms, and networking capabilities. This integration presents both great opportunities and intricate challenges. A prime example is autonomous driving systems, which are revolutionizing transportation but also raising public concerns about safety and reliability. The fundamental question my research seeks to address is:

> how can we ensure the security, safety, and reliability of modern cyber-physical systems?

The technical approach is to synergize software security methodologies, AI robustness analysis, and network system design to enhance multiple layers of such systems, with careful consideration of their interaction with real-world physical environments. This multifaceted approach is vital as the security of the overall system is determined by its weakest link. Recognizing that few security solutions are free of limitations, it becomes clear that focusing on a single layer of the system is inadequate for comprehensive protection. The scope of my research encompasses three key components:

- **Correctness of control software**. The control software is the chassis of the whole CPS and its vulnerabilities could compromise overall safety. My research builds tools or frameworks to systematically ensure the correctness of CPS software, e.g., compliance with safety specifications, through advanced program analysis, formal verification, and systematic design strategies.
- **Robustness of AI components**. Nowadays, AI algorithms are the backbone of many CPS functionalities. My research in this area explores the vulnerabilities of these algorithms to adversarial attacks and seeks to enhance their robustness by designing robustness mechanisms and improved AI pipelines.
- **Security and reliability of multi-agent collaboration**. CPSs often involve collaboration via computer networks and thus expose various security surfaces such as data integrity and network reliability problems. My research here delves into exploring practical attacks and mitigation measures.

1 Correctness of Control Software

Research results:

- SIGMETRICS 2021 [4]. A software analysis framework exposing software bugs of autonomous driving software causing traffic rule violations.
- RAID 2022 [8]. A formal verification framework detecting and repairing improper configurations of industrial control systems.
- IV 2022 [9]. A source code instrumentation tool automatically enforcing safety policies in autonomous driving software.

The correctness of CPS software directly influences the safety of the entire system. Ensuring the correctness of such software is challenging because (1) the outputs of CPS software manifest directly in the physical world, unlike conventional software engineering in only digital spaces, and (2) the complex system comprising numerous interconnected modules significantly complicates debugging or validation. To address the first challenge, my research explores formal methods to reason about the physical world impact of the code logic. For the second challenge, I built automated pipelines to reduce error-prone manual efforts.

Formal reasoning of CPS safety. AVChecker [4] 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 [8] 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 [9], 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

Research results:

- CVPR 2022 [5]. A new security threat where the attacker triggers road accidents by fooling the trajectory prediction model of autonomous vehicles.
- Arxiv preprint [2]. A benchmark evaluating robustness of point cloud 3D object detection.
- ICLR 2024 [3]. A novel and robust object detection pipeline for autonomous driving.

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 interpret vast and varied sensory data accurately, adapt to unpredictable road conditions, and make split-second decisions to ensure safety. My research adopts multi-faceted approaches towards robust AI: (1) exploring security vulnerabilities that malicious attackers could exploit; (2) building comprehensive benchmarks for evaluating model robustness, and (3) designing novel robust AI pipelines.

Adversarial attacks against trajectory prediction. This pivotal research [5] 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.

Benchmarking robustness of 3D point cloud recognition. We create ModelNet40-C [2], the first comprehensive benchmark for assessing the robustness of deep neural networks against corruptions in 3D point cloud data, which is a type of sensor data widely used by AVs. This benchmark, featuring 15 common and realistic corruptions, provides an essential tool for evaluating the performance of state-of-the-art models in safety-critical applications. Our provided codebase and dataset are publicly available.

Robust perception pipeline. We design CALICO [3], a robust perception pipeline empowered by contrastive learning on both camera and LiDAR data. Each sensor captures distinct aspects of the environment: cameras provide texture and color information, while LiDAR offers precise depth and spatial data. By using contrastive learning, these varying perspectives are all simultaneously involved in the training process, enabling the model to learn rich and complementary representations from diverse data modalities. Such an approach improves the model's robustness against data variations and adversarial attacks that might be less apparent in each sensor modality.

3 Security and reliability of multi-agent collaboration

Research results:

- USENIX Security 2024 [6]. The new data fabrication attack and mitigation in collaborative perception for autonomous driving.
- Mobicom 2023 [7]. A new efficient and robust protocol of collaborative perception that selectively shares the most critical data with tolerance of asynchronous sensors.

The exploration of security and reliability issues in CPS multi-agent collaboration is of paramount importance in the academic and technological realms. 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. My research explores attacks and defenses of such systems, as well as proposes new robust protocols.

Message fabrication in collaborative perception. In this pioneering research [6], 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%), both in a simulation world and in the real-world CAV testing facility (Mcity [1]).

Robust data sharing protocol in collaborative perception. Our system RAO [7] 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. Also, 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 achieves these advancements while maintaining minimal latency and low data transmission overheads.

4 Future Research Directions

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 depth of my research by developing certified protection that integrates the technology across multiple layers and fields. 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 large language models in CPS**. My ongoing research explores emerging technologies that enhance 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, as previous studies have demonstrated that LLMs and VLMs are vulnerable to adversarial attacks. A future research direction is thoroughly understanding this emerging security

threat and securing CPS against these vulnerabilities. My experience in AI robustness analysis provides a strong foundation for advancing this research direction.

- **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 [4, 8], there remains a gap in certifying AI-based controllers to validate entire systems. AI certification can provide hard guarantees on specific behaviors of AI models, offering stronger and more explainable security properties. Given the challenges of scalability and resource demands 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 machine learning experts to drive this effort forward.

- **Hardware-related security measures**. The critical role of hardware in CPSs introduces novel attack models and security properties. Advanced hardware-based security measures, such as trusted execution environments, can provide robust assurances of data integrity, for instance, securing communication in connected vehicles against adversaries in AV software. Additionally, I plan to conduct thorough analysis and experiments on real-world devices, such as on-vehicle sensors and roadside units. This approach will lend greater realism to our attack models and yield findings with more substantial and practical impacts. My prior experience in security analysis on actual devices and collaboration with hardware engineers has laid a foundation for this research direction.

References

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

[2] Sun, J., Zhang, Q., Kailkhura, B., Yu, Z., Xiao, C., and Mao, Z. M. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296* (2022).

[3] 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).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.