Building Safe and Trustworthy Autonomous Systems

As autonomous systems are being increasingly integrated into various aspects of our lives, from self-driving cars to medical diagnosis, ensuring the safety, security and reliability of these systems becomes of utmost importance. The overarching goal of my research is to *develop the foundations for safe and trustworthy autonomous systems*. My works cut across theory, algorithm, and system design, and have appeared at top venues in the areas of machine learning (ML), formal methods, design automation, and cyber-physical systems. My research has also attracted funding from government agencies such as NSF, DARPA, IARPA, ONR and industry partners such as Toyota and Intuit. During my time at Boston University, my research has evolved along the following thrusts. Descriptions of selected research projects can also be found at `https://sites.bu.edu/depend/research/`.

A. Robust, Safe, and Secure Deep Learning Systems

The primary driving force behind modern autonomous systems is deep learning. These are highly parameterized, complex neural networks that can be trained from data to play a variety of roles, ranging from image recognition, through text generation, to making control decisions. Despite their impressive performance, the data-driven nature of these systems cause them to be highly *sensitive to changes in the input*, e.g., tiny perturbations imperceptible to the human eyes can change a network's classification result, irrelevant features in the input can derail a network's output, small errors in the state estimates can cause a significant change in the control output of a neural-network controller, adversarially implanted trigger patterns can force a network to produce a specific prediction regardless of the other input features, etc. In this thrust, through judicious *combinations of formal methods and machine learning*, my team develops novel techniques and tools that enable efficient analysis and training of deep neural networks, specifically addressing challenges related to robustness, safety and security.

Impact: Our research has produced new theories, techniques, and state-of-the-art results on verifying and enhancing **adversarial robustness** [3, 19, 13, 14], as well as improving the robustness of deep reinforcement learning agents **against task-irrelevant input features** [20]. Our studies on **neural Trojans** [28, 27] have received coverage by the WIRED magazine [8]. In addition, we have developed a series of techniques and tools (ReachNN, ReachNN*, POLAR) for **reachability and safety analysis of neural-network controlled systems** (NNCSs) [4, 17, 12, 5, 18]. NNCSs are closed-loop control systems with neural networks acting as the controllers. Our latest tool, POLAR [5], features novel polynomial overapproximation techniques and has state-of-the-art performance in terms of efficiency and tightness of reachable set computation across a diverse set of benchmarks.

B. Specification-Guided Imitation Learning

Imitation learning (IL) is a powerful learning paradigm that enables machines, such as robots or artificial intelligence (AI) systems, to learn from demonstrations provided by human experts or expert agents. It has found numerous applications across various domains ranging from

robotics, through autonomous driving, to text generation. However, as the effectiveness of imitation learning hinges on the quality of demonstrations, it faces several outstanding challenges. Demonstrations, especially human demonstrations on real systems, can be inadequate, partial, imperfect, environment-specific, or suboptimal. In addition, a fundamental problem in inverse reinforcement learning (IRL)-based IL is *reward ambiguity* where many reward functions can explain the same demonstrations, but only some of them will produce good policies. Misaligned rewards can result in poor performance and undesirable behaviors. In this thrust, we explore the *synergies between formal specifications and demonstrations* in IL. We aim to overcome the fundamental limitations of IL such as *overfitting to demonstrations* and *reward ambiguity*.

Impact: We have developed a **novel counterexample-guided framework** that incorporates formal safety specifications in apprenticeship learning (IRL-based IL) [30]. This framework provides a principled way of simultaneously guaranteeing safety and achieving good performance. In addition, we have studied **specification-guided IL in a supervised learning setting**, and shown that we can incrementally improve a low-performance, safe policy to achieve high performance while maintaining safety [29]. For IRL-based IL, we have pioneered the works of using **parameterized programs and automata to specify the reward space** [32, 31]. Moreover, our recent work identifies new theoretical conditions on the reward space and develops a new meta IRL-IL algorithm that can exploit these conditions to **avoid reward misspecification despite the presence of reward ambiguity** [33]. This work represents a significant step towards solving the long-standing challenge of reward ambiguity in IL and IRL.

C. Reliable and Secure Multi-Agent Systems

Recent trends indicate an ever-increasing adoption of autonomous *multi-robot systems* in warehouse automation. Beyond warehouse automation, multi-robot systems also have many other exciting applications, such as shape formation, search and rescue, surveillance and reconnaissance, cooperative target tracking and monitoring, collective transport, etc. On the road, *connected vehicle* technologies hold the promise of resolving long-lasting problems in transportation networks such as accidents, congestion, unsustainable energy consumption and environmental pollution. On the flip side, these multi-agent systems are susceptible to a wide range of attacks and faults, such as actuator or sensor failures, networking faults, denial-of-service attacks, and attacks that can be launched by compromised, malicious agents within the group. In this thrust, we leverage the *unique cyber-physical characteristics* of these multi-agent systems such as their dynamical constraints and physical sensing capabilities to develop novel frameworks to ensure them to operate in a safe, secure and reliable manner.

Impact: Our research has produced breakthrough results in **Byzantine resilience** of robot swarms with hundreds and thousands of robots [25]. In this general setting of Byzantine faults, an unknown subset of the robots can behave arbitrarily or maliciously, such as providing conflicting information or intentionally spreading false data. Compared with the prior state-of-the-art, our new decentralized blocklist protocol generalizes to applications not implemented via the Linear Consensus Protocol and is adaptive to the number of Byzantine robots while simultaneously reducing the required network connectivity. We have also spearheaded the study on a new class of attacks called **plan-deviation attacks** where compromised robots deviate unobserved from planned trajectories in multi-robot systems [23, 24, 26]. For connected and automated vehicles, our recent works have demonstrated **optimal and resilient control and coordination** of such vehicles at traffic bottleneck points [15, 9, 16, 10]. In addition, we have developed novel methods to ensure **reliable communication** between agents in energy-constrained settings [21, 22].

D. Extensible and Adaptive Cyber-Physical Systems

A longstanding problem in the design of cyber-physical systems (CPSs) is the inability to cope with software and hardware evolutions over the lifetime of a design. A fundamental reason for this is that small changes in resource usage can cause big and unexpected changes in the timing and ultimately functionality of the system. In addition to the extensibility challenge at design time, CPSs often need to operate in uncertain environments where the presence of noise and disturbances can degrade performance and even jeopardize the safety of the system. Thus, it is important to develop techniques that enable CPSs to adapt to such uncertainties at runtime. In this thrust, we develop a new *extensibility-driven design* framework that incorporates extensibility as a first-class design objective. In addition, we develop novel techniques that enable CPSs to *adapt to a range of uncertainties and errors* such as sensing noise, software/hardware execution disturbances, and timing violations.

Impact: Our research has made significant headways in **extensibility-driven designs** [37, 35]. In particular, we have developed **a new paradigm for handling weakly-hard constraints** which can be used to model a wide variety of intermittent errors [34, 38, 7, 2, 6, 37], with applications to transportation systems [36, 35], in-vehicle architecture design [11, 37], networked systems [6], and autonomous driving [7, 38]. Our research also paves the way for a new system design methodology that can proactively change system parameters or resource allocations to improve its performance while preserving system safety and timing correctness [34, 1].

Summary of Research Approach and Research Agenda

In my research, I study problems whose solutions can have a transformational impact on an area. I am interested in discovering and formulating new problems, finding new ways to attack an open problem, and ultimately developing new theories or defining a new research area. My principle is to first seek fundamental understandings of the problems that I study, and then develop research ideas that are grounded in theory and supported by empirical observations. Many of the problems that I study are multi-faceted in nature. Thus, I also actively forge and lead collaborative efforts to develop multi-disciplinary and cross-cutting approaches. In addition, I am committed to open-source developments. My group regularly releases code and software tools to encourage community engagements, broader dissemination and collaboration, and to foster transparency and further innovations (`https://github.com/BU-DEPEND-Lab`).

My research agenda is anchored on the following research questions. We want autonomous systems to be trustworthy, but what, precisely, does "trustworthiness" mean? The current, prevalent data-centric view of ML and AI is at odds with the traditional, formal way of specifying input-output behaviors of a system. As a result, ensuring that an autonomous system produces behaviors in accordance with its designer's intent remains an open challenge. Can we reconcile these two different viewpoints? Can we create new paradigms where formal specifications and data can play synergistic roles in learning and system design? What types of guarantees can we achieve with such integrations? How do we address the scalability challenges of applying formal reasoning to large deep learning models? Can we make AI systems simultaneously safe, robust, secure, explainable and achieve high performance? In the pursuit of answering these questions, I also place special emphasis on bringing theory to practice, and work on real systems or prototype systems developed based on realistic applications. Lastly, as part of my research agenda and with broader societal impact in mind, I aim to help foster and grow the community in building better and more trustworthy autonomous systems.

References

The names of the students advised by me are underlined.

[1] Xin Chen, Jiameng Fan, Chao Huang, Ruochen Jiao, **Wenchao Li**, Xiangguo Liu, Yixuan Wang, Zhilu Wang, Weichao Zhou, and Qi Zhu. Safety-assured design and adaptation of connected and autonomous vehicles. In *Chapter in Machine Learning and Optimization Techniques for Automotive Cyber-Physical Systems*. Springer, 2023.

[2] Chao Huang, **Wenchao Li**, and Qi Zhu. Formal verification of weakly-hard systems. In *International Conference on Hybrid Systems: Computation and Control (HSCC)*, 2019.

[3] Chao Huang, Fan, Jiameng, Xin Chen, **Li, Wenchao**, and Qi Zhu. Divide and slide: Layer-wise refinement for output range analysis of deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(11):3323–3335, 2020.

[4] Chao Huang, Fan, Jiameng, **Li, Wenchao**, Xin Chen, and Qi Zhu. Reachnn: Reachability analysis of neural-network controlled systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s):1–22, 2019.

[5] Chao Huang, Jiameng Fan, Xin Chen, **Wenchao Li**, and Qi Zhu. Polar: A polynomial arithmetic framework for verifying neural-network controlled systems. In *The 20th International Symposium on Automated Technology for Verification and Analysis (ATVA)*, 2022.

[6] Chao Huang, Kacper Wardega, **Wenchao Li**, and Qi Zhu. Exploring weakly-hard paradigm for networked systems. In *Workshop on Design Automation for CPS and IoT (DESTION)*, April 2019.

[7] Chao Huang, Shichao Xu, Zhilu Wang, Shuyue Lan, **Wenchao Li**, and Qi Zhu. Opportunistic intermittent control with safety guarantees for autonomous systems. In *ACM/IEEE Design Automation Conference (DAC)*, July 2020.

[8] Will Knight. Tainted data can teach algorithms the wrong lessons. `https://www.wired.com/story/tainted-data-teach-algorithms-wrong-lessons/`, November 2019.

[9] Ehsan Sabouni, H M Sabbir Ahmad, Wei Xiao, Christos G. Cassandras, and **Wenchao Li**. Merging control in mixed traffic with safety guarantees: A safe sequencing policy with optimal motion control. In *The 26th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[10] Ehsan Sabouni, H M Sabbir Ahmad, Wei Xiao, Christos G. Cassandras, and **Wenchao Li**. Optimal control of connected automated vehicles with event-triggered control barrier functions: a test bed for safe optimal merging. In *The 7th IEEE Conference on Control Technology and Applications (CCTA)*, 2023.

[11] **Wenchao Li**, Léonard Gérard, and Natarajan Shankar. Design and verification of multi-rate distributed systems. In *2015 ACM/IEEE International Conference on Formal Methods and Models for Codesign (MEMOCODE)*, pages 20–29, September 2015.

[12] Fan, Jiameng, Chao Huang, **Li, Wenchao**, Xin Chen, and Qi Zhu. Towards verification-aware knowledge distillation for neural-network controlled systems. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8, 2019.

[13] Feisi Fu and **Wenchao Li**. Sound and complete neural network repair with minimality and locality guarantees. In *International Conference on Learning Representations (ICLR)*, 2022.

[14] Feisi Fu, Zhilu Wang, Jiameng Fan, Yixuan Wang, Chao Huang, Qi Zhu, Xin Chen, and **Wenchao Li**. Reglo: Provable neural network repair for global robustness properties. In *NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML)*, 2022.

[15] H M Sabbir Ahmad, Ehsan Sabouni, Wei Xiao, Christos G. Cassandras, and **Wenchao Li**. Evalu-

ations of cyber attacks on cooperative control of connected and autonomous vehicles at bottleneck points. In *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*, 2023.

[16] H M Sabbir Ahmad, Ehsan Sabouni, Wei Xiao, Christos G. Cassandras, and **Wenchao Li**. Trust-aware resilient control and coordination of connected and automated vehicles. In *The 26th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[17] Jiameng Fan, Chao Huang, Xin Chen, **Wenchao Li**, and Qi Zhu. Reachnn*: A tool for reachability analysis of neural-network controlled systems. In *The 18th International Symposium on Automated Technology for Verification and Analysis (ATVA)*, 2020.

[18] Jiameng Fan and **Wenchao Li**. Safety-guided deep reinforcement learning via online gaussian process estimation. *International Conference on Learning Representation (ICLR), Workshop on Safe Machine Learning: Specification, Robustness, and Assurance (SafeML)*, May 2019.

[19] Jiameng Fan and **Wenchao Li**. Adversarial training and provable robustness: A tale of two objectives. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 7367–7376. AAAI Press, 2021.

[20] Jiameng Fan and **Wenchao Li**. Dribo: Robust deep reinforcement learning via multi-view information bottleneck. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[21] Kacper Wardega and **Wenchao Li**. Application-aware scheduling of networked applications over the low-power wireless bus. In *Design, Automation and Test in Europe Conference (DATE)*, March 2020.

[22] Kacper Wardega, **Wenchao Li**, Hyoseung Kim, Yawen Wu, Zhenge Jia, and Jingtong Hu. Opportunistic communication with latency guarantees for intermittently-powered devices. In *Design, Automation and Test in Europe Conference (DATE)*, March 2022.

[23] Kacper Wardega, Roberto Tron, and **Wenchao Li**. Masquerade attack detection through observation planning for multi-robot systems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2019.

[24] Kacper Wardega, Roberto Tron, and **Wenchao Li**. Resilience of multi-robot systems to physical masquerade attacks. In *IEEE Workshop on the Internet of Safe Things (SafeThings)*, May 2019.

[25] Kacper Wardega, Max von Hippel, Roberto Tron, Cristina Nita-Rotaru, and **Wenchao Li**. Byzantine resilience at swarm scale: A decentralized blocklist protocol from inter-robot accusations. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2023.

[26] Kacper Wardega, Max von Hippel, Roberto Tron, Cristina Nita-Rotaru, and **Wenchao Li**. Hola robots: Mitigating plan-deviation attacks in multi-robot systems with co-observations and horizon-limiting announcements. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2023.

[27] Panagiota Kiourti, **Wenchao Li**, Anirban Roy, Karan Sikka, and Susmit Jha. Misa: Online defense of trojaned models using misattributions. In *Annual Computer Security Applications Conference (ACSAC)*, pages 570–585, New York, USA, 2021.

[28] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and **Wenchao Li**. Trojdrl: Evaluation of backdoor attacks on deep reinforcement learning. In *ACM/EDAC/IEEE Design Automation Conference (DAC)*, July 2020.

[29] Weichao Zhou, Ruihan Gao, BaekGyu Kim, Eunsuk Kang, and Wenchao Li. Runtime-safety-guided policy repair. In Jyotirmoy Deshmukh and Dejan Ničković, editors, *Runtime Verification (RV)*, pages 131–150, Cham, 2020. Springer International Publishing.

[30] Weichao Zhou and **Wenchao Li**. Safety-aware apprenticeship learning. In *Proceedings of the 30th International Conference on Computer-Aided Verification (CAV)*, July 2018.

[31] Weichao Zhou and **Wenchao Li**. A hierarchical bayesian approach to inverse reinforcement learning with symbolic reward machines. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[32] Weichao Zhou and **Wenchao Li**. Programmatic reward design by example. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[33] Weichao Zhou and **Wenchao Li**. Pagar: Imitation learning with protagonist antagonist guided adversarial reward. *arXiv preprint arXiv:2306.01731*, 2023.

[34] Zhilu Wang, Chao Huang, Hyoseung Kim, **Wenchao Li**, and Qi Zhu. Cross-layer adaptation with safety-assured proactive task job skipping. *ACM Transactions on Embedded Computing Systems (TECS)*, 2021.

[35] Bowen Zheng, Chung-Wei Lin, Hengyi Liang, Shinichi Shiraishi, **Wenchao Li**, and Qi Zhu. Delay-aware design, analysis and verification of intelligent intersection management. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–8, May 2017.

[36] Bowen Zheng, **Wenchao Li**, Peng Deng, Léonard Gérard, Qi Zhu, and Natarajan Shankar. Design and verification for transportation system security. In *52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6, June 2015.

[37] Qi Zhu, Hengyi Liang, Licong Zhang, Debayan Roy, **Li, Wenchao**, and Samarjit Chakraborty. Extensibility-driven automotive in-vehicle architecture design: Invited. In *Proceedings of the 54th Annual Design Automation Conference (DAC)*, DAC '17, pages 13:1–13:6, New York, NY, USA, 2017. ACM.

[38] Qi Zhu, **Wenchao Li**, Hyoseung Kim, Yecheng Xiang, Kacper Wardega, Zhilu Wang, Yixuan Wang, Hengyi Liang, Chao Huang, Jiameng Fan, and Hyunjong Choi. Know the unknowns: Addressing disturbances and uncertainties in autonomous systems. In *International Conference on Computer Aided Design (ICCAD)*, November 2020.