Enhancing Safety of Robotic Systems with Perception Enabled Components

Kaustav Chakraborty

Ph.D. Candidate, University of Southern California, Los Angeles

My research is driven by the rapid advancements in autonomous systems, particularly those that rely on vision as a primary sensing modality for navigation and control. The widespread use of visual feedback from devices such as LiDARs and cameras, in robotic applications, from autonomous vehicles to indoor robots, presents both significant opportunities and substantial challenges, especially regarding safety and reliability. This document serves as a broad overview of the key papers that have shaped my research trajectory, providing critical insights into these challenges from different but complementary perspectives.

One foundational work that profoundly influenced my interest in vision-based autonomy is *"Combining Optimal Control and Learning for Visual Navigation in Novel Environments"*, by Bansal et al [1]. This paper introduces an approach where a deep neural network processes visual input to predict intermediate goals, which are then tracked by a classical optimal control planner. Integrating *learning-based perception* with *model-based control* enables navigation through complex, previously unseen environments. However, it also raises essential questions about the *safety guarantees* of such hybrid systems, particularly when visual perception encounters unfamiliar scenarios or *adversarial* conditions. This work inspired me to explore the safety vulnerabilities inherent in vision-driven robotic systems.

Building on this motivation, my research [2] provides a methodological framework for assessing the safety of vision-based controllers. This work formulates the problem of identifying *system-level failures* as a Hamilton-Jacobi (HJ) reachability problem. Computing the Backward Reachable Tube (BRT) systematically determines the set of initial states that lead to unsafe conditions under a given vision-based control policy. This approach underscores the need for the *stress testing* of *safety-critical systems*, particularly autonomous vehicles, where rare but potentially catastrophic events must be identified and mitigated.

The concept of computing reachable sets via solving the HJ partial differential equation (PDE) offers a promising avenue for both optimal control and safety analysis of complex visual controllers. Further advancements, such as DeepReach [3], have enabled the integration of physics-inspired machine-learning techniques, allowing my work to scale to higher-dimensional systems while maintaining computational efficiency, leading to the design of fallback controllers targeting online [4] safety, while culminating in an algorithm [5] that *closes the loop* between detecting and mitigating failures for both *runtime* and *design-time* safety.

A unifying thread in my work is re-framing safety from a *component-level safety* analysis to consider the *closed-loop system-level* impact of vision failures. A minor error in image processing (component-level failure) can propagate through the system, leading to unsafe control actions and compromising the overall mission of an autonomous agent (system-level failure). Addressing this challenge, our work [6] introduces SPARQ (Safety Evaluation for Perception and Recovery Q-network), a real-time, data-driven *monitor* that treats safety as a risk estimator over partially observed processes. Unlike conventional *anomaly-detectors* [7] that are often limited to a component-level view and are too computationally demanding [8] for online use, SPARQ reasons about the counterfactual consequences of perception errors, estimating how visual uncertainty affects the likelihood of future safety violations. Its probabilistic approach avoids the over-conservatism of traditional reachability methods [9], striking a more effective balance between safety and operational efficiency.

Beyond serving as an effective runtime safety mechanism, SPARQ revealed a broader insight: data-driven techniques can form the foundation for deploying safety reasoning in complex, dynamic, and multi-agent environments. Building on this key idea, our work FORCE-OPT [10] extends the data-driven safety philosophy into a more rigorous, reachability-theoretic framework — by using trajectory predictors [11] to construct *probabilistic Forward-Reachable Sets*. This formulation grounds the uncertainty captured by these learned prediction models within the semantics of reachability analysis, enabling the safety evaluation of complex, *end-to-end autonomy* stacks that integrate multiple sensing and planning modalities. More importantly, FORCE-OPT translates the benefits from reachability theory to account for unmodelled system states, such as complex interactions arising from the diverse behavior of agents in a multi-agent setting. Properly accounting for these factors is crucial in real-world applications such as self-driving, where a misprediction arising from uncertain driver interactions can lead to catastrophic outcomes.

Through my research, I aim to develop methodologies that enhance the safety of autonomous systems with vision-based controllers, ensuring that these systems can reliably operate in complex, real-world environments. The intersection of learning-based perception, control theory, and formal safety analysis continues to inspire my work, as I look forward to further contributing to this evolving field. An exciting and natural extension of this direction lies in the emerging class of *foundation models for embodied reasoning*, such as *Vision-Language-Action (VLA)* and *Vision-Language Models (VLMs)*, which are rapidly redefining how robots perceive, plan, and interact with their environments. While these models show remarkable generalization and reasoning capabilities, they also introduce *new and poorly understood failure modalities* that classical safety frameworks cannot yet capture.

A major thrust of my future research is to develop algorithms that can *automatically uncover, characterize, and reason* about the failure modes of advanced robotic systems, linking representational uncertainties to system-level safety guarantees. I envision a unified framework where formal reachability theory, data-driven uncertainty quantification, and semantic reasoning coalesce, allowing safety to become an intrinsic property of intelligence rather than a constraint imposed afterward. Ultimately, my goal is to create autonomous systems that not only act intelligently but also understand when and why they might fail and how to remain safe when they do.

References

[1] S. Bansal, V. Tolani, S. Gupta, J. Malik, and C. Tomlin, "Combining optimal control and learning for visual navigation in novel environments," in *CoRL*. PMLR, 2020, pp. 420–429.

[2] K. Chakraborty and S. Bansal, "Discovering closed-loop failures of vision-based controllers via reachability analysis," *IEEE RAL*, 2023.

[3] S. Bansal and C. J. Tomlin, "Deepreach: A deep learning approach to high-dimensional reachability," in *ICRA*. IEEE, 2021.

[4] A. Gupta*, K. Chakraborty*, and S. Bansal, "Detecting and mitigating system-level anomalies of vision-based controllers," in *2024 IEEE ICRA*. IEEE, 2024, pp. 9953–9959.

[5] K. Chakraborty, A. Gupta, and S. Bansal, "Enhancing safety and robustness of vision-based controllers via reachability analysis," *Under review; arXiv preprint arXiv:2410.21736*, 2024.

[6] K. Chakraborty*, Z. Feng*, S. Veer, A. Sharma, B. Ivanovic, M. Pavone, and S. Bansal, "System-level safety monitoring and recovery for perception failures in autonomous vehicles," *Acccepted IEEE ICRA 2025*, 2024.

[7] P. Antonante, H. Nilsen, and L. Carlone, "Monitoring of perception systems: Deterministic, probabilistic, and learning-based fault detection and identification," *arXiv preprint arXiv: 2205.10906*, 2022.

[8] P. Antonante, S. Veer, K. Leung, X. Weng, L. Carlone, and M. Pavone, "Task-aware risk estimation of perception failures for autonomous vehicles," *arXiv preprint arXiv:2305.01870*, 2023.

[9] S. Topan, K. Leung, Y. Chen, P. Tupekar, E. Schmerling, J. Nilsson, M. Cox, and M. Pavone, "Interaction-dynamics-aware perception zones for obstacle detection safety evaluation," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 1201–1210.

[10] K. Chakraborty*, Z. Feng*, S. Veer*, A. Sharma, W. Ding, S. Topan, B. Ivanovic, M. Pavone, and S. Bansal, "Safety evaluation of motion plans using trajectory predictors as forward reachable set estimators," *In submission; arXiv preprint arXiv:2507.22389*, 2025. [Online]. Available: https://arxiv.org/abs/2507.22389

[11] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide, and C. Pal, "Latent variable sequential set transformers for joint multi-agent motion prediction," *arXiv preprint arXiv:2104.00563*, 2021.