From Gambits to Assurances: Game-Theoretic Integration of Safety and Learning for Interactive Robotics

Haimin Hu

Princeton University
Princeton, NJ, USA
haiminh@princeton.edu

Abstract

Autonomous robots are becoming more versatile and widespread in our daily lives. From autonomous vehicles to companion robots for senior care, these human-centric systems must demonstrate a high degree of reliability in order to build trust and, ultimately, deliver social value. How safe is safe enough for robots to be wholeheartedly trusted by society? Is it sufficient if an autonomous vehicle can avoid hitting a fallen cyclist 99.9% of the time? What if this rate can only be achieved by the vehicle always stopping and waiting for the human to move out of the way? I argue that, for trustworthy deployment of robots in human-populated space, we need to complement standard statistical methods with **clear-cut robust safety assurances** under a vetted set of operation conditions. We need **runtime learning** to minimize the robot's performance loss during safety-enforcing maneuvers by reducing its inherent uncertainty induced by its human peers, for example, their intent (does a human driver want to merge, cut behind, or stay in the lane?) or response (if the robot comes closer, how will the human react?). We need to **close the loop** between the robot's learning and decision-making so that it can optimize efficiency by anticipating how its ongoing interaction with the human may affect the evolving uncertainty, and ultimately, its long-term performance.

Previous and Current Work

My vision is to enable human-centered robotic systems that can be built, deployed, and verified with safety assurances under minimal performance loss. The core of my program is to **plan robot motion in the joint space of both physical and information states**, actively ensuring safety and improving efficiency as robots navigate uncertain, changing environments and interact with humans.

Guaranteed Safe HRI with Active Learning. Robots that interact with humans must behave under *verifiable* safety assurances. Under an operational design domain (ODD)—a specification of the robot's deployment environment and failure conditions—strict safety guarantees in human–robot interaction (HRI) may be obtained with *safety filter*, a supervisory control scheme that overrides the task policy to safeguard against the *worst possible* human behavior and external disturbance. However, if the robot's task policy is only driven by its goal and neglects potential "close-call" interactions with the human, it may unwittingly keep triggering safety overrides, needlessly hurting the robot's performance.

To systematically reconcile safety and performance in designing human-centered autonomy pipelines, my key idea is to equip a safety filter with a probabilistic scenario-based *task policy* for predicting a wide spectrum of *multi-modal* interactions between the robot and nearby humans (Hu, Nakamura, and Fisac 2022; Hu et al. 2023a). These scenario predictions enable the robot to *preempt* future costly safety filter interventions and, where possible, adjust its course of action to avoid the risk of having to apply an inefficient last-minute maneuver. My approach transforms conventional probabilistic safety certificates (i.e., gambits) into a performance–computation tradeoff under strict safety guarantees.

While effective at guaranteeing safety, existing safety filters predominantly reason only in the physical space, ignoring the robot's ability to *learn while interacting*, instead assuming static information throughout safety intervention. Building on belief-based game-theoretic planning formalism (Hu et al. 2023a), I developed the first safety filter that closes the safety–learning loop for interactive robotics (Hu* et al. 2023). The key idea is to perform a worst-case game-theoretic safety analysis in an *augmented* state space, which encompasses both physical interactions and the robot's *belief* encoding its uncertainty. Crucially, this framework enables formal robot safety analysis *in closed-loop* with *generative AI models* that predict multi-modal interactions.

Provably Safe and Convergent Robot Learning. Computing a safety-enforcing controller—the key element of a safety filter—is a fundamental open problem for robots with high-dimensional, nonlinear dynamics. On the other hand, recent success in deep learning presents an exciting opportunity to scale up robot safety analysis. I pioneered one of the first deep learning approaches to synthesize *from scratch* a control barrier function (CBF)—one of the most popular safety filters used in robotics (Robey* et al. 2020). For multi-agent problems (e.g., human–robot interaction) where safety must be analyzed in an adversarial setting, interaction-agnostic training can lead to severe oscillatory behaviors, preventing the algorithm from converging to a useful policy. By integrating deep reinforcement learning (RL) with game theory, I designed the first robust-RL-based neural safety synthesis algorithm that is *provably*

convergent (Wang* et al. 2024). The resulting safety filter consistently outperforms the prior state-of-the-art on a 36-dimensional quadrupedal locomotion task. This approach also enables approximate safety analysis in the joint belief–physical space (Hu* et al. 2023), where the overall state space is 200-dimensional.

Despite their promises in scalability, neural safety filters are inherently challenging to yield safety assurances *by design* due to their black-box nature. My insight is that robot safety can be certified by rapidly validating these "untrusted" neural controllers at runtime. I developed one of the first polynomial-time algorithms that efficiently computes a strict, reasonably tight over-estimate of the forward reachable tube for dynamical systems in *closed-loop* with neural network controllers (Hu et al. 2020). The robot can then use this tube within a model-predictive safety filter framework to construct a certified safe "bubble" at runtime, enabling recursive safety.

Scaling Interactive Robot Decision Making. Scaling up decision-making for interactive robotics is challenging as the increase in agent numbers leads to combinatorially many interaction scenarios. To address this demand, I have been leading an interdisciplinary research effort. Collaborating with optimization experts, I proposed a tree-search-based algorithm that computes the *socially optimal* order of play and Stackelberg equilibrium strategy for general N-robot trajectory games (Hu* et al. 2024), yielding ~ 5000 times faster computation than the brute-force approach and 35% reduction in task completion time compared to a state-of-the-art (order-agnostic) dynamic game solver.

Research Agenda

Bridging Dynamic Games and Foundation Models. In recent years, generative AI backed by foundation models (FMs) has begun to revolutionize traditional robot decision-making pipelines. In particular, these models have showed an unprecedented capability to generalize across multiple domains 'zero-shot', showing exciting promise for complex, large-scale applications such as autonomous driving. However, the black-box policies built atop FMs pose significant challenges to guarantee safety in closed-loop, and may struggle to adapt to different user specifications in real time. In recent work, I proposed to blend the robot's generative pre-trained reference policy with a model-based game policy via penalizing their KL divergence, allowing engineers to encode safety and value alignment through the design of dynamic game solvers while inheriting the strong performance provided by the data-driven reference policy (Lidard* et al. 2024). This approach produces realistic and robust policies without the need to manually define the game cost that models interactions. Overall, this idea opens a promising avenue of research that combines the generalization capabilities of FMs with the guarantees and properties of dynamic games, unlocking useful, *provable* features for learned robot policies, ranging from human-in-the-loop safety guarantees (Hu, Nakamura, and Fisac 2022; Hu et al. 2023a) to rapid alignment and disambiguation (Hu and Fisac 2022; Hu et al. 2023b, 2024).

Human-Centric Smooth Safety. The appealing property of smooth and minimal intervention associated with filter-aware motion planning (Hu, Nakamura, and Fisac 2022; Hu et al. 2023a) is not only useful for autonomous driving, but also for emerging human-centric robotics applications. For example, a central challenge in human–robot collaborative fabrication is the need to guarantee safety without ever stopping the robot, as interruptions can lead to poor-quality outcomes, such as uneven surfaces or incomplete structures. This calls for a task policy that *minimally triggers* the safety filter intervention. Building on my expertise in human-predictive safety analysis, I plan to develop a novel *layered* safety analysis framework: The inner layer guards against catastrophic failures (e.g., collisions) using last-resort physical control overrides, while the outer layer minimizes the need for physical interventions by providing *multi-modal* cognitive cues (e.g., visual highlights, haptic nudges, and audio alerts) to guide the human away from inner layer activation surface.

References

Hu, H.; DeCastro, J.; Gopinath, D.; Rosman, G.; Leonard, N. E.; and Fisac, J. F. 2024. Think Deep and Fast: Learning Neural Nonlinear Opinion Dynamics from Inverse Dynamic Games for Split-Second Interactions. *arXiv preprint arXiv:2406.09810*.

Hu*, H.; Dragotto*, G.; Zhang, Z.; Liang, K.; Stellato, B.; and Fisac, J. F. 2024. Who plays first? Optimizing the order of play in Stackelberg games with many robots. *Robotics: Science and Systems (R:SS)*.

Hu, H.; Fazlyab, M.; Morari, M.; and Pappas, G. J. 2020. Reach-SDP: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. *Conference on Decision and Control (CDC)*.

Hu, H.; and Fisac, J. F. 2022. Active uncertainty reduction for human-robot interaction: An implicit dual control approach. *Algorithmic Foundations of Robotics (WAFR)*.

Hu, H.; Isele, D.; Bae, S.; and Fisac, J. F. 2023a. Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach. *The International Journal of Robotics Research (IJRR)*.

Hu, H.; Nakamura, K.; and Fisac, J. F. 2022. SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction. *IEEE Robotics and Automation Letters (RA-L)*.

Hu, H.; Nakamura, K.; Hsu, K.-C.; Leonard, N. E.; and Fisac, J. F. 2023b. Emergent coordination through game-induced nonlinear opinion dynamics. *Conference on Decision and Control (CDC)*.

Hu*, H.; Zhang*, Z.; Nakamura, K.; Bajcsy, A.; and Fisac, J. F. 2023. Deception Game: Closing the safety-learning loop in interactive robot autonomy. *Conference on Robot Learning (CoRL)*.

Lidard*, J.; Hu*, H.; Hancock, A.; Zhang, Z.; Contreras, A. G.; Modi, V.; DeCastro, J.; Gopinath, D.; Rosman, G.; Leonard, N.; Santos, M.; and Fisac, J. F. 2024. Blending data-driven priors in dynamic games. *Robotics: Science and Systems (R:SS)*.

Robey*, A.; Hu*, H.; Lindemann, L.; Zhang, H.; Dimarogonas, D. V.; Tu, S.; and Matni, N. 2020. Learning control barrier functions from expert demonstrations. *Conference on Decision and Control (CDC)*.

Wang*, J.; Hu*, H.; Nguyen, D. P.; and Fisac, J. F. 2024. MAGICS: Convergent Neural Synthesis of Robot Safety. *Algorithmic Foundations of Robotics (WAFR)*.