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

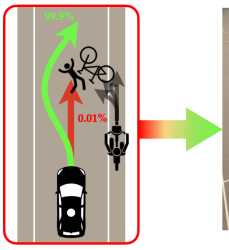
Haimin Hu

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?

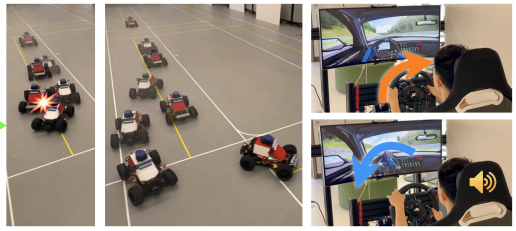
I argue that, for trustworthy deployment of robots in human-populated space, we need to complement standard statistical methods with **clear-cut robust safety assurances** under a vetted set of operation conditions as well established as those of bridges, power plants, and elevators. We need **runtime 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.

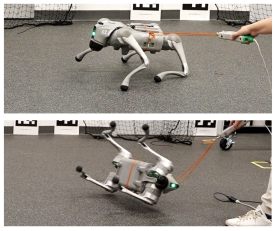
My vision is to enable interactive robotic systems that can be built, deployed, and verified with safety assurances under minimal performance loss. Towards this goal I have developed new algorithms and theorems centered around dynamic game theory, integrating insights from robust optimal control, deep reinforcement learning, generative AI, and numerical optimization. The core of my program is to **plan robot motion in the joint space of both physical and information states**, actively ensuring safety as robots navigate uncertain, changing environments and interact with humans. A consistent principle throughout my research is to ensure that my methods can be validated with hardware tests and that they are reproducible by independent experts. The key contributions of my work include:

1. **Trustworthy human–robot interaction:** planning safe and efficient trajectories by closing the computation loop between interaction and runtime learning that actively reduces the robot's uncertainty about the human [1–7].
2. **Verifiable neural safety analysis for complex robotic systems:** learning robust neural controllers for robots with high-dimensional dynamics; guaranteeing their training-time convergence and deployment-time safety [8–11].
3. **Scalable game-theoretic planning under uncertainty:** collaborating with neuroscientists and operations research experts to develop new game-theoretic methods for complex and uncertain human–robot systems [12–16].



Safety Gambits


1. **Trustworthy Human–Robot Interaction:** safe and smooth human-interactive driving (*left*); safe AI coaching for car racing (*right*)


2. **Neural synthesis of robot safety:** robust quadrupedal locomotion


3. **Scalable game-theoretic planning:** real-time coordination of delivery trucks

Impact. Together, these contributions lay the foundation for next-generation interactive robotic systems deployed with verifiable assurances and real-time adaptability in uncertain, unstructured, and human-populated environments. My work is nominated for the *Roberto Tempo Best Paper Award* at the IEEE Conference on Decision and Control (CDC), and has attracted global attention from industry leaders like Toyota Research Institute and Honda Research Institute who build on my research. In recognition of my contributions, I have been named as a *Human–Robot Interaction Pioneer* (rising young researcher) by IEEE and ACM, and appointed as an *Associate Editor* of IEEE Robotics and Automation Letters (RA-L), a rare honor for a Ph.D. student.

Agenda. As a professor, I intend to lead a research agenda at the intersection of robotics and artificial intelligence, with a clear focus on enhancing public trust in human-centered robotics through transparent safety assurances and cutting-edge capabilities. I will equip my lab with testbeds that I have worked extensively with, including autonomous vehicles [1–4, 7, 12–15, 17], quadrotors [12, 18, 19], and legged robots [8, 11], to carry out both theoretical and experimental work that advances the fields. I will leverage my external collaborators across the industry (Toyota, Honda) and government (NSF, ONR, DARPA) to ensure that my research drives the success of academia and society at large. **The long-term goal of my lab is to develop comprehensive safety principles that contribute to shaping regulatory standards and enhancing public trust in human-centered autonomy.**

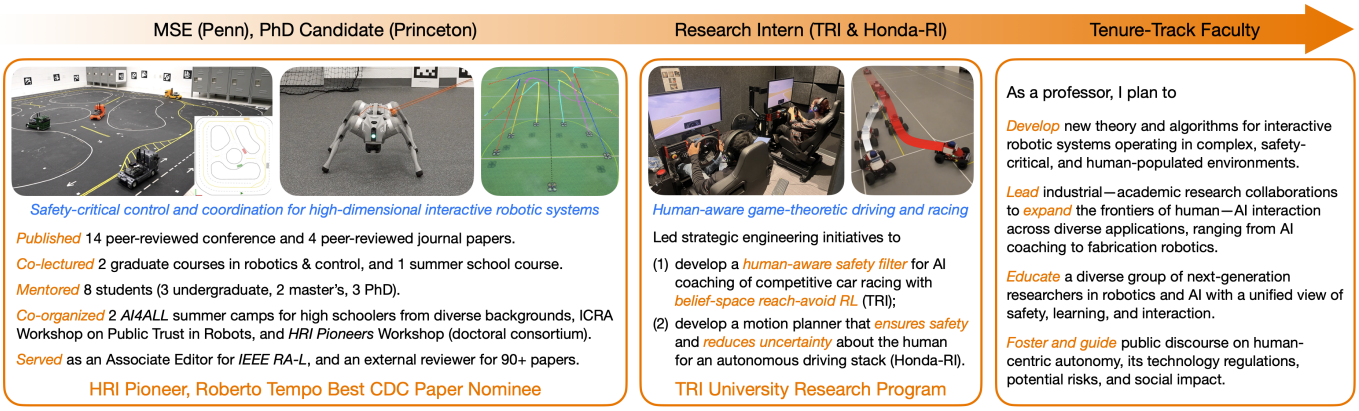


Figure 1: My ongoing career trajectory, delineating past and anticipated contributions to research, education, and outreach.

Previous and Current Work

Guaranteed Safe Human–Robot Interaction 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 may be obtained with *safety filters* [6], a supervisory control scheme that overrides the nominal robot action produced by a task policy to safeguard against the *worst possible* human behavior and external disturbance. However, if the robot's task policy is solely goal-driven and disregards the safety filter during "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, my key idea is to equip a safety filter with a *task policy* that predicts a wide range of possible interactions, induced by factors such as the human's goal or alertness, between the robot and nearby humans [1]. These *multi-modal* 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 (i.e., gambits) into *prediction-centric performance tuning* under *strict* safety guarantees. For example, in an autonomous driving experiment in collaboration with the Honda Research Institute, I demonstrated that my strategy can balance the small cost of slowing down the ego vehicle around a nearby loitering human-driven car against the unlikely, but significant performance loss of an emergency stop if the human abruptly changes the lane and cuts in front of the robot. Importantly, this will *never compromise safety* by risking a catastrophic high-speed collision [3].

Beyond proactive safety, can the robot *actively reduce its uncertainty* about its human peer's intent (e.g., whether to perform a lane change) to make an even more informed decision? Runtime learning (i.e., inference) algorithms, ranging from classical Bayesian inference to contemporary neural trajectory predictors, allow robots to reduce such uncertainty with new information. However, the commonly adopted predict-then-plan pipeline prevents the robot from exploiting its *future* ability to purposefully act and acquire new observations. To this end, I proposed for the first time to model human–robot interaction as a *dual control game*—a game-theoretic variant of the principled dual control theory—allowing the robot to *automatically* trade off nominal performance with reducing its uncertainty via runtime learning [2]. This work has been invited for an *special-issue submission* to IJRR. In this extended article, I showed that, by integrating active learning with filter-awareness, the resulting planner can outperform the prior state-of-the-art by a statistically significant margin in extensive simulated urban driving studies on the Waymo Open Motion Dataset [3]. Notably, this learning-in-the-loop planning formalism is generalizable to collaborative settings, e.g., *apprenticeship learning*. I am working with Toyota to extend this work for building an AI coaching system that teaches amateurs to drive more safely and competitively in car racing [17].

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. This simplification can lead to overly conservative robot behaviors, such as the freezing robot problem, and—in extreme cases—catastrophic safety failures. Building on my dual control game formalism [3], I developed the first safety filter that closes the safety–learning loop for interactive robotics [4]. 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, which represent the state-of-the-art capability to predict multi-modal multi-agent behaviors and is widely used by high-stakes robotics applications such as autonomous vehicles and assistive robots.

Provably Safe and Convergent Learning-based Robot Control. Computing a safety-enforcing controller—the key element of a safety filter—is a fundamental open problem for robots with high-dimensional, nonlinear dynamics: state-of-the-art finite-element methods only scale to 4–5 state variables; other analytical methods require structural assumptions or case-by-case manual derivation. 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 [10, 11]. 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* [8]. 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 [4], 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 [9]. The robot can then use this tube within a model-predictive safety filter [6] to construct a certified safe "bubble" at runtime [19], enabling recursive safety assurances. I also devised a sampling-based algorithm for rapidly certifying the validity of a learned CBF within a region of the robot's operating space [10].

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 at Princeton, 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 [12], yielding $\sim$ 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. In another vein, I have worked with neuroscientists and control theorists to develop methodologies for *resolving ambiguity* in *time-sensitive* multi-agent interactions, from the proverbial corridor deadlock to a busy highway toll station. My game-theoretic planner integrates a nonlinear opinion dynamics (NOD) model, a mathematical abstraction of agents' mutual influence, consensus, and disagreement. Leveraging the bifurcation phenomenon of NOD, I proved that ambiguity in decision-making can be swiftly resolved at an exponential rate, a rare result in dynamic games [13, 14]. In particular, my work [13] has been nominated for the *Roberto Tempo Best CDC Paper Award*.

Research Agenda

With an eye towards a future where humans can unquestionably embrace the presence of robots around them, I envision a *general-purpose* safety framework that defines the *regulatory standard* and *performance benchmarks* of next-generation human-centered robotic systems. Towards this vision, I plan to explore the following research directions:

Bridging Dynamic Games and Foundation Models. In recent years, generative AI backed by foundation models (FMs) has begun to revolutionize the traditional decision-making pipelines in robotics. In particular, these models have demonstrated 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 Kullback–Leibler (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. I have derived an *analytical*, computationally efficient *closed-loop* Nash equilibrium strategy under linear dynamics and Gaussian reference, and extended this solution concept to general nonlinear, multi-modal Markov games leveraging scenario optimization [15]. This approach produces realistic and robust policies without the need to manually define the game cost, thereby mitigating the notorious issue of *reward hacking* associated with hand-crafted costs, rendering more natural robot behavior. I used the solver out-of-the-box for Waymax-based simulated autonomous driving involving up to 14 closely interacting agents, leading to (i) improved safety compared to the SOTA data-driven policy, and (ii) closer resemblance to human behaviors than the SOTA game-based policy. Overall, this idea of guiding model-based interactive strategies with data-driven prior knowledge 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 [1, 3, 4] to rapid alignment and disambiguation [2, 13, 14, 17].

Human-Centric Smooth Safety. The appealing property of smooth and minimal intervention associated with filter-aware motion planning [1, 3] 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 (or even slowing down) the robot, as many high-precision digital fabrication tasks are *continuous*; interruptions can lead to poor-quality outcome, such as uneven surfaces or incomplete structures, requiring costly and time-consuming rework. Similar need for smooth safety intervention arises in other applications such as surgical robotics, where halting can increase patient risk, and human–swarm interaction, where minor deviations by a single robot from the nominal plan can propagate and cause significant delays for the entire swarm, triggering unnecessary human correction. 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. I believe that this kind of layered approach may be a game-changer for human–robot collaboration, freeing robots from human-isolated fences and enabling safe, seamless collaborations.

Accessible Safety Stacks for Human–Robot Interaction. I was recently selected to be a Human–Robot Interaction Pioneer (rising young researcher) for my contributions to trustworthy human–robot interaction through the lens of game theory and belief-space reasoning [5]. As my work gains recognition, researchers and engineers increasingly reach out to me with interest in applying game-theoretic safety filters to their human–robot interaction projects. Oftentimes, I found that while they had a reasonable high-level understanding of the main ideas, they were ultimately unsure about which papers to read or which code repositories to begin with. This indicates that safety filters and safe human–robot interaction techniques remain largely inaccessible for non-experts. Given my expertise, I am well-equipped to lead the effort in *democratizing* this emerging yet crucial subfield of robotics. As an initial step, I have published a review paper with my colleagues, where I have elucidated the fundamental structures of various, seemingly different safety filters and revealed a set of unifying conditions under which a filter can guarantee safety even in highly interactive settings [6]. This paper has quickly become the go-to reference for safety filters. My next steps will prioritize integrating the latest safety frameworks [3, 4] and synthesis algorithms [8, 10] into standardized software, e.g., a ROS2 package, which can be installed with a single command and used plug-and-play. I am also excited to collaborate with computer vision experts to account for sensor imperfections in the design of safety filter pipelines.

My very long-term goal is to collaborate with policy experts, providing a crystal-clear answer to *what guarantees we can offer—and demand—around the deployment of AI-enabled autonomous robots*. The safety assurances and corresponding societal implications of these systems must be as well understood as those of bridges, power plants, and elevators. As roboticists, we must ensure our work is fully accessible to policymakers, regulators, and the general public. To this end, I have organized the inaugural Workshop on Public Trust in Autonomous Systems to be held at ICRA 2025. This event brings together technical and regulatory experts for a focused day-long discussion, targeting new insights on what it would take to establish rigorous foundations for public trust in autonomous systems. I anticipate continued collaboration with these experts for many years to come.

Safety, Alignment, and Trustworthiness for AI Beyond Robotics. Although my research has primarily been focused on robotic systems, I expect my insights to be applicable to other human-centered AI systems. While generative AI such as large language models (LLMs) has recently made monumental successes, ensuring the correct operation of these systems is equally crucial. However, there has been increasing social concern regarding the misuse of these systems, ranging from mind manipulation via deceptive information to exploitation by malicious users to influence public opinion and alter political outcomes. It is of utmost urgency to ensure that these human-interactive AI systems operate in a safe and transparent manner. I believe my expertise in human-centric robotics positions me well to address these emerging AI safety challenges, as they share some of the key traits already explored in my research. For example, in my recent collaboration with Thomas L. Griffiths (Psychology and Computer Science, Princeton University), I have proposed to modify reinforcement learning from human feedback (RLHF)—the state-of-the-art value alignment technique at the core of LLM training—with hindsight feedback [20]. This enhancement significantly *reduces the interaction uncertainty* during human feedback and *mitigates misalignment*, as evidenced by a large-scale human subject study. I have also proved mathematically that hindsight feedback is guaranteed to improve alignment compared to RLHF. In the long run, I will leverage more insights from my work on studying deceptive interactions in robotics [4] to develop new algorithms that detect and prevent manipulative behaviors of generative AI models. In addition, I will develop a principled framework that leverages aligned foundation models for planning complex, human-interactive robotics tasks.

References

[1] H. **Hu**, K. Nakamura, and J. F. Fisac. "SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction". *IEEE Robotics and Automation Letters*, 2022. Presented at ICRA'22.

[2] H. **Hu** and J. F. Fisac. "Active uncertainty reduction for human-robot interaction: An implicit dual control approach". *Algorithmic Foundations of Robotics (WAFR)*, 2022, **Invited extension for IJRR special issue.**

[3] H. **Hu**, D. Isele, S. Bae, and J. F. Fisac. "Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach". *The International Journal of Robotics Research (IJRR)*, 2024.

[4] H. **Hu***, Z. Zhang*, K. Nakamura, A. Bajcsy, and J. F. Fisac. "Deception Game: Closing the safety-learning loop in interactive robot autonomy". *Conference on Robot Learning (CoRL)*, 2023, Short version received **Best Presentation Award** at RSS'24 Safe Autonomy Workshop.

[5] H. **Hu**. "Doxo-Physical Planning: A new paradigm for safe and efficient human-robot interaction under uncertainty". *Human-Robot Interaction Pioneers Workshop*, 2024.

[6] K.-C. Hsu, H. **Hu**, and J. F. Fisac. "The safety filter: A unified view of safety-critical control in autonomous systems". *Annual Review of Control, Robotics, and Autonomous Systems*, 2023.

[7] H. **Hu**, Y. Pu, M. Chen, and C. J. Tomlin. "Plug and play distributed model predictive control for heavy duty vehicle platooning and interaction with passenger vehicles". *Conference on Decision and Control (CDC)*, 2018.

[8] J. Wang*, H. **Hu***, D. P. Nguyen, and J. F. Fisac. "MAGICS: Adversarial RL with Minimax Actors Guided by Implicit Critic Stackelberg for Convergent Neural Synthesis of Robot Safety". *Algorithmic Foundations of Robotics (WAFR)*, 2024.

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

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

[11] L. Lindemann, H. **Hu**, A. Robey, H. Zhang, D. V. Dimarogonas, S. Tu, and N. Matni. "Learning hybrid control barrier functions from data". *Conference on Robot Learning (CoRL)*, 2020.

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

[13] H. **Hu**, K. Nakamura, K.-C. Hsu, N. E. Leonard, and J. F. Fisac. "Emergent coordination through game-induced nonlinear opinion dynamics". *Conference on Decision and Control (CDC)*, 2023, Nominated for the **Roberto Tempo Best Paper Award**.

[14] H. **Hu**, J. DeCastro, D. Gopinath, G. Rosman, N. E. Leonard, and J. F. Fisac. "Think deep and fast: Learning neural NOD from inverse dynamic games for split-second interactions". *International Conference on Robotics and Automation (ICRA)*, 2024 (under review).

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

[16] H. **Hu**, K. Gatsis, M. Morari, and G. J. Pappas. "Non-cooperative distributed MPC with iterative learning". *IFAC World Congress*, 2020.

[17] H. **Hu**, D. D. Oh, J. Lidard, G. Rosman, J. DeCastro, D. Gopinath, N. E. Leonard, and J. F. Fisac. "Learning human-aware safety filters for safe and smooth AI coaching for car racing". *IEEE Transactions on Robotics*, 2024 (in prep).

[18] H. **Hu**, X. Feng, R. Quirynen, M. E. Villanueva, and B. Houska. "Real-time tube MPC applied to a 10-state quadrotor model". *American Control Conference (ACC)*, 2018.

[19] M. Chen*, S. L. Herbert*, H. **Hu**, Y. Pu, J. F. Fisac, S. Bansal, S. Han, and C. J. Tomlin. "FaSTrack: A modular framework for real-time motion planning and guaranteed safe tracking". *IEEE Transactions on Automatic Control*, 2021.

[20] K. Liang, H. **Hu**, R. Liu, T. L. Griffiths, and J. F. Fisac. "RLHS: Mitigating misalignment in RLHF with hindsight simulation", 2024 (submitted to a conference), Short version accepted to NeurIPS 2024 Safe Generative AI Workshop.