Gambits or Assurances? Towards Robust and Verifiable Intelligence for Human-Centered Robotics

Haimin Hu, Princeton University

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

Fig. 1: **My research makes three key contributions towards interactive robotic systems that can be built, deployed, and verified with safety assurances.** *Left:* I proposed a game-theoretic framework that allows robots to plan *verifiably* safe and efficient trajectories around people by closing the computation loop between interaction and runtime learning [1–6]. *Middle:* I developed new algorithms for learning robust neural controllers for robots with high-dimensional dynamics, with *theoretical guarantees* on their training-time convergence and deployment-time safety [7–12]. *Right:* I designed novel and *scalable* game-theoretic motion planners for complex and uncertain human–robot systems [13–15].

Intelligent robotic systems are becoming more versatile and widespread in our daily lives. From autonomous vehicles to companion robots for senior care, these human-centered 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 (Figure 1)? 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 (*"safety gambits"*) with *verifiable 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 or response. 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.

Related work. Interactive robot motion planning is naturally modeled as a non-cooperative dynamic game [16] due to agents' coupled, heterogeneous objectives as in, *e.g.*, autonomous driving [17, 18], collaborative manipulation [19], and physical human–robot interaction (HRI) [20]. The interaction uncertainty further complicates the problem as a partially-observable stochastic game (POSG) [21]. While efficient and scalable solvers have been developed for deterministic games [17, 18, 22], extending those to POSG requires active information gathering to reduce the interaction uncertainty. Existing approaches such as [23–26] rely on heuristic information-gathering mechanisms to reduce uncertainty, which requires manual tuning to balance with the robot's nominal performance, and is oftentimes difficult to scale up. In addition, these approaches occasionally yield contrived robot behaviors governed by manually designed cost functions and may not seamlessly incorporate data-driven policies that closely mimic realistic human behaviors [27, 28], rendering the feasibility of their deployment around humans dubious.

Research vision. I aim to develop *verifiable* decision-making algorithms that ensure *safety* for HRI while minimally affecting the robot's performance. Towards this goal I have developed new algorithms and theorems centered around dynamic game theory, integrating insights from control systems safety, reinforcement learning (RL), and generative AI. *The core of my program is to plan robot motion in the joint space of both physical and belief 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.

I. Research Contributions

My research, summarized in Figure 1, advances the theory and practice of human-centered robotics in three directions:

Robust HRI through belief-space safety filters. Robots that interact with humans must behave under *verifiable* safety assurances. Strict safety guarantees may be obtained with *safety filters* [4, 29], a supervisory control scheme that overrides the robot's task policy to safeguard against unlikely, but safety-critical human behaviors. However, if the robot's task policy is solely goal-driven and disregards the safety filter during "close-call" interactions, it may *unwittingly* keep triggering 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 *belief-space game-theoretic task policy* that predicts a wide range of possible human–robot interaction modes, induced by factors such as the human's goal or alertness [1, 2, 5, 6]. These predictions enable the robot to *preempt* future costly safety filter interventions and, where possible, adjust its course of action *from early on* to avoid the risk of having to apply an inefficient last-minute maneuver. This approach can complement conventional probabilistic safety [30, 31] (i.e., gambits), leading to prediction-centric *performance tuning* under *strict, verifiable* safety guarantees.

While effective at ensuring safety, existing safety filters [32–35] 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 [36], and—in extreme cases—catastrophic safety failures. Building on my belief-space game-theoretic planning framework [2, 6], I developed the first safety filter that closes the safety–learning loop for (human-centered) interactive robotics [3]. The key idea is to perform a *robust* game-theoretic safety analysis in an *augmented* state space, which encompasses both physical interactions and the robot's *belief* encoding the uncertainty about other agents (e.g., its human peers). Crucially, this method enables, for the first time, formal reach–avoid safety analysis *in closed-loop* with *generative AI* models (*e.g.*, [28, 37]), which can efficiently predict multi-modal interaction scenarios at scale.

Provably safe and convergent neural safety filters. 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 [32] only scale to 4–5 state variables; other analytical methods require structural assumptions or case-by-case manual derivation [34]. 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 [8, 9]. For multi-agent problems (e.g., human–robot interaction), where safety must be analyzed in a robust sense, interaction-agnostic training can lead to severe oscillatory behaviors, preventing the algorithm from converging to a useful policy. By integrating deep RL with game theory, I designed the first multi-agent neural safety synthesis algorithm that is *provably convergent* [11]. The resulting safety filter consistently outperforms the prior state-of-the-art [38, 39] on a 36-dimensional quadrupedal locomotion task.

Despite their promises in scalability, neural safety filters can rarely 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 [7]. The robot can then use this tube within a model-predictive safety filter [33, 35] to construct a certified safe "bubble" at runtime [10, 39], enabling recursive safety assurances.

Scaling up interactive robot decision making. In multi-agent settings (i.e., $N \gg 2$), the increase in agent numbers (N) generally leads to combinatorially more interaction scenarios. Leveraging insights from integer programming games [40], I proposed a leader-follower [16, Sec. 7] algorithm that efficiently computes the *socially optimal* interaction strategy [14]. For $N = 10$, the algorithm, on average, yields ~ 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 [22].

II. Future Directions

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 and AI systems. Towards this vision, I plan to explore the following two research directions:

Bridging dynamic games and foundation models. Generative AI backed by foundation models (FMs) has begun to revolutionize the traditional decision-making pipelines in robotics [41]. These models have demonstrated an unprecedented capability to generalize across multiple domains zero-shot. However, the black-box policies built atop FMs pose significant challenges to verify and guarantee safety *in closed-loop*. I plan to leverage dynamic game theory to *blend the robot's generative pre-trained reference policy with a model-based game policy*, which would allow engineers to encode safety and prior knowledge (e.g., robot dynamics) through the design of dynamic game solvers while inheriting the strong performance provided by the data-driven reference policy. This approach would also produce 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.

AI safety beyond physical HRI. While generative AI such as large language models has recently made monumental successes, ensuring the correct operation of these systems is equally crucial—there has been increasing social concern regarding malicious use of these AI systems to manipulate human minds via, for example, fake news or exaggerating information [42]. I believe my expertise in human-centered robotics positions me well to address these emerging AI safety challenges, as they share some of the key traits already explored in my research. In particular, I plan to distill insights from my work on studying deceptive behaviors in HRI [3] to develop new algorithms that can *detect and prevent manipulative behaviors of generative AI models*. In addition, I plan to leverage *aligned* AI models for planning complex and safety-critical HRI tasks in real-world settings.

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.

[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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] D. D. Oh, J. Lidard, H. Hu, H. Sinhmar, E. Lazarski, D. Gopinath, E. S. Sumner, J. A. DeCastro, G. Rosman, N. E. Leonard *et al.*, "Safety with Agency: Human-Centered Safety Filter with Application to AI-Assisted Motorsports," *Robotics: Science and Systems (R:SS)*, 2025.

[13] 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.

[14] 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.

[15] H. Hu, J. F. Fisac, N. E. Leonard, D. Gopinath, J. DeCastro, and G. Rosman, "Think deep and fast: Learning Neural NOD from inverse dynamic games for split-second interactions," *International Conference on Robotics and Automation (ICRA)*, 2025.

[16] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[17] J. F. Fisac, E. Bronstein, E. Stefansson, D. Sadigh, S. S. Sastry, and A. D. Dragan, "Hierarchical game-theoretic planning for autonomous vehicles," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9590–9596.

[18] A. Zanardi, E. Mion, M. Bruschetta, S. Bolognani, A. Censi, and E. Frazzoli, "Urban driving games with lexicographic preferences and socially efficient Nash equilibria," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4978–4985, 2021.

[19] Y. Zhao, B. Huang, J. Yu, and Q. Zhu, "Stackelberg strategic guidance for heterogeneous robots collaboration," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4922–4928.

[20] Y. Li, G. Carboni, F. Gonzalez, D. Campolo, and E. Burdet, "Differential game theory for versatile physical human–robot interaction," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 36–43, 2019.

[21] E. A. Hansen, D. S. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic games," in *AAAI*, vol. 4, 2004, pp. 709–715.

[22] D. Fridovich-Keil, E. Ratner, L. Peters, A. D. Dragan, and C. J. Tomlin, "Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1475–1481.

[23] D. Sadigh, N. Landolfi, S. S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state," *Autonomous Robots*, vol. 42, no. 7, pp. 1405–1426, 2018.

[24] R. Tian, L. Sun, M. Tomizuka, and D. Isele, "Anytime game-theoretic planning with active reasoning about humans' latent states for human-centered robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4509–4515.

[25] Z. Sunberg and M. J. Kochenderfer, "Improving automated driving through POMDP planning with human internal states," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20 073–20 083, 2022.

[26] S. Wang, Y. Lyu, and J. M. Dolan, "Active Probing and Influencing Human Behaviors Via Autonomous Agents," in *Proceedings of (ICRA) International Conference on Robotics and Automation*, May 2023, pp. 1514–1521.

[27] J. Ngiam, V. Vasudevan, B. Caine, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal *et al.*, "Scene transformer: A unified architecture for predicting future trajectories of multiple agents," in *International Conference on Learning Representations*, 2021.

[28] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion Transformer with Global Intention Localization and Local Movement Refinement," *Advances in Neural Information Processing Systems*, 2022.

[29] K. P. Wabersich, A. J. Taylor, J. J. Choi, K. Sreenath, C. J. Tomlin, A. D. Ames, and M. N. Zeilinger, "Data-driven safety filters: Hamilton-Jacobi reachability, control barrier functions, and predictive methods for uncertain systems," *IEEE Control Systems Magazine*, vol. 43, no. 5, pp. 137–177, 2023.

[30] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[31] M. C. Campi, S. Garatti, and F. A. Ramponi, "A general scenario theory for nonconvex optimization and decision making," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4067–4078, 2018.

[32] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, "Hamilton-Jacobi reachability: A brief overview and recent advances," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 2242–2253.

[33] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 7130–7135.

[34] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. IEEE, 2019, pp. 3420–3431.

[35] O. Bastani, "Safe reinforcement learning with nonlinear dynamics via model predictive shielding," in *2021 American control conference (ACC)*. IEEE, 2021, pp. 3488–3494.

[36] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 797–803.

[37] S. Shi, L. Jiang, D. Dai, and B. Schiele, "MTR++: Multi-Agent Motion Prediction with Symmetric Scene Modeling and Guided Intention Querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[38] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2817–2826.

[39] K.-C. Hsu, D. P. Nguyen, and J. F. Fisac, "ISAACS: Iterative Soft Adversarial Actor-Critic for Safety," in *Proceedings of the 5th Conference on Learning for Dynamics and Control*, 2023.

[40] M. Carvalho, G. Dragotto, A. Lodi, and S. Sankaranarayanan, "Integer programming games: a gentle computational overview," in *Tutorials in Operations Research: Advancing the Frontiers of OR/MS: From Methodologies to Applications*. INFORMS, 2023, pp. 31–51.

[41] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *The International Journal of Robotics Research*, p. 02783649241281508, 2023.

[42] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.