Towards Risk Aware Racing Agents: Learning Adaptive Policies in Competitive Racing Games

Kehlani Fay
Dept. of Mathematics
Harvey Mudd College
kfay@g.hmc.edu

Victor Shia
Dept. of Engineering
Harvey Mudd College
vshia@g.hmcedu

***Abstract*— Creating autonomous agents capable of assessing risk-to-safety trade offs in uncertain environments is a key barrier in modern autonomous robotics, preventing autonomous agents from safely operating in the real world. As real world observations are noisy and robots must frequently operate with incomplete state information, autonomous agents must infer and predict future states of other agents, including non-cooperative and competitive agents. Safe planning and decision making under these conditions remains an ongoing challenge in autonomous robotics. This work investigates enabling autonomous agents to develop risk awareness under novel human agent behaviors in a competitive two player racing game through simulation in order to explore prediction of non-cooperative agent's future actions and adaptation to change in an agent's strategy. In contrast with traditional work, which often relies on simplified models of agent processing, this work begins initial modeling of dual agent reasoning and outlines an initial approach and human trials for benchmarking risk levels.**

I. INTRODUCTION

Modern autonomous robots struggle to operate alongside humans while ensuring physical safety. With the increasing integration of autonomous robots into real-world and human-centric spaces, the need to safely interact and predict other agent's actions has become critical to enabling robots to further deploy into multi-agent environments. Current limitations of safety based methods include trade offs between performance and provably safe methods alongside often strong assumption to guarantee safety or performance [1] [2] [3]. In addition, methods in multi-agent decision making struggle with limited ability to interpret nonverbal cues [4] [5] and difficulty of predicting future actions in noisy and partial state information environments [6] [7]. Finding safe and high performance frameworks that overcome these limitations remains an open challenge.

In cases of autonomous driving, autonomous agents interact with dynamic observations and often need to make quick decisions to achieve real-world success. To model strategic behavior between agents and handle inherit uncertainties in real world autonomous driving, several game-theoretic approaches have been introduced. These games include non-cooperative models of multi-agent games[8], Stackelberg games in which decisions are made in a leader-follower style [9][10], and stochastic [11]. These methods have allowed for initial work and analysis in competitive strategic games such as lane merging, handling intersections, and racing.

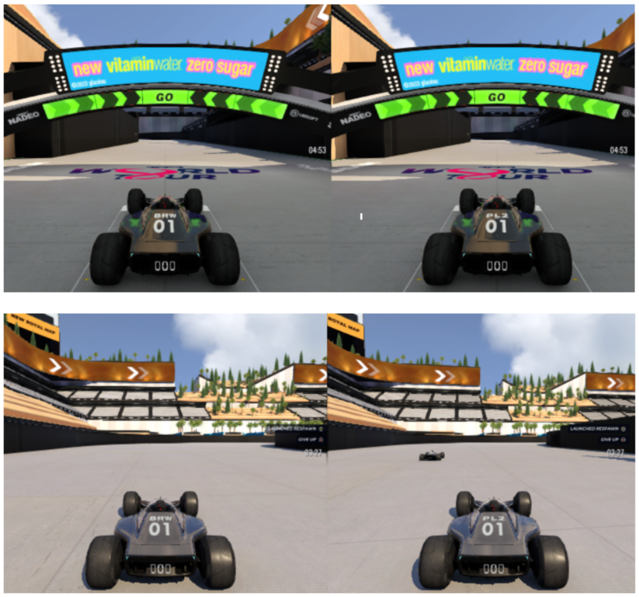

Fig. 1: The Trackmania racing environment with two opponents side by side with the simulated robot policy (left) and human player (right). Two opponents are trained in series using Soft Actor Critic (SAC) policy to create an adaptive and risk aware racing policy which competes with novel human inputs at test time. Top: The simulated starting configuration of both agents with robot (left) and human (right). Bottom: The deployed robot policy and human policy during training.

While earlier game-theoretic models have created models for rationality and behavior modeling, additional work in multi-agent machine learning has been used to learn inherent rewards and adapt to changing observations. While policies learn through trial and error and do not have strong safety guarantees, multi-agent reinforcement learning (MARL) approaches have potential to adapt to novel observations, generalize to novel scenarios, and handle high uncertainty through domain randomization. In order to explore the added benefit of risk awareness and adaptive MARL methods, this work introduces a novel, Kalman inspired, MARL framework for predicting future agents actions, incorporating dual agent risk awareness, and adaptation to novel strategies.

To quantify safety to performance trade offs and high uncertainty of returns, initial research in utilizing risk to robotic safety have been explored. While risk has long been explored in the context of finance for over two decades [12], recent work has introduced risk to autonomous robotics for collision avoidance. Prior methods have evaluated risk metrics under high uncertainty [13], robustness to perturbations [14] [15], and comparisons of risk limitations through axioms [12]. These works have laid the foundation for modeling risk awareness in interactive agents.

Prior work in autonomous racing has been explored both in simulation and real world. Prior simulation methods have utilized game-theoretic approaches in combination with model-based controllers and learning to race through reinforcement learning methods [16]. While current methods are largely based in simulation, the recent advent of the Indy Autonomous Challenge has allowed initial racing methods to be deployed to real albeit via a single agent today [17].

This work aims to address these challenges in competitive multi-agent racing games through learning non verbal cues of human players to inform a risk aware reinforcement policy. By modeling both agents as risk aware and rational, this work explores how predicting human actions can inform risk to safety tradeoffs while maintaining competitive policies. In addition, this work aims to explore the application of a Kalman inspired framework to explore more accurate predictions of an opponents future actions and risk alongside adaption to changes in opponent behavior.

II. Related Works

Prior work on autonomous vehicle racing traditionally focus on game-theoretic methods, however recent approaches have shifted towards multi-agent reinforcement learning methods. Methods include human data collection to inverse learn rewards in combination with dynamical systems modeling [18], mutli-agent game formulations through Stackelberg [19] or Stochastic games [20], and initial methods in multi-agent racing [21]. These approaches have shown learning competitive behaviors is possible but often do not take into account safety-to-risk trade offs for multi-agent games. These methods also often lack adaptive behavior to changing strategies or dynamic risk tolerances.

Multi-Agent Safety Recent methods in human-robot safety have largely adopted probabilistic models of agent behavior to predict future actions with two main camps of thought: provably safe and approximately safe. While provably safe methods offer safety guarantees, the resulting framework is often overly conservative with large trade offs in performance [22]. In Li et al. A provably safe motion planner was deployed to safely dress a human using a robotic arm. However, this method relied on changing the definition of safety to include collision safe which boosted performance. Similar work in provably safe methods often rely on reachability analysis [23], [24] with often strong assumptions on agent behavior or distribution shift detection methods for single agents [14]. While approximately safe methods lack guarantees, they allow more flexibility in human modeling and are not overly conservative. In Fisac et al. [25], real-time model confidence was used to shift the predicted human action distribution based on how rational the person behaved. Similar statistical models have been used with success across crowd navigation [26] and autonomous driving with high success [27].

Risk Several risk metrics exist for measuring risk to reward trade offs. However, when transferring risk profiling to robotics, some risk metrics have competitive advantages due to inherent properties such as Subadditivity and Monotonicity [12]. Risk assessments which meet these properties well include Entropic Value at Risk [28] and Conditional Value at Risk (CVaR) [29]. Prior work in risk aware autonomous driving have used varying definitions of how to measure risk, including exceeding safety bounds with uncertainty [30], agent interaction [20], and probability of loss [31]. This work utilizes CVaR due to its interpretability and meeting desired axiom properties.

Competitive Racing Autonomous racing has taken off in recent years with advancement of multi-agent reinforcement learning methods and game-theoretic deployments. Multi-agent racing frameworks have used constrained dynamic potential games [32], reinforcement learning methods [33], and iterative best response [16] to improve performance. Wu et al. [34] combined risk awareness with reinforcement learning using reshaped rewards to encourage exploration. This paper extends current setups through simulating extending both agents with risk awareness. In addition, this ongoing work seeks to add adaptability to novel strategies through a Kalman inspired framework [35].

III. Methodology

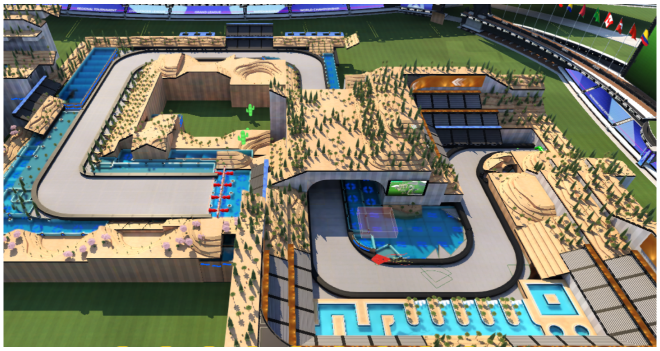

Fig. 2: The custom race track for testing two player races. Both agents start on the fir right hand corner and end on the left hand side. The track is designed to be wide enough to compete but with tight corners for testing risk awareness.

To test and deploy risk-aware policies, a deterministic racing game, Tackmania, with physics real simulation features for vehicle drift, performance across variable track environments (i.e. ice, dirt, gravel), and acceleration and deceleration was chosen. This is a popular racing game with deterministic physics, allowing for replays of the same input output pairs. As Trackmania is intended for single agent races or multi-agent without collisions, the game was adapted with

Fig. 3: The first person view of human and robot players. The Trackmania gym environment is equipped with both lidar or image training for observations.

a custom plugin to send state information of both players and added a collision plugin which created simulated collision physics. A custom race track was created in order to allow space for two agents to race side by side. In order to train the policy, a Trackmania gym environment through TMRL [36] was adapted for two player gymnasium training and then transferred to the racing game environment once initial success benchmarks were met for testing. This open source framework allows for visual observations and initial open source testing against additional players with available online competitions for bench marking algorithms.

The objective is to create a racing policy that can operate under minimal state information of a human player and reliably compete against a human opponent with awareness of risk to safety tradeoffs. To accomplish this, a Kalman-inspired method is used, predicting the opponent's risk, adaptation value, and future actions, correcting these predictions based on actual robot actions, and feeding these predictions into a policy trained on simulated two-player games.

Starting with the robot player, the framework takes in a state history of past states of both players. This history is then passed to a neural net ϕ which outputs an initial estimate of the opponents future risk c, updated state ($\bar{x}_H$), and adaptation value α. These are summed together with the correction parameter from the module ζ. The adjusted predicted future state, history of states, and adjusted risk and adaptation parameters are passed into the base policy for training. The policy outputs an action for the robot and the state of the robot is then updated in sync with the output from the human policy. The updated value and the prior expected value for the human state are finally passed to the ζ module which outputs an antipated correction for each term given the last outcomes.

In order to train a robot policy that will not overfit to limited human player runs, a simulated player is created for training. Similarly, the simulated human player takes in a history of the past states of both players. Φ then outputs the estimated risk and future state of the robot. However, the adaptation value is not given to the human as it is assumed humans are risk aware but not predictive of change in opponents. The estimated risk and estimated future state of the robot are passed to a base policy which outputs the human's action. This allows direct comparison in performance of the human framework and additional modules of the robot framework. At deployment, a real human faces the robot.

For the first loop of the framework, an initial guess of the opponent's future state, risk, and estimated change in risk are given based on average values from human two player trials. This is passed to the base policy with no change from ζ. The base policy then outputs a first action for the robot. Likewise, for the human policy, initial risk and future state of the robot are set. The same process also occurs for the simulated human agent.

Policies are trained using curriculums on risk thresholds and varying adaptation (reflected as change in risk) for the human opponent. Φ is trained through collected two player human racing trials which manually compute the CVAR risk value, change in risk a, and record the future action at each time step. Phi is trained with MSE on each value. Once accurate values of Φ are output, the base policy is then trained for both human and robot pairs. The polciies train in sequence with the curriculum on the human player. After a stable base policy is created with high performance, the policy is frozen and the neural net module ζ is trained to output corrections on the outputs of Φ.

During training, the robot and simulated human model is trained with an estimated position of both players, estimated velocity of the human opponent, and simulated lidar inputs in order to create a policy that can work with limited information. The policy trains with domain randomization on these observations with increased randomization range over time for each measurement as training continues.

Risk Measurement In order to calculate risk of each agent a history of the past 5 actions are analyzed using Conditional Value at Risk. Using a confidence threshold of .90 in order to evaluate the worst 10 percent of potential outcomes. P(x) represents the probability distribution of a given return. Var represents the value at risk. For computing an initial probability of return is computed by the progress along the track minus the safety cost of inter agent distance and proximity to racetrack boundaries. The probability of return is estimated from distance of the agent to boundaries and inter agent distance.

$$CVaR = \frac{1}{1-c}\int_{-1}^{VaR} xp(x)\,dx \tag{1}$$

Rewards The reward of the human and robot are modeled as the same but with different observations, where the robot

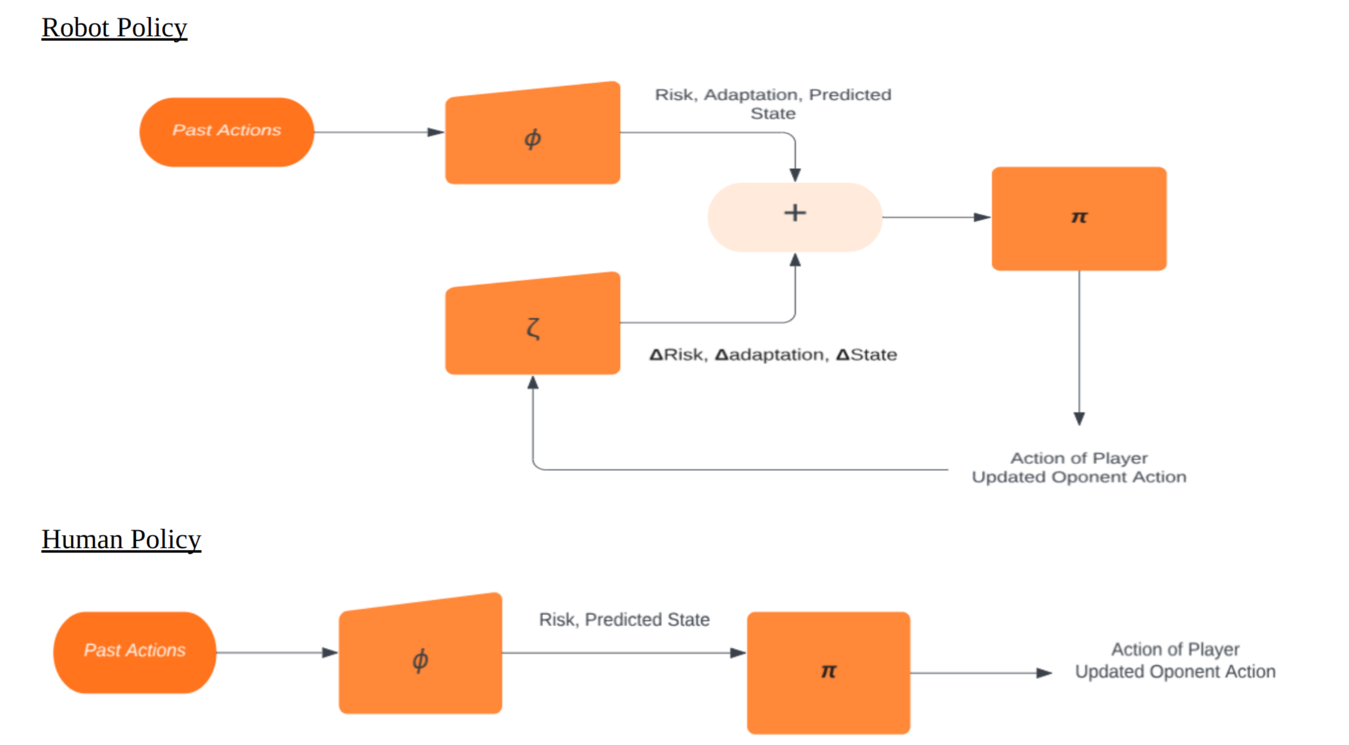

Fig. 4: The policy for Human and Robot Players side by side. Φ and ζ are neural networks which output risk, adaptation, and predicted oponent state and change in these values respectively.

has an additional input of expected adaptation of the human player. This allows the robot to anticipate change in risk for novice players throughout the game and account for change in strategy. The base reward can be computed as R

$$R = \sum_{n=0}^{max_epochs} T(x_r, x_h) - S(x_h, x_r, \bar{x}_o) + I(x_r, x_h, \bar{x}_o) \tag{2}$$

Here the track reward, T(), rewards the agent for distance along the track and distance ahead of the human player where x_r and x_h represent the robot and human state respectively. Using centerline points placed evenly cross the track, the closest point is returned as the current robot and human's distance. T() returns the initial distance of the robot plus the value between these two distances.

The safety cost is returned based on the anticipated future action of the opponent, current state of both players, estimated velocities, and risk. Players at a high speed have a high cost associated scaled by the agents risk for the taken action. Additionally, the safety cost for the distance between agents and distance to the racetrack boundaries is modeled through a sigmoid curve where agents close together have a higher cost than agents farther apart. This is also scaled by the risk of the taken action. Lastly, the difference between the opponent's expected position, $\bar{x}_o$, and the agents position is taken and scaled. The sum of these factors complete the safety cost.

Lastly, an information reward is given which takes in the state of both players and the opponent's expected reward. This returns a reward for the accuracy of the predicted value of the opponent's state in order to encourage slight exploration and information gain of the output action.

Curriculum In order to train the robot policy, risk and change in risk parameters are kept at a constant average at initialization. Upon winning half or more of 50 consecutive races, the human policy increases in range of risk thresholds and change in risk for each episode. This continues until a maximum bound of risk aversion, a policy which tries to get as far from the robot as possible while achieving the goal, and aggression, a policy that does not care about the robot's safety, are met.

IV. Initial Results

As this work is in progress, all results are preliminary and remain ongoing. Thus far, two main components have been completed including an initial survey of human two player races for analyzing the role of risk used to create benchmarks to test the trained policies on and initial training of a robot and human policy.

Initial Human Trials In order to find initial measures for curriculum training and methods to benchmark the policy, 50 unique human opponent runs were recorded with full state information. Through evaluating human runs from two player opponents, closely tied races showed signs of similar risk values and similar changes in risk. Races typically resulted in leveling out at a constant risk threshold. New players often had continuously changing risk (increasing) as opponents learned controls and learned aggressive driving. This indicates risk is a good metric for encouraging diverse behavior and may be useful to better predict future agent actions.

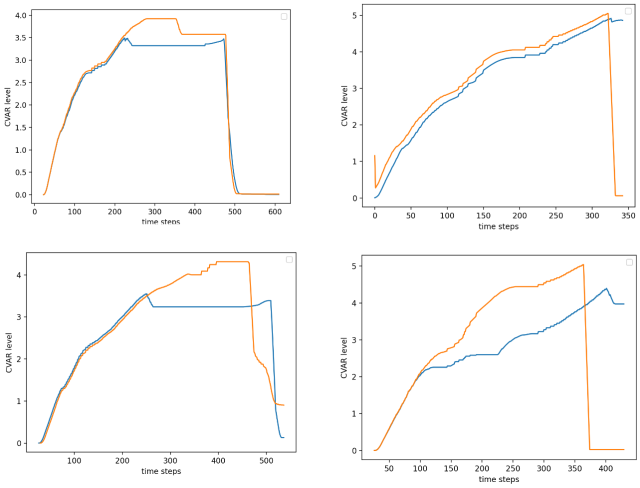

Fig. 5: Examples of the calculated CVaR for a two player human v. human games. New players and continually changing risk level and close matches resulted in similar risk over time profiles. Top Left: Two players with risk levels increasing throughout the game as speed increases until the which level out. Bottom Left: Players risk levels during a match witch start initially close and then separates out. Top Right: Two novice players with continually increasing risk resulting in a very close match. Bottom Right: A match with two new players with increasing risk but the more aggressive agent wins the match.

Initial Training Thus far, an initial policy has been deployed for both human and robot players and trained in simulation. Fine tuning for risk awareness and safety parameters, further curriculum development, and improved rewards are ongoing. Initial training of the neural net Φ had approximately 12% error for calculated risk, 43 % error for adaptation, and 24% error for future state. Further work is being done to explore how to get more accurate adaption parameters and collecting more human trials in order to collect a more accurate future state.

V. Future Work

As this work is in progress, the initial policy will be trained using additional information from human trials and deployed in Phase 1 and 2. The current framework including control input with collisions has been created and adapted in the Trackmania environment in order to allow the policy. Initial training is in progress. An ablation study of observations and risk reward optimization will be performed in addition to studying behavior under competitive behaviors (blocking, edging, and risk change). Finally, the policy will be deployed against real human opponents to evaluate what policy will work.

In studying accuracy of simulated models, human benchmarks to models will compared with analysis on performance between variable risk thresholds and change in risk. Additionally, variable risk metrics will be compared alongside variable methods for risk costs and models of human and robot players. Benchmarks against similar robot MARL techniques will race against the phase 2 policy with visual observation as input. A variety of base policies will be deployed and benchmarked in order to ensure optimal performance.

Lastly, rationality models will be explored to compare performance of simplified single agent reward modeling as opposed to dual agent reasoning. Additional policies will be tested using with fine tuned rewards through hyper parameter tuning with variable definitions of safety.

Acknowledgment

Thank you to Simon Ramstedt and fellow collaborators who created an initial open source work for training reinforcement algorithms in Trackmania.

References

[1] R. Tian, L. Sun, A. V. Bajcsy, M. Tomizuka, and A. D. Dragan, "Safety assurances for human-robot interaction via confidence-aware game-theoretic human models," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11 229–11 235, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238226680

[2] Z. Qin, D. Sun, and C. Fan, "Sablas: Learning safe control for black-box dynamical systems," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1928–1935, 2022.

[3] A. Majumdar and M. Goldstein, "Pac-bayes control: Synthesizing controllers that provably generalize to novel environments," 06 2018.

[4] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, pp. 1–34, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:150521093

[5] J. P. Vasconez, L. Guevara, and F. A. Cheein, "Social robot navigation based on hri non-verbal communication: A case study on avocado harvesting," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 957–960. [Online]. Available: https://doi.org/10.1145/3297280.3297569

[6] L.-Y. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. F. Moura, and M. Veloso, "Teaching robots to predict human motion," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 562–567.

[7] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. M. Dokas, "Safety bounds in human robot interaction: A survey," *Safety Science*, vol. 127, p. 104667, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216294346

[8] A. Liniger and J. Lygeros, "A noncooperative game approach to autonomous racing," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 3, pp. 884–897, 2020.

[9] K. Ji, M. Orsag, and K. Han, "Lane-merging strategy for a self-driving car in dense traffic using the stackelberg game approach," *Electronics*, vol. 10, p. 894, 04 2021.

[10] S. Dennis, F. Petry, and D. Sofge, "Game theory approaches for autonomy," *Frontiers in Physics*, vol. 10, 10 2022.

[11] Y. Jeong, "Stochastic model-predictive control with uncertainty estimation for autonomous driving at uncontrolled intersections," *Applied Sciences*, vol. 11, p. 9397, 10 2021.

[12] A. Majumdar and M. Pavone, *How Should a Robot Assess Risk? Towards an Axiomatic Theory of Risk in Robotics*, 01 2020, pp. 75–84.

[13] J. Bernhard and A. Knoll, "Risk-constrained interactive safety under behavior uncertainty for autonomous driving," 02 2021.

[14] A. Farid, S. Veer, and A. Majumdar, "Task-driven out-of-distribution detection with statistical guarantees for robot learning," 06 2021.

[15] J. Cheng, K. Huang, and Z. Zheng, "Can perturbations help reduce investment risks? risk-aware stock recommendation via split variational adversarial training," 04 2023.

[16] Z. Wang, R. Spica, and M. Schwager, "Game theoretic motion planning for multi-robot racing," 09 2018.

[17] A. Wischnewski, M. Geisslinger, J. Betz, T. Betz, F. Fent, A. Heilmeier, L. Hermansdorfer, T. Herrmann, S. Huch, P. Karle, F. Nobis, L. Ögretmen, M. Rowold, F. Sauerbeck, T. Stahl, R. Trauth, M. Lienkamp, and B. Lohmann, "Indy autonomous challenge – autonomous race cars at the handling limits," 02 2022.

[18] D. Sadigh, N. Landolfi, S. Sastry, S. Seshia, and A. 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, 10 2018.

[19] J. H. Yoo and R. Langari, "A game-theoretic model of human driving and application to discretionary lane-changes," *ArXiv*, vol. abs/2003.09783, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214612208

[20] M. Wang, N. Mehr, A. Gaidon, and M. Schwager, "Game-theoretic planning for risk-aware interactive agents," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 6998–7005.

[21] J. Herman, J. Francis, S. Ganju, B. Chen, A. Koul, A. Gupta, A. Skabelkin, I. Zhukov, M. Kumskoy, and E. Nyberg, "Learn-to-race: A multimodal control environment for autonomous racing," 03 2021.

[22] S. Li, N. Figueroa, A. Shah, and J. Shah, "Provably safe and efficient motion planning with uncertain human dynamics," 07 2021.

[23] J. Chen, J. Li, C. Fan, and B. Williams, "Scalable and safe multi-agent motion planning with nonlinear dynamics and bounded disturbances," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 11 237–11 245, 05 2021.

[24] A. Bajcsy, S. Bansal, E. Bronstein, V. Tolani, and C. J. Tomlin, "An efficient reachability-based framework for provably safe autonomous navigation in unknown environments," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 1758–1765.

[25] J. Fisac, A. Bajcsy, S. Herbert, D. Fridovich-Keil, S. Wang, C. Tomlin, and A. Dragan, "Probabilistically safe robot planning with confidence-based human predictions," 06 2018.

[26] P. Trautman, J. Ma, R. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human-robot cooperation," *The International Journal of Robotics Research*, vol. 34, 02 2015.

[27] R. Thakkar, A. Samyal, D. Fridovich-Keil, Z. Xu, and U. Topcu, "Hierarchical control for multi-agent autonomous racing," 02 2022.

[28] A. Ahmadi Javid, "Entropic value-at-risk: A new coherent risk measure," *Journal of Optimization Theory and Applications*, vol. 155, 12 2012.

[29] D. Ormoneit and R. Neuneier, "Conditional value at risk," 03 1999.

[30] J. Bernhard and A. Knoll, "Risk-constrained interactive safety under behavior uncertainty for autonomous driving," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 63–70.

[31] K. Mokhtari and A. R. Wagner, "Don't get into trouble! risk-aware decision-making for autonomous vehicles," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 1570–1577.

[32] Y. Jia, M. Bhatt, and N. Mehr, "Rapid: Autonomous multi-agent racing using constrained potential dynamic games," in *2023 European Control Conference (ECC)*, 2023, pp. 1–8.

[33] J. Herman, J. Francis, S. Ganju, B. Chen, A. Koul, A. Gupta, A. Skabelkin, I. Zhukov, M. Kumskoy, and E. Nyberg, "Learn-to-race: A multimodal control environment for autonomous racing," 03 2021.

[34] L.-C. Wu, Z. Zhang, S. Haesaert, Z. Ma, and Z. Sun, "Risk-aware reward shaping of reinforcement learning agents for autonomous driving," 06 2023.

[35] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[36] S. Ramstedt and AndrejGobeX, "Tmrl: Trackmania reinforcement learning github," 2021.