Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ34366

Data-Driven Modeling of Driving Behavior via Inverse Reinforcement Learning

Xingyi Wei[1*†], Chenkang Wang[1†], Yiran Xing[2†], Yinjia Zhu[3†], Yiming Pan[4†]

[1]*Maple Leaf International School-Xi'an, Xi'an, China*
[2]*Shanghai HWS school, Shanghai, China*
[3]*Changwai Bilingual School, Changzhou, China*
[4]*Shanghai HD school, Shanghai, China*
[*]*Corresponding Author. Email:3246552807@qq.com*
[†]*These authors contributed equally to this work and should be considered as co-first authors*

Abstract. In recent years, the rapid development of intelligent transportation systems (ITS) and autonomous driving has made human driving behavior modeling accurate and critical for improving traffic safety, efficiency, and autonomous system adaptability. Traditional rule-based or utility-centric models, however, fail to handle the complexity, randomness, and scenario dependence of real driving. Thus, our study aims to explore inverse reinforcement learning (IRL), a data-driven method, for driving behavior modeling. We first reviewed major IRL variants such as Maximum Margin IRL, MaxEnt Deep IRL, GAIL, Bayesian IRL and IAL and analyze their strengths like MaxEnt Deep IRL's adaptability to large state space and limitations in ITS. We then proposed two frameworks: 1) an AOAT strategy based on MaxEnt IRL which uses HighD data set and reduces lateral deviations by 42.91%-55.35% vs fixed-weight schemes; 2) a multi-agent framework integrating multi-modal data, using Bradley-Terry regression and PPO algorithm for real-time traffic signal optimization. Finally, we discussed IRL's challenges such as data set reliance, poor reward function interpretability, high computation and cost and proposed future directions including standardized datasets, hybrid reward structures and algorithm optimization. This study proves IRL's value for human-centric modeling, laying a foundation for safer, more adaptable ITS.

Keywords: Intelligent Transportation Systems (ITS), Inverse Reinforcement Learning (IRL), Driving Behavior Modeling, Anthropomorphic Obstacle Avoidance Trajectory, Multi-Agent Traffic Optimization

1. Introduction

Over the last ten years, intelligent transportation systems (ITS) have turned into a big talking point in research, with the main goal of making getting around safer, more efficient, and better for the planet. As self-driving tech keeps advancing fast, it's become more important than ever to get a handle on how people really drive and build spot-on models that capture that behavior. Old-school methods—like systems that follow strict rules and models centered on real-world use—have given

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ34366

us some helpful insights, but they often struggle when it comes to the messiness, unpredictability, and situation-specific side of actual driving. To tackle these problems, inverse reinforcement learning (IRL), which uses real-life data, has been getting more and more notice. IRL digs up the hidden "payoffs" behind the driving actions we can see, letting us figure out things like how much drivers care about staying safe, what makes them feel comfy, and how much risk they're okay with taking. This gives us a more structured way to mimic the decisions people make when they're behind the wheel.

IRL has also been put to work successfully in areas like lane changes, following other vehicles, and interactions between several cars at once [1-3]. These wins show that IRL can bridge two big parts of transportation research: guessing how people will drive and figuring out why they drive that way.

But even with this headway, there are still major hurdles. These issues don't just come from the limits of today's IRL algorithms—they also have to do with how well these algorithms work in real traffic. Current IRL studies vary a great deal in three key spots: the starting assumptions they use, the choices made when building the algorithms, and the ways they're tested and judged.

This makes it hard to reach conclusions that are consistent or can be applied widely. While some IRL methods—such as maximum entropy IRL, deep IRL, and Bayesian IRL—have proven useful, we still don't fully grasp how their pros and cons measure up in different traffic situations [4, 5]. What's more, IRL depends a lot on large, high-quality sets of data about driving paths. This reliance causes problems: data might be scarce, have mistaken, or produce results that don't work well when moved from one scenario to another [6].

Interpretability is another issue. Even though IRL can reveal the hidden intentions of drivers, the "reward functions" it learns (which describe what drives decisions) are often abstract. This makes it hard to turn them into traffic rules or ideas that people can easily understand [7, 8]. Additionally, IRL doesn't fully incorporate knowledge specific to transportation—such as traffic laws, cultural habits related to driving, or how people perceive risk. These further limits how it can be used in real-world settings [5].

The future of IRL in transportation research will depend on solving these key challenges and making its future development path clearer. Despite the growing diversity of IRL methods, universal issues in efficiency, generalizability, and interpretability—especially pronounced in transportation contexts—persist. Progress will likely rely on methodological innovation, enhanced interpretability, and deeper integration of domain-specific knowledge. Only with such advances can IRL effectively support behavior-aware mobility systems and human-centered autonomous driving. Beyond theoretical contributions, this research agenda holds significant practical importance: a robust, real-world IRL framework could underpin tools for safety evaluation of autonomous vehicles, policy testing, and adaptive control. Ultimately, such developments would promote more trustworthy and socially inclusive intelligent transportation systems.

The remainder of this article is structured as follows. First, it provides a comparative review of major IRL methods, highlighting their advantages, drawbacks, and applicability in traffic environments. Then it examines how real-world driving scenarios can uncover heterogeneity in driving factors and behavioral preferences, with implications for both predictive modeling and policy design. Next, current applications and evaluation strategies are analyzed, identifying persistent limitations such as data dependency and lack of universality. Finally, the article outlines future research directions and presents a developmental roadmap for IRL in intelligent transportation, offering actionable insights for both researchers and practitioners.

DOI: 10.54254/2755-2721/2026.TJ34366

2. Literature review

2.1. The model of Inverse Reinforcement Leaning

Inverse reinforcement learning (IRL), as a method for inferring latent reward functions from expert demonstrations, has shown broad application prospects in recent years in fields such as robotics, autonomous driving, and intelligent decision-making. Its research methods are diverse, including Maximum Margin Inverse Reinforcement Learning (Maximum Margin IRL), Maximum Entropy Deep Inverse Reinforcement Learning (MaxEnt Deep IRL), Generative Adversarial Imitation Learning (GAIL), Bayesian Inverse Reinforcement Learning (Bayesian IRL), and Inverse Apprenticeship Learning (IAL).

2.2 Maximum margin Inverse Reinforcement Learning

Maximum Margin Inverse Reinforcement Learning (Maximum Margin IRL) is an inverse reinforcement learning method based on structured prediction and the principle of margin maximization. The core idea of this method is to find a reward function under which the performance of the expert policy is significantly better than any other policy, thereby creating a "maximum margin" distinction.

This method typically assumes that expert behavior is "optimal" in some sense and seeks a reward function through an optimization process that maximizes the gap between the cumulative reward of the expert policy and other policies. This not only helps the learner better imitate expert behavior but also maintains robustness in the presence of noise or suboptimal demonstrations.

2.3 Maximum entropy deep Inverse Reinforcement Learning

Maximum Entropy Deep Inverse Reinforcement Learning (MaxEnt Deep IRL) is a methodological approach that extends the principles of Maximum Entropy Inverse Reinforcement Learning, forming a structured framework for learning from demonstrations.

This method utilizes neural networks to model reward structures and, by employing a differentiable objective function, facilitates the efficient training of deep architectures, thereby eliminating the need for manual tuning of reward parameters.

This framework is particularly well-suited for tackling complex tasks in extensive state spaces, such as robotic manipulation processes and autonomous driving scenarios. Additionally, this solution features high adaptability: customized neural networks can be adjusted to meet various task requirements while leveraging the same training cost function to optimize performance.

In performance evaluation benchmarks, MaxEnt Deep IRL delivers results comparable to state-of-the-art methods. Although it requires a larger amount of training data, its structural complexity is independent of the quantity of demonstration samples. This trait makes it highly suitable for robotic lifelong learning contexts—where intelligent systems continuously evolve as they acquire new data [9, 10].

Then: Generative Adversarial Imitation Learning (GAIL), an advanced state of-the-art technical approach which naturally merges Generative Adversarial Networks (GANs) basics with those from the Inverse Reinforcement Learning (IRL) domain. Application of the technique involves training two linked models, namely the generator whose primary function is to simulate and replicate the behavior of expert agents while the other model (i.e. the discriminator) is developed to differentiate

DOI: 10.54254/2755-2721/2026.TJ34366

the real behavior data collected from the experts against the imitated behavior output of a learning agent.

When the generator evolves to imitate expert more closely, then the discriminator evolves to perceive discrepancies. The contest is directed toward the generator's ability to imitate experts perfectly. Consequently, the learning is not merely copying the expert's actions, but generalize expert's actions to unseen situations as well [11].

2.4 Bayesian Inverse Reinforcement Learning

Bayesian Inverse Reinforcement Learning (Bayesian IRL), which introduces probabilistic reasoning into the process [11]. Bayesian methods generally integrate prior knowledge with data or evidence. Bayesian IRL offers three key advantages: first, an optimal policy is not required. Second, it avoids assuming infallible decision-making by experts and explicitly accounts for uncertainties in expert behavior. Third, like other Bayesian methods, it can incorporate external information about the problem into the reward function via the prior distribution. Thus, instead of guessing the expert's exact motives, it can model the probability of different motives being the driving force behind the behavior. This is especially useful when dealing with noisy or incomplete data [12].

2.5 Inverse apprenticeship learning

Inverse Apprenticeship Learning (IAL), a hybrid approach. It builds on Inverse Reinforcement Learning (IRL) by focusing on mimicking expert behavior over time. It not only infers a reward function but also makes the learner's behavior closely match that of the expert. You can think of it as "learning on the job." In this method, the learner observes the expert and gradually adjusts its behavior to approximate the expert's performance as closely as possible over time—often without a fully accurate reward function.

Although inverse reinforcement learning continues to advance methodologically, it still faces several core challenges. IRL infers reward functions from expert demonstration data, but the sample efficiency problem is particularly prominent. Traditional IRL algorithms require large amounts of high-quality expert trajectory data, yet obtaining such data in real-world scenarios is extremely costly, and the generalization of the learned reward functions is limited.

2.6 The limitations of inverse reinforcement

The contradiction between the interpretability and generalizability of reward functions is another major challenge. Reward functions learned by IRL often overfit the specific behavioral patterns in the expert data, making it difficult to adapt to dynamic changes in the environment. For example, in a medical decision-making scenario, a reward function learned from expert data at one hospital might fail at another hospital due to differences in patient populations. Although existing methods improve generalizability by introducing meta-learning frameworks or latent variable models, the computational complexity increases significantly.

Simultaneously, reliance on the quality of expert data is also a significant issue. If expert data contains systematic biases (e.g., driving habits specific to a cultural context), IRL can amplify these biases. Furthermore, optimal strategies may differ among experts, leading to instability in reward function learning.

Multi-task transfer obstacles also pose a difficulty. Currently, in cross-task knowledge transfer, the structure of reward functions varies greatly between different tasks, leading to difficulties in

DOI: 10.54254/2755-2721/2026.TJ34366

representation sharing; during transfer, old and new policies might interfere with each other, causing policy interference that affects the agent's judgment. Although recent research attempts to construct abstract reward representations shared between tasks through hierarchical inverse reinforcement learning, the granularity selection for time abstraction and state abstraction still lacks theoretical guidance.

In summary, inverse reinforcement learning must still address a series of core problems in methodological innovation and application promotion, including generalizability, efficiency, reliability, and transferability.

3.Methodlogy

3.1 Driving factors and behavioral preferences

Naturalistic driving data serve as crucial and indispensable resources for learning about and understanding driver behavior. Additionally, such data play a key role in extracting and generating realistic simulation scenarios—these scenarios can either be directly derived from the data or randomly generated by meeting the statistical criteria obtained from it. Another vital application of naturalistic driving data lies in analyzing human drivers' behaviors: this involves figuring out how drivers operate vehicles on real-world roads and how they respond to various road events

Driving often involves four key steps: surrounding monitoring, predicting, decision making and maneuver executing. Many factors could affect these steps and influencing a driver's driving performance and safety [13].

Driving often involves car-following behavior—for instance, during rush-hour highway driving, drivers adjust their vehicle's speed and distance by balancing the desire to shorten travel time and prioritize safety. Thus, car-following models must be developed to improve traffic safety. Early driving behavior models focused on car-following, describing how a vehicle (follower) responds to the vehicle ahead (leader), with the follower assumed to react to the leader's actions. Recently, microscopic traffic simulation models have driven the development of general acceleration models —these capture not only drivers who closely follow leaders but also those who do not—and sparked interest in lane-changing behavior.

General acceleration models define multiple driving regimes (e.g., free-flow, emergency, and various car-following types like acceleration/deceleration or reactive/non-reactive) and assume distinct behaviors for each. Lane-changing models usually have two components: decision-making and execution. However, existing classifications struggle to balance the two, leading to rigid behavioral structures—for example, overtaking in MLC scenarios is difficult.

However, although the aforementioned methods can identify and model driving styles to a certain extent, they have a significant limitation: even among drivers of the same style, there are notable differences in operational details, risk perception, and decision-making preferences. If these differences are overlooked, the control system will fail to fully replicate the natural driving behavior of individuals, leading to deviations.

3.2 Strategy for anthropomorphic obstacle avoidance trajectory (AOAT) in adaptive driving scenarios

In real-world driving scenarios, factors such as unpredictable behavior of traffic participants, complex driving environments, and differences in driver styles pose significant challenges to the adaptability and acceptability of trajectory planning.

DOI: 10.54254/2755-2721/2026.TJ34366

To address this issue, WU et al., considering both the acceptability and adaptability of obstacle avoidance trajectory planning strategies, proposed an Anthropomorphic Obstacle Avoidance Trajectory (AOAT) planning scheme for adaptive driving scenarios based on Maximum Entropy Inverse Reinforcement Learning.

As shown in Figure 1, the strategy adopted by AOAT is divided into two phases: offline training and online optimization. During the offline training phase, a large number of expert obstacle avoidance trajectories and corresponding driving scene information are first extracted from the HighD natural driving dataset. Subsequently, Maximum Entropy Inverse Reinforcement Learning technology is used to extract the feature weights of the trajectory optimization function from the EDOAT (Note: Likely means expert demonstration obstacle avoidance trajectories) data. Next, key driving scene information affecting the obstacle avoidance trajectory is extracted, and a mapping model between the driving scene and the trajectory feature weights is constructed. Finally, the parameters w of this model is obtained through a multivariate nonlinear fitting method. During the online optimization phase, the optimization function for the obstacle avoidance trajectory is reconstructed based on the driving scene information E and the mapping model obtained in the offline training phase.

Regarding data selection, because the HighD dataset was collected by drones over typical road sections, primarily including two typical scenarios: ordinary highways and ramp entrances, and this model focuses on the obstacle avoidance behavior characteristics of vehicles on regular straight road sections, free lane-changing trajectory data from unobstructed road sections ahead and driving trajectory data from ramp entrances were eliminated during the trajectory extraction process. Furthermore, to improve the effectiveness of the designed obstacle avoidance trajectories, Wu et al. discarded trajectory data where surrounding vehicles were present but outside the capture field of view, as well as trajectory data where surrounding vehicles left the capture field of view before the obstacle avoidance maneuver was completed [14].

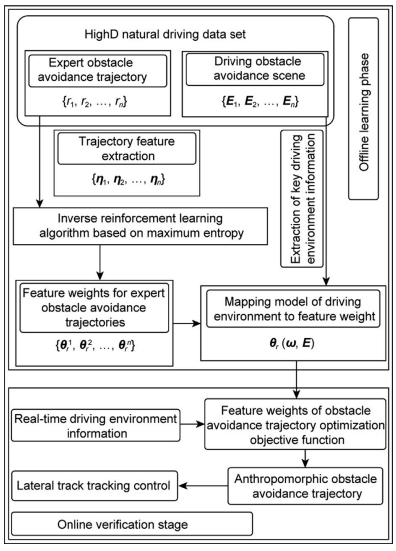

Figure 1. AOAT planning strategy for adaptive driving scenarios

The driving scene-weight mapping constructed through multivariate nonlinear fitting demonstrated excellent performance. The fitting evaluation metrics showed that the coefficients of determination R^2 for weights θ_1 and θ_2 reached 0.9853 and 0.9424, respectively, with root mean square errors at a low level, indicating that the model can accurately capture the nonlinear relationship between dynamic scene features and weight parameters; Comparison results in 42 test scenarios showed that the trajectories generated by dynamically adjusting weights using this

DOI: 10.54254/2755-2721/2026.TJ34366

mapping model had significantly smaller differences from expert demonstration trajectories in terms of lateral position, lateral velocity, and lateral acceleration characteristics compared to the fixed weight scheme. The difference magnitudes were reduced to 49.04%, 42.91%, and 55.35% of those of Scheme 1, respectively, fully demonstrating the effectiveness and superiority of this model in enhancing trajectory anthropomorphism and scene adaptability [14].

3.3 Global optimization of multi-agent, multi-objective collaboration

The aforementioned research on trajectory optimization for vehicle obstacle avoidance demonstrates the significant potential of data-driven methods in understanding and replicating human driving behavior. However, the challenges of intelligent transportation systems are not solely concentrated at the single-vehicle intelligence level; they also involve considering global optimal scheduling for multiple agents and multiple objectives across the entire road network, which is a major bottleneck in our current urban traffic management. Inspiration from the alignment of IRL and Large Language Models (LLM) [15, 16] gave the Sun team an idea: could they try to use a multi-objective reward function, inferred from multimodal traffic data and historical expert decisions, to train an agent that can generate optimal adaptive signal control strategies for intersections in real time, thereby achieving the best traffic throughput, safety, and sustainability for the entire road network.

To better solve this problem, Sun et al. designed a data-driven IRL integration framework. This method first constructs a dense perception network covering the city's main intersections, integrating multimodal perception data such as high-definition cameras, radars, inductive loops, and environmental sensors to form a high-dimensional traffic state representation, while simultaneously collecting human expert intervention records from traffic management centers as expert demonstration data [17, 18]. Next, Sun et al. used a deep reward network based on the Bradley-Terry regression model [19, 17] to establish a multi-objective reward function from these demonstration data. This function can simultaneously encode dimensions such as traffic efficiency, safety, emission reduction, and emergency priority. This step crucially translates implicit human preferences into optimizable quantitative metrics.

Subsequently, based on the learned reward function, they used the PPO algorithm [20] to train a decision generator (i.e., a policy network) using a Transformer architecture. This network serves as a core component of the system: it receives real-time multimodal traffic flow states as input, considers the interactions between multiple approaches, computes several feasible future phase timing plans for each intersection, and performs long-term benefit prediction [21]. Furthermore, a reward-guided decoding mechanism was proposed to further enhance decision-making performance [22]: each generation unit is scored instantly, and the optimal decoding path is selected during the inference process, providing stronger resistance to interference in complex scenes. This step is also integrated into the online learning process, allowing for continuous fine-tuning and optimization based on data samples obtained from each experiment, thus continually improving itself [23].

The results from simulations and real road sections presented by Sun et al. show that the multi-objective collaborative optimization significantly outperforms existing traffic systems and can effectively improve traffic efficiency. This research proposes a new paradigm for achieving the goals of next-generation intelligent transportation systems.

DOI: 10.54254/2755-2721/2026.TJ34366

4. Discussion and conclusion

4.1 Discussion

The outcomes of this research bring to light both the promise of using inverse reinforcement learning (IRL) in intelligent transportation and the lasting challenges it encounters. Through an examination of prior research, real-world driving data, and existing modeling approaches, several important insights emerge.

Firstly, the application of IRL demonstrates a significant potential in capturing the heterogeneity of driver behavior. Traditional models usually oversimplify human decision-making by assuming a uniform behavioral pattern, while real life provides a framework to reveal the potential reward structures that reflect different motivations such as safety awareness, comfort, and risk tolerance. This advantage is particularly evident in trajectory planning, where an anthropomorphic model trained on natural driving data can generate obstacle avoidance behaviors that are more in line with human habits. However, the effectiveness of these methods is closely related to the availability and quality of the data sets used. As observed in our method, reliance on resources such as the HighD dataset introduces issues of scalability, privacy, and context coverage, limiting the model's generalization across different driving environments.

Secondly, although the latest developments such as maximum entropy real-life, deep real-life, and adversarial imitation learning have expanded the methodological toolkit, interpretability remains a core issue. The reward functions learned through these methods are typically abstract and difficult to translate into explicit traffic policies or driver training guidelines. This limitation reduces the practical value of systems based on real-world environments in areas that require transparency and accountability, such as regulatory decisions and the public adoption of autonomous driving technologies. Closing this divide might call for a mixed strategy that brings real-world practice together with field expertise, traffic rules, and research into human behavior—all to make the end results easier to understand and put into action.

Third, when we talk about the experiment results, there's an obvious conflict between how well something adapts and how feasible it is in terms of computing. On one hand, the research findings show that this real-scene-based learning method can roughly cut the trajectory deviations of human drivers by half. And with this reduction, these technologies naturally become more acceptable, and users' trust in them also goes up. On the other hand, the huge computing demands arising during offline training and online planning have become a major roadblock, preventing these technologies from being put to use in real-time scenarios. In traffic environments, this problem is particularly prominent—road conditions change quickly and risks are extremely high. In such cases, both drivers and autonomous driving systems have to make quick decisions to ensure everyone's safety. To break through these real-time limitations, two key things need to be done in the future: first, make the learning algorithms more efficient, and second, enhance the computing power at the edge. Even so, challenges remain: issues with depending on data, being able to explain how things work, applying findings broadly, and computing costs all need to be fixed before IRL can go from being a promising research area to something that can be reliably used on a large scale.

But whether it can change the field will depend on how well researchers deal with these current shortcomings. Future progress will probably focus on three areas: (1) building standard datasets that protect users' privacy and cover different driving environments; (2) designing reward systems that make sense and fit with the field's expert knowledge; (3) making real-time planning more efficient in terms of computing. Fixing these problems won't only make IRL models more dependable—they'll also help these models mesh better with self-driving systems that prioritize human needs, as

DOI: 10.54254/2755-2721/2026.TJ34366

well as with how policies get made. In the end, this will help create a transportation system that's safer, more flexible, and trusted by everyone.

4.2 Conclusion

This study explores the application of inverse reinforcement learning (IRL) in intelligent transportation systems (ITS), focusing on addressing challenges in human driving behavior modeling and obstacle avoidance trajectory planning. Key conclusions are as follows: Methodologically, diverse IRL approaches exhibit distinct strengths in ITS scenarios—Maximum Margin IRL ensures robustness in distinguishing expert policies, MaxEnt Deep IRL adapts to complex large-state-space tasks, GAIL generalizes well to unknown scenarios, and Bayesian IRL handles expert behavior uncertainty. Yet, universal issues like sample inefficiency (reliance on costly large-scale high-quality trajectory data) remain unresolved, with no single method able to tackle all core problems. In practical application, the proposed anthropomorphic obstacle avoidance trajectory (AOAT) planning strategy achieves targeted progress. By using IRL for offline learning of expert-driven obstacle avoidance trajectories and a trajectory expectation feature-matching algorithm, it avoids cumbersome manual parameter tuning; meanwhile, taking ego vehicle speed and speed difference with the lead vehicle as core scenario parameters enhances trajectory adaptability to dynamic traffic, reducing deviations from human driving habits and improving ride comfort and safety. However, IRL's large-scale application in ITS still faces bottlenecks: overreliance on specific datasets (e.g., HighD) limits model generalization across different traffic contexts, abstract reward functions lack interpretability (hindering translation into actionable policies), and computational inefficiency restricts real-time deployment in dynamic high-risk scenarios. Future efforts should focus on three directions: building standardized, privacy-preserving multi-scenario datasets to boost generalization; developing hybrid reward structures integrating domain knowledge to enhance interpretability; and optimizing algorithm efficiency via lightweight network design and edge computing to enable real-time planning. In all, IRL offers a robust framework for human-centric ITS development, and addressing its current limitations will drive the creation of safer, more adaptable transportation systems.

Acknowledgement

Xingyi Wei, Chenkang Wang, Yiming Pan, Yiran Xing, Yinjia Zhu contributed equally to this work and should be considered co-first authors.

References

[1] Liu, S., Li, X., Chen, J., Guo, C., Wu, J., Luo, Q. and Ma, H. (2025). Individualized Driving Intention Prediction With Inverse Reinforcement Learning. IEEE Transactions on Intelligent Transportation Systems, pp.1–15. doi: https: //doi.org/10.1109/tits.2025.3543553.

[2] Geng, M., Cai, Z., Zhu, Y., Chen, X. and Lee, D.-H. (2023). Multimodal Vehicular Trajectory Prediction With Inverse Reinforcement Learning and Risk Aversion at Urban Unsignalized Intersections. IEEE Transactions on Intelligent Transportation Systems, pp.1–14. doi: https: //doi.org/10.1109/tits.2023.3285891.

[3] Ghoul, T. and Sayed, T. (2021). Real-Time Safety Optimization of Connected Vehicle Trajectories Using Reinforcement Learning. Sensors, 21(11), p.3864. doi: https: //doi.org/10.3390/s21113864.

[4] Bhattacharyya, R.P., Wulfe, B., Phillips, D.L., Kuefler, A., Morton, J.R., Ransalu Senanayake and Kochenderfer, M.J. (2023). Modeling Human Driving Behavior Through Generative Adversarial Imitation Learning. 24(3), pp.2874–2887. doi: https: //doi.org/10.1109/tits.2022.3227738.

DOI: 10.54254/2755-2721/2026.TJ34366

[5] Deshpande, S., Rahee Walambe, Kotecha, K., Ganeshsree Selvachandran and Abraham, A. (2025). Advances and applications in inverse reinforcement learning: a comprehensive review. Neural Computing and Applications. doi: https: //doi.org/10.1007/s00521-025-11100-0.
[6] Abdel Madjid, A., Fergani, B. and Saidouni, D. (2016). Trajectory Prediction for Autonomous Driving: Progress, Limitations, and Future Directions. [online] Arxiv.org. Available at: https: //arxiv.org/html/2503.03262v1.
[7] Morton, J. and Kochenderfer, M.J. (2017). Simultaneous policy learning and latent state inference for imitating driver behavior. arXiv (Cornell University). doi: https: //doi.org/10.1109/itsc.2017.8317738.
[8] Arora, S. and Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. Artificial Intelligence, 297, p.103500. doi: https: //doi.org/10.1016/j.artint.2021.103500.
[9] Wulfmeier, Markus, et al. "Maximum Entropy Deep Inverse Reinforcement Learning." ArXiv.org, 11 Mar. 2016, arxiv.org/abs/1507.04888. Accessed 29 Sept. 2023.
[10] Maximum Entropy Inverse Reinforcement Learning Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213bziebart@cs.cmu.edu, amaas@andrew.cmu.edu, dbagnell@ri.cmu.edu, anind@cs.cmu.edu
[11] Wang, Wenshuo, et al. "Modeling and Recognizing Driver Behavior Based on Driving Data: A Survey." Mathematical Problems in Engineering, vol. 2014, 2014, pp. 1–20, https: //doi.org/10.1155/2014/245641.
[12] Yadav, Amit. "Inverse Reinforcement Learning for Human-Behavior Modeling." Medium, Biased-Algorithms, 7 Oct. 2024, medium.com/biased-algorithms/inverse-reinforcement-learning-for-human-behavior-modeling-f0f383358763.7
[13] "Modeling and Recognizing Driver Behavior Based on Driving Data: A Survey." Mathematical Problems in Engineering, vol. 2014, 2014, pp. 1–20, https: //doi.org/10.1155/2014/245641. priority_high
[14] Wu, J., Yan, Y., Liu, Y., & Liu, Y. (2024). Research on Anthropomorphic Obstacle Avoidance Trajectory Planning for Adaptive Driving Scenarios Based on Inverse Reinforcement Learning Theory. Engineering, *33*, 133–145. https: //doi.org/10.1016/j.eng.2023.07.018
[15] Sun, H., & van der Schaar, M. (2024). Inverse-alignment: Inverse reinforcement learning from demonstrations for llm alignment. arXiv preprint arXiv: 2405.15624.
[16] Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.
[17] Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. Advances in Neural Information Processing Systems, 29.
[18] Springall, A. (1973). Response surface fitting using a generalization of the bradley-terry paired comparison model. Journal of the Royal Statistical Society Series C: Applied Statistics, 22(1), 59-68.
[19] Sun, H., et al. (2024b). Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. arXiv preprint arXiv: 2411.04991.
[20] Schulman, J., et al. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347.
[21] Chan, A. J., et al. (2024). Dense reward for free in reinforcement learning from human feedback. arXiv preprint arXiv: 2402.00782.
[22] Deng, H., & Raffel, C. (2023). Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. arXiv preprint arXiv: 2310.09520.
[23] Xiong, W., et al. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. arXiv preprint arXiv: 2312.11456.