Optimizing Urban Route Choice for Autonomous Vehicles using Multi-Agent Reinforcement Learning

Doctoral Consortium

Anastasia Psarou
Jagiellonian University
Krakow, Poland
anastasia.psarou@doctoral.uj.edu.pl

ABSTRACT

Autonomous vehicles (AVs) have already been introduced in some cities worldwide, and understanding how they could effectively learn optimal routing strategies, that is, routes to travel from an origin to a destination point in a traffic network, is essential. In my Ph.D., I use Multi-Agent Reinforcement Learning (MARL) to model AV routing decisions in a microscopic setting, shared with human drivers, where AVs learn to select routes that minimize their costs (e.g., travel time) given the currently observed state of the traffic network. This paper provides a brief overview of my work and focuses on two main contributions. First, we show that when multiple independent learning AVs simultaneously learn routing strategies in a traffic network, they may destabilize it by increasing human and AV travel times, as the state-of-the-art MARL algorithms used to train their routing decisions require long training iterations to converge to the optimal solution. Second, we study a solution to this problem by introducing a social component into the AV's reward functions, building on prior work on socially aware incentives in multi-agent systems. We show that this can accelerate convergence to the system-optimal solution and benefit individual agents in this routing game.

KEYWORDS

Autonomous vehicles, Multi-agent reinforcement learning, Route choice

ACM Reference Format:
Anastasia Psarou. 2026. Optimizing Urban Route Choice for Autonomous Vehicles using Multi-Agent Reinforcement Learning: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/JTZD1181

1 INTRODUCTION

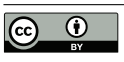
Automation has advanced to the point where Autonomous Vehicles (AVs) are being deployed in some cities around the world, such as San Francisco [4]. As a result, routing decisions that were traditionally made by selfish human drivers [20] are likely to be determined by learning algorithms. Multi-Agent Reinforcement Learning (MARL) could be a suitable method for modeling optimal routing strategies of AVs, as it can scale to large numbers of agents and routes (actions), unlike game-theoretic equilibrium solvers [19]. Additionally, with MARL, we can study the problem in a microscopic setting, unlike Operations Research (OR) approaches such as Frank-Wolfe [7], which take a macroscopic perspective by modeling aggregate traffic flows rather than treating each agent as an independent decision-maker.

A standardized framework to facilitate the study of the routing decisions of AVs and human drivers in shared traffic environments was lacking in the literature. To address this gap, we contributed *RouteRL* [2], a MARL framework for modeling and simulating the collective routing decisions of human-driven and autonomous vehicles. In *RouteRL*, AVs are modeled as reinforcement learning (RL) agents, and human drivers learn routing strategies using state-of-the-art behavioral human learning models [8]. RouteRL is formulated as an Agent Environment Cycle (AEC) game [18] in which agents act sequentially according to their departure times. Once an agent selects an action (route), the selected route is simulated in SUMO, a microscopic traffic simulator, [10], to obtain the agent's realized travel time, which typically serves as the reward to be minimized. By modeling congestion at the microscopic level, SUMO enables the study of interactions among agents under realistic traffic conditions.

In my PhD, I have been using *RouteRL* to study how AV routing decisions affect human and AV travel times. First, in [11], we showed that when multiple AV agents are introduced into a traffic network and learn routing strategies simultaneously, they require many training iterations to converge to the optimal solution, thereby negatively affecting overall system travel times. Subsequently, in [12], we showed that incorporating a social component into the AV's reward function, which accounts for the AV's impact on other agents' travel times, can be beneficial not only for the system-wide performance but also for individual AV agents. I discuss these contributions in Sections 2 and 3, respectively.

2 COLLABORATION BETWEEN CITY AND ML COMMUNITY FOR EFFICIENT AV ROUTING

In this study [11], we present evidence highlighting the importance of collaboration between city authorities and the ML community for the evaluation and monitoring of AI-based routing algorithms deployed by car companies. At the same time, continued efforts by the ML community are necessary to improve the robustness, fairness, and practical applicability of current algorithms. Without such collaboration between city authorities and the ML community, autonomous routing decisions could worsen congestion as AV penetration rates increase in urban areas.



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JTZD1181

To support our claim, we initially study the routing decisions of AVs and human drivers in a simple network with two routes (two actions) that intersect, creating a priority scheme in which vehicles on the shorter route give way to those on the longer route. We assume that, initially, only human drivers are included in the system, modeled as utility maximizers, and that, after some learning iterations, their collective routing behavior converges to a Wardrop equilibrium [20]. We then extend the scenario to a mixed-traffic setting, in which a subset of drivers travel using AVs. AVs use various state-of-the-art MARL algorithms, including the Independent Deep Q-Learning (IDQN) [17], Independent Proximal Policy Optimization (IPPO) [5], Independent Soft Actor Critic (ISAC) [9], Multi-Agent Soft Actor Critic (MASAC), Multi-Agent PPO (MAPPO) [21], Value Decomposition Networks (VDN) [15], QMIX [13] to learn optimal routing strategies. Our results indicate that some algorithms do not find the optimal solution, while others require many training iterations, corresponding to several years of real-world commuting, to converge to the optimal solution.

To more accurately model real-world conditions, we introduce non-determinism into the simple network's traffic flow and demand by varying agents' departure times. Additionally, we enable human drivers to adapt their decisions in response to the disequilibrated system, whereas previously we assumed their choices were fixed during AV training. We further study the problem in the real-world traffic network of Ingolstadt [1, 3]. Our results show that, even after extensive training, agents in the simple network fail to converge to the optimal solution when non-determinism is introduced. Additionally, during MARL training in the Ingolstadt network, agents do not converge to the optimal solution even after many episodes, and the system is destabilized (as evidenced by the variability in travel times).

3 SOCIAL AWARENESS OF AV ROUTING

In subsequent work [12], we address the convergence issue discussed in Section 2 by showing that incorporating a social component into the selfish travel-time-based reward of the MARL-enabled AV agents can achieve faster convergence without altering the equilibrium solutions, which are preserved. Specifically, we complement the selfish reward with a counterfactual based on the marginal cost that each AV imposes on other drivers in the system. The marginal cost is defined as the sum of the travel-time differences experienced by all agents in the system when agent i is present in the simulation, relative to when that agent is removed. We demonstrate the effectiveness of this new reward formulation in the simple traffic network discussed in Section 2, which includes two alternative routes (discrete actions) for each AV agent to choose between. The IDQN, MAPPO, and Upper Confidence Bound (UCB) algorithms [16] are used to train the routing decisions of the AV agents under the proposed reward formulation, resulting in faster convergence to both the System Optimal (SO) and individually optimal solutions, which coincide in this network.

Subsequently, to support our claim that our new reward formulation achieves faster convergence to the SO solution and is also beneficial for the individual AV agents in a routing game, we apply it to AV agents routing in the real-world traffic network of Saint-Arnoult from the URB benchmark [1], where the SO and individually optimal solutions do not coincide. Using the UCB algorithm to train AV routing decisions, we show that the total system and total AV group travel times are reduced when we introduce the marginal cost into the AV agents' rewards. Additionally, more than 50% of the AV agents achieve shorter individual travel times when trained with the marginal-cost-based reward. This result emphasizes that incorporating socially oriented behavior into AV routing decisions can be efficient for improving both individual and system-level performance in future transportation systems.

To compute the marginal cost, additional simulation runs are performed, each excluding a single AV agent. As a result, each iteration requires a number of additional simulations equal to the number of AV agents, leading to computational overhead as the AV population grows. Consequently, a key limitation of this work is the cost of computing the marginal cost.

4 FUTURE WORK

In future work, I plan to address the computational challenge discussed in Section 3 by reducing the number of simulation runs required to compute marginal costs. One possible direction is to estimate the marginal cost rewards only for agents whose decisions are likely to affect many others in the system, rather than computing them for every AV agent. This could lead to a more efficient and scalable approach to training AV routing policies with MARL.

Complementing my work on efficient AV routing, I am currently studying resource (route) allocation through *Karma* [6], a non-monetary mechanism for efficient and equitable resource allocation. In *Karma* economies, individuals compete for scarce resources by participating in auctions and bidding with an artificial currency. Prior work has used *Karma* to alleviate congestion by pricing specific road segments [14]. They showed that *Karma* can achieve system travel times comparable to those achieved by monetary pricing schemes while being fairer, since monetary pricing advantages wealthier individuals who can afford to pay. By contrast, *Karma* allocates resources based on urgency without discriminating based on income. However, bidding in *Karma* economies can be challenging. Hence, I am currently investigating learning-based bidding agents for allocating routes in Karma economies under realistic traffic conditions, with the aim of assessing whether, in this setting, they can achieve efficient and more equitable outcomes than monetary pricing.

ACKNOWLEDGMENTS

My PhD research is financed by the European Union within the Horizon Europe Framework Programme (ERC Starting Grant CO-eXISTENCE no. 101075838).

REFERENCES

[1] Ahmet Onur Akman, Anastasia Psarou, Michał Hoffmann, Łukasz Gorczyca, Łukasz Kowalski, Paweł Gora, Grzegorz Jamróz, and Rafał Kucharski. 2025. URB – Urban Routing Benchmark for RL-equipped Connected Autonomous Vehicles. In *Advances in Neural Information Processing Systems*.

[2] Ahmet Onur Akman*, Anastasia Psarou*, Łukasz Gorczyca, Zoltán György Varga, Grzegorz Jamróz, and Rafał Kucharski. 2025. RouteRL: Multi-agent reinforcement learning framework for urban route choice with autonomous vehicles. *SoftwareX* 31 (2025), 102279. https://doi.org/10.1016/j.softx.2025.102279

[3] James Ault and Guni Sharon. 2021. Reinforcement learning benchmarks for traffic signal control. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[4] Evelyn Cheng. 2025. *Robotaxis in 2025: Waymo plots global expansion as Zoox, Tesla roll to the starting line.* https://www.cnbc.com/2025/12/16/waymo-amazon-zoox-tesla-robotaxi-expansion.html Accessed: 2026-01-21.

[5] Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *CoRR* abs/2011.09533 (2020). arXiv:2011.09533 https://arxiv.org/abs/2011.09533

[6] Ezzat Elokda, Saverio Bolognani, Florian Dörfler, and Heinrich H. Nax. 2026. Dynamic Resource Allocation with Karma: An Experimental Study. arXiv:2404.02687 [econ.GN] https://arxiv.org/abs/2404.02687

[7] Masao Fukushima. 1984. A modified Frank-Wolfe algorithm for solving the traffic assignment problem. *Transportation Research Part B: Methodological* 18, 2 (1984), 169–177.

[8] Christian Gawron. 1998. *Simulation-based traffic assignment.* Ph.D. Dissertation. Universität zu Köln.

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290 [cs.LG] https://arxiv.org/abs/1801.01290

[10] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic Traffic Simulation using SUMO, In The 21st IEEE International Conference on Intelligent Transportation Systems. *IEEE Intelligent Transportation Systems Conference (ITSC).* https://elib.dlr.de/124092/

[11] Anastasia Psarou, Ahmet Onur Akman, Łukasz Gorczyca, Michał Hoffmann, Grzegorz Jamróz, and Rafał Kucharski. 2025. Collaboration Between the City and Machine Learning Community is Crucial to Efficient Autonomous Vehicles Routing. arXiv:2502.13188 [cs.MA] https://arxiv.org/abs/2502.13188

[12] Anastasia Psarou, Łukasz Gorczyca, Dominik Gaweł, and Rafał Kucharski. 2025. Autonomous vehicles need social awareness to find optima in multi-agent reinforcement learning routing games. arXiv:2510.11410 [cs.MA] https://arxiv.org/abs/2510.11410

[13] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *CoRR* abs/1803.11485 (2018). arXiv:1803.11485 http://arxiv.org/abs/1803.11485

[14] Kevin Riehl, Anastasios Kouvelas, and Michail A. Makridis. 2024. Karma economies for sustainable urban mobility – a fair approach to public good value pricing. *npj Sustainable Mobility and Transport* 1, 14 (Dec. 2024). https://doi.org/10.1038/s44333-024-00014-4

[15] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. *CoRR* abs/1706.05296 (2017). arXiv:1706.05296 http://arxiv.org/abs/1706.05296

[16] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA.

[17] Ming Tan. 1997. *Multi-agent reinforcement learning: independent vs. cooperative agents.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–494.

[18] J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis Santos, Rodrigo Perez, Caroline Horsch, Clemens Dieffendahl, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. 2021. PettingZoo: Gym for Multi-Agent Reinforcement Learning. arXiv:2009.14471 [cs.LG] https://arxiv.org/abs/2009.14471

[19] Theodore L. Turocy. 2001. *Gambit: Software Tools for Game Theory, Version 0.2007.01.30.* Technical Report 01-01. Texas A&M University Department of Economics. http://www.gambit-project.org/doc/gambit01.pdf

[20] J. G. Wardrop. 1952. ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH. https://api.semanticscholar.org/CorpusID:131127018

[21] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *CoRR* abs/2103.01955 (2021). arXiv:2103.01955 https://arxiv.org/abs/2103.01955