Autonomous Vehicles need Social Awareness to Find Optima in Multi-Agent Reinforcement Learning Routing Games.

Extended Abstract

Anastasia Psarou
Jagiellonian University, Faculty of Mathematics and Informatics
Krakow, Poland
anastasia.psarou@uj.edu.pl

Łukasz Gorczyca
Jagiellonian University, Faculty of Mathematics and Informatics
Krakow, Poland
lukasz.gorczyca@student.uj.edu.pl

Dominik Gaweł
University of Warsaw, Faculty of Mathematics, Informatics, and Mechanics
Warsaw, Poland
dg448617@students.mimuw.edu.pl

Rafał Kucharski
Jagiellonian University, Faculty of Mathematics and Informatics
Krakow, Poland
rafal.kucharski@uj.edu.pl

ABSTRACT

Previous work has shown that when multiple selfish Autonomous Vehicles (AVs) simultaneously learn optimal routing strategies using Multi-Agent Reinforcement Learning (MARL), they may require a significant amount of time to converge to the optimal solution, equivalent to years of real-world commuting. We demonstrate that moving beyond the selfish component in the reward significantly relieves this issue. In particular, we introduce a reward signal based on the marginal cost matrix, which quantifies the impact of each individual action (route-choice) on the system (total travel time). This formulation reduces training time and improves convergence reliability. Experiments on both a toy network and the real-world Saint-Arnoult network show that the proposed reward improves individual and system travel times over the selfish reward baseline, and in the toy network, enables agents to reach the optimal solution faster, indicating that incorporating social awareness (i.e., including marginal costs in routing decisions) can enhance both system-wide and individual outcomes in future urban systems with AVs.

KEYWORDS

Multi-agent reinforcement learning; autonomous vehicles; urban route choice; marginal cost matrix

ACM Reference Format:
Anastasia Psarou, Łukasz Gorczyca, Dominik Gaweł, and Rafał Kucharski. 2026. Autonomous Vehicles need Social Awareness to Find Optima in Multi-Agent Reinforcement Learning Routing Games.: Extended Abstract. 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/FTHN6981

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/FTHN6981

1 INTRODUCTION

Every day, selfish human drivers [13] make routing decisions to travel from an origin to a destination point through a traffic network. For instance, they commute from home to work by selecting among multiple possible routes. The integration of Autonomous Vehicles (AVs) into transportation systems will introduce new dynamics, as AVs may coordinate, learn, or communicate [5], in contrast to human drivers' selfish routing behavior. To model AV routing decisions, Multi-Agent Reinforcement Learning (MARL) is employed, as in prior works [2, 7, 11, 12]. Yet despite its promise, MARL presents challenges when applied to real-world traffic environments. Prior research has shown that when multiple *selfish* AVs simultaneously learn routing strategies using state-of-the-art MARL algorithms, they may destabilize traffic networks, as these algorithms require long training iterations, equivalent to several years of real-world commuting, to converge to the optimal solution (system optimal or individually optimal), or in some cases, fail to converge at all [10].

In this work, we show that incorporating a social component in the reward of the AV agents accelerates convergence to the optimal solution. Specifically, including the marginal cost in the immediate reward of the AVs brings the system faster to the System Optimal (SO) solution (as shown in Fig. 2 using the network in Fig. 1). This results in shorter average travel times for both AVs and human agents, once the training process is complete. Our results are demonstrated in a small two-route yield (TRY) network (Fig. 1) and in the real-world network of Saint-Arnoult.

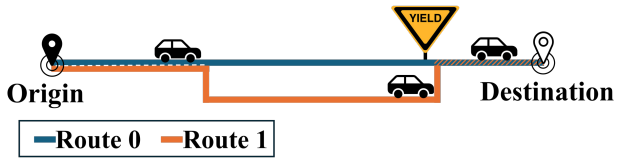

Figure 1: We simulate the routing decisions of 10 AV and 12 human agents, who choose between two alternative routes in the Two-Route Yield (TRY), using the RouteRL framework [3] integrated with the SUMO traffic simulator [8].

Algorithm 1 Marginal cost matrix calculation for joint action $\mathbf{u}_T$

Require: Simulation run T, parameters Θ, agent set $\mathcal{I}$, joint action $\mathbf{u}_T$, travel times t_T
Ensure: Marginal cost matrix M

$M_{i,j}(u) \leftarrow \mathbf{0}^{\mathcal{N}\times\mathcal{N}}$
for $j \in \mathcal{I}$ **do**
  $\mathcal{I}^{(-j)} \leftarrow \mathcal{I} \setminus \{j\}$ ▷ Remove agent j
  $\mathbf{u}_T^{(-j)} \leftarrow \mathbf{u}_T \setminus \{u_j\}$
  env $\leftarrow$ TrafficEnvironment$(\mathcal{I}^{(-j)}, \Theta)$
  Run simulation $\mathcal{E}$ with env under actions $\mathbf{u}_T^{(-j)}$
  for $i \in \{1, \ldots, A_m\}$ **do**
    $t_{\mathcal{E}}(i) \leftarrow$ travel time of agent i in run $\mathcal{E}$
    $M_{i,j}(u) \leftarrow t_{\mathcal{E}}(i) - t_T(i)$
  end for
end for
return $M_{i,j}(u)$

2 MARGINAL COST MATRIX

In this study, we define the marginal cost matrix as $M_{i,j}(\mathbf{u})$ (Algorithm 1), a square matrix, where each entry at position (i, j) is the travel time difference experienced by agent i when agent j is present compared to when agent j is removed from the system. The marginal travel time, denoted as $m_j(\mathbf{u})$ (Eq. 1), is transformed using the hyperbolic tangent function to produce a bounded, sign-preserving value. We denote the travel time of an agent i as $e_i(\mathbf{u})$ and the travel time of agent i when an agent j is removed from the system is denoted as $e_i(\mathbf{u}^{-j})$, where $\mathbf{u}$ is the joint action of the system. The marginal reward, $m_j(\mathbf{u})$, is multiplied by a coefficient β and summed with the negative of the travel time experienced by agent j, produced by the SUMO simulation.

$$m_j(\mathbf{u}) = \sum_{\substack{i=1 \\ i\neq j}}^{N} \tanh\Big(e_i(\mathbf{u}) - e_i(\mathbf{u}^{-j})\Big) = \sum_{i=1}^{N} M_{i,j}(\mathbf{u}) \tag{1}$$

To construct the marginal cost matrix, we rerun the environment for each joint action $\mathbf{u}$, once for each removed AV agent. As a result, this requires an additional simulation for each AV in the system. For example, in the TRY network of Fig. 1, which has 2 discrete actions (routes), a scenario with 10 agents yields 1024 joint actions. If each joint action is encountered during training, an additional 1024×10 environment runs will be needed (this system contains 10 AV agents).

We consider two marginal cost scenarios: the *AV group marginal*, in which AVs account for their impact on other AVs, and the *system marginal*, in which they account for their impact on all drivers, including both humans and AVs. Notably, when the marginal is introduced in the selfish travel-time based reward, we obtain an interplay between the selfish inclination of utility-maximizing travelers to arrive at the destination as fast as possible, and the ideal cooperation to make the system as efficient as possible (see more in [6]). Here, we exploit this property in a system where user equilibrium (UE) is also System-Optimal (SO) (TRY network) (which is rarely the case in complex systems).

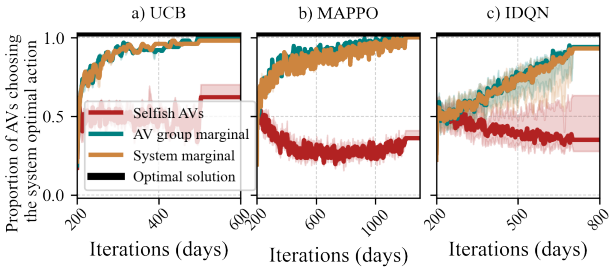

Figure 2: Introducing the marginal travel time in the reward of the AV agents accelerates convergence to the optimal solution both in the *AV group*, and *system marginal* scenarios.

3 RESULTS

TRY network. We demonstrate that under our proposed reward formulation, AV agents using different MARL algorithms converge faster to the optimal solution, as depicted in Fig. 2. In Fig. 2(a, b), the joint action is already close to optimal after about 100 iterations. As shown in Fig. 2, both *AV group marginal* and *system marginal* approaches require a similar amount of time to converge to the optimal solution. The algorithms used are Upper Confidence Bounds (UCB) [4], Multi-Agent Proximal Policy Optimization (MAPPO) [14], and Independent Deep Q-learning (IDQN) [9].

Table 1: Both the system-wide and AV group travel times are reduced when the marginal cost is incorporated into the rewards of the selfish AV agents.

	Selfish AVs	AV group marginal	System marginal
System	27493.92 (15.55)	**27494.81 (2.96)**	**27488.84 (0.74)**
AV group	1855.05 (2.55)	**1851.31 (1.58)**	**1847.85 (3.37)**

Saint-Arnoult real-world network. We present results from a bigger real-world network, the Saint-Arnoult from the URB benchmark [1], where the SO and UE solutions do not coincide. In the experiment, we consider 111 traveling agents, each selecting from an action space of three available routes (resulting in a huge joint action space of 3^{111}). We consider that after a period of human training, 10 of the human agents switch to traveling with AVs, learning the optimal route choice for 300 iterations using the UCB algorithm, and subsequently evaluating the learned policy for an additional 10 iterations. In Table 1, we observe that with a relatively short training (300 iterations, i.e., as much as in the simple TRY network, whereas we are now in a much more complex setting), the system and the AV group average travel times are reduced when we introduce the social component in the reward of the AVs. This result emphasizes that incorporating socially oriented behavior in AVs' routing can be an efficient way to improve both individual and system-level performance in future transportation systems.

ACKNOWLEDGMENTS

This work is financed by the European Union within the Horizon Europe Framework Programme (ERC Starting Grant COeXISTENCE 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, Zoltán György Varga, Grzegorz Jamróz, and Rafal Kucharski. 2024. Impact of Collective Behaviors of Autonomous Vehicles on Urban Traffic Dynamics: A Multi-Agent Reinforcement Learning Approach. In *Seventeenth European Workshop on Reinforcement Learning*. https://openreview.net/forum?id=88zP8xh5D2

[3] 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

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 2-3 (2002), 235–256. https://doi.org/10.1023/A:1013689704352

[5] Daniel J. Fagnant and Kara Kockelman. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice* 77 (2015), 167–181. https://doi.org/10.1016/j.tra.2015.04.003

[6] Michał Hoffmann, Michał Bujak, Grzegorz Jamróz, and Rafał Kucharski. 2025. Wardropian Cycles make traffic assignment both optimal and fair by eliminating price-of-anarchy with Cyclical User Equilibrium for compliant connected autonomous vehicles. arXiv:2507.19675 [eess.SY] https://arxiv.org/abs/2507.19675

[7] Daniel A. Lazar, Erdem Bıyık, Dorsa Sadigh, and Ramtin Pedarsani. 2021. Learning How to Dynamically Route Autonomous Vehicles on Shared Roads. arXiv:1909.03664 [math.OC] https://arxiv.org/abs/1909.03664

[8] 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/

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. arXiv:1312.5602 [cs.LG] https://arxiv.org/abs/1312.5602

[10] 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

[11] Luiz A. Thomasini, Lucas N. Alegre, Gabriel O. Ramos, and Ana L. C. Bazzan. 2023. RouteChoiceEnv: a Route Choice Library for Multiagent Reinforcement Learning. In *Adaptive and Learning Agents Workshop at AAMAS*.

[12] Kagan Tumer and Adrian Agogino. 2006. Agent Reward Shaping for Alleviating Traffic Congestion. (01 2006).

[13] J G WARDROP. 1952. ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH. *Proceedings of the Institution of Civil Engineers* 1, 3 (May 1952), 325–362. https://doi.org/10.1680/ipeds.1952.11259

[14] 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