Position: Collaboration Between the City and the Machine Learning Community is Crucial to Efficient Autonomous Vehicles Routing

Anastasia Psarou [1 2] **Ahmet Onur Akman** [1 2] **Łukasz Gorczyca** [1] **Michał Hoffmann** [1] **Grzegorz Jamróz** [1] **Rafał Kucharski** [1]

Abstract

Autonomous vehicles (AVs) are operating on public roads in several cities. Assuming they use Multi-Agent Reinforcement Learning (MARL) for simultaneous route optimization, higher AV penetration rates may degrade traffic networks' system-wide performance. We study AV routing decisions in a traffic environment shared with human drivers. Our experiments with standard MARL algorithms reveal that, both in simplified and complex networks, policies often fail to converge to an optimal solution or require long training iterations. This convergence issue is amplified by the fact that we cannot rely entirely on simulated training, as there are no accurate models of human routing behavior. In addition, real-world training in cities risks disrupting urban traffic systems, increasing externalities, such as CO_2 emissions, and introducing non-stationarity as human drivers will adapt unpredictably to AV behaviors. **In this position paper, we argue that city authorities must collaborate with the ML community to monitor and critically evaluate the routing algorithms proposed by car companies, ensuring fair, system-efficient algorithms that maintain, or even improve, the performance of urban traffic networks.**

1. Introduction

In urban traffic networks, human drivers every day make routing decisions (Arnott et al., 1990) to arrive at their destinations as fast as possible (Bovy & Hoogendoorn-Lanser, 2005). With the deployment of Autonomous Vehicles (AVs) in some cities worldwide, such as San Francisco (Cheng, 2025), these routing decisions may be delegated to algorithms that aim to maximize the reward by selecting the optimal action given the traffic network's current state. Classically, this routing problem is formulated as a game-theoretical problem (Correa et al., 2004), where humans independently maximize their perceived payoffs.

In mixed systems, where AVs increasingly share the roads with human drivers, they will influence the complex social dynamics of rational yet non-deterministic human driving behavior (Rahmati et al., 2019) for a while, until human driving ceases. However, as we argue, current machine learning (ML) systems will not have (as of today) a sufficiently detailed model of urban mobility capable of training routing algorithms. Furthermore, as we demonstrate, specifically during training, the joint actions of AVs may lead to suboptimal solutions, resulting in costs for all users in the capacitated system with limited resources. Moreover, AVs may take actions different from those of human drivers, which will likely trigger human adaptations that will change their routing policies in response to AVs' actions. This impact is negligible to some level, and single or a few AVs routing recklessly during training will not disequilibrate the system. Yet the critical mass of AVs can be quickly reached as AVs become broadly available (as little as 15% of AVs in our experiment disequilibrate the system). This aligns with projections indicating rapid growth in AV deployment, with the number of commercial AVs for ridesharing expected to reach a few million by 2030 (Sachs, 2024).

This position paper argues that city authorities must begin actively collaborating with the ML research community to monitor and evaluate AI-based routing algorithms for AVs deployed by car companies. In parallel, the ML community should focus on continuously improving existing algorithms to ensure they are robust, fair, and suitable for real-world deployment. Without regulation, autonomous collective routing could exacerbate congestion and introduce chaos into urban traffic systems, which are a shared public resource. We support this position by demonstrating that, in simplified and real-world networks, when multiple AVs simultaneously learn

[1]Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland [2]Doctoral School of Exact and Natural Sciences, Jagiellonian University, Kraków, Poland. Correspondence to: Anastasia Psarou <anastasia.psarou@uj.edu.pl>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

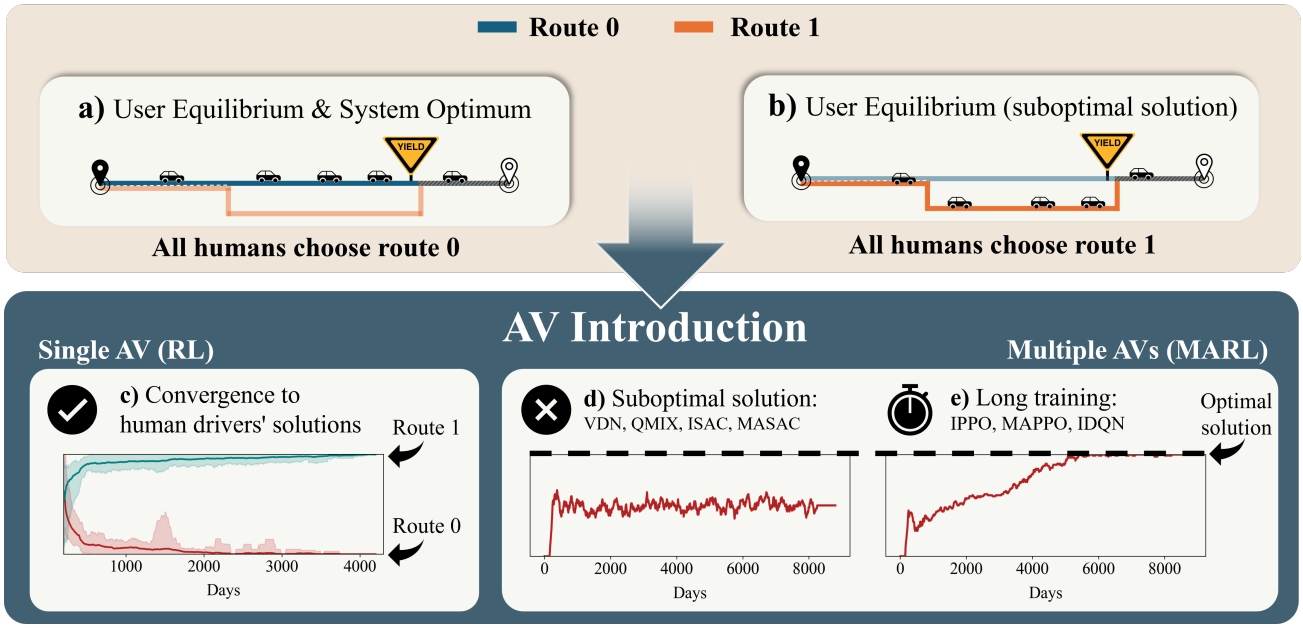


Figure 1. **Overview.** We demonstrate how MARL fails to find optimal routing policies on a simple topology of **T**wo **R**outes: shorter with no priority (**Y**ield sign) and longer with priority (`TRY` network). We simulate 22 human drivers who may reach two equilibria: (a) optimal when all drivers use the shorter route 0, and (b) suboptimal when all use the longer route 1 (the first three humans always choose route 1 in this solution). If we replace a single human driver with an AV, any standard RL algorithm will quickly find the optimal routing policy (c). However, when multiple AVs (10) learn simultaneously, many **MARL algorithms fail to converge to the optimal policy** (d) or require hundreds of days (episodes) until they find it (e) in this trivial case as well as in real complex systems (Ingolstadt, Saint-Arnoult).

routing policies using MARL, they will either disrupt the road networks and fail to find optimal solutions or learn long enough to affect the system's performance. Our argument is supported by experiments conducted on a toy network (Figure 1a, b) and complemented with experiments on the real-world Ingolstadt and Saint-Arnoult networks (Figure 7, 8), using a portfolio of MARL algorithms. We analyze and show that:

- Human drivers, by maximizing their payoffs, stabilize the system into two equilibria (Figure 1(a, b)).
- RL finds the optimal route for a single AV (Figure 3). However, when multiple human drivers are replaced by AVs (acting selfishly or cooperatively), MARL either **fails to converge or needs lengthy training** to find the optimal solution (Figure 4(a, b)).
- The simulators of urban mobility are not ready to serve as virtual environments to train MARL, and training in the real world will be at the cost of the system's performance (Section 3.3).
- Optimal equilibrium state can easily transition to suboptimal (with worse performance for each agent) when non-determinism (e.g., in the traffic simulation) and/or human adaptation is introduced as another source of non-stationarity (Figure 4c).
- Centralization can, in some cases, speed up convergence to optimal policies, but at the cost of privacy (Figure 11b).
- In the real-world Ingolstadt, Saint-Arnoult networks, multiple simultaneously routing AVs can disrupt the system (Figure 5, 9).

These phenomena can hinder the massive potential of AVs to contribute to sustainability (Taiebat et al., 2018), efficiency (Talebpour & Mahmassani, 2016), and optimality (Zhou et al., 2024) of urban mobility. While Connected and Autonomous Vehicles (CAVs) promise novel routing strategies that allow the reduction of total and individual costs (travel times) and system costs (like total delays) (Jamróz et al., 2025) and its externalities (CO_2, NO_x, safety, noise, etc.) (Kopelias et al., 2020) - these benefits can be limited by the challenges described above.

2. Call to Action

Based on experiment findings, to safely and reliably exploit the opportunities that AVs and MARL offer to the future urban traffic systems, **we call for**:

- **The introduction of a regulatory framework**, developed in collaboration with the ML community, requiring car companies to submit their routing algorithms for certification before deployment in public road networks.
- **The development and deployment of monitoring systems** that track collective routing behaviors and can allow authorities to detect issues, like inequitable travel time

distributions or prolonged system-wide congestion.

- **The data-driven development of urban traffic simulators**, led jointly by the ML and transportation research, to realistically reproduce human route choice (in the presence of real-time information) and its adaptation to dynamic environmental changes.
- **The continuous development and improvement of ML algorithms** that can robustly handle non-stationarity, scale with agent populations, and adapt to real-world traffic conditions.
- **Broad experimental studies** to benchmark routing algorithms in diverse traffic scenarios, identify failure modes that inform regulatory policies, and safe algorithm design, using benchmarks like the Urban Routing Benchmark (URB) (Akman et al., 2025a).

An expanded discussion of this Section is included in Appendix B.

3. Background

3.1. Agent Environment Cycle (AEC) game

We abstract the daily (repeated) route choices made by humans and AVs in capacitated networks (limited resources) as a repeated congestion game (Holzman & Law-Yone, 1997). This can be formalized as a one-cycle Agent Environment Cycle (AEC) game (Terry et al., 2021), defined in (Akman et al., 2025a) as a tuple: $\langle \mathcal{S}, \mathcal{I}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \{r_i\}_{i\in\mathcal{I}}, \{O_i\}_{i\in\mathcal{I}}, \upsilon \rangle$. At each episode (day or iteration), a finite set $\mathcal{I}$ of AV agents act sequentially in order υ of departure time, each selecting a route from their action space. Let $N = |\mathcal{I}|$ denote the number of AV agents, and $\mathcal{H}$ the set of human agents. We consider a finite set of states $\mathcal{S}$. At each episode t, each AV agent $i \in \mathcal{I}$, receives an observation $o_i^t \in O_i$ from the environment, and selects an action $a_i^t \in \mathcal{A}_i$, resulting in a joint action $\mathbf{a}^t = (a_1^t, a_2^t, \ldots, a_N^t)$. Then, receives an immediate reward r_i^{t+1}, that is a value of function r_i in episode $t+1$.

3.2. Multi-agent reinforcement learning (MARL)

The route assignment is a combinatorial optimization problem, for which a plausible solution is to employ ML-equipped AVs and specifically use *MARL*, where each agent (AV) learns an optimal policy to select the best route in the currently observed state of the road network. MARL involves multiple agents interacting within a shared environment, where each agent's actions can influence the environment's state. This makes the environment non-stationary from a single agent's perspective. Using a single-agent RL algorithm to learn value functions of *joint* actions would eliminate the non-stationarity (Claus & Boutilier, 1998) but would not scale well when the size of the action space grows exponentially with the number of agents (Lu et al., 2024).

One approach is to train a set of *independent learners* (IL) where each learner treats other agents as part of the environment. We use Independent Deep Q-Learning (IDQN) as the initial baseline (Mnih et al., 2015), followed by the Independent Soft Actor-Critic (ISAC), the multi-agent version of the SAC algorithm (Haarnoja et al., 2018), and the Independent Proximal Policy Optimization (IPPO) algorithm, which has shown benchmark performance in a variety of problems (Yu et al., 2022; Papoudakis et al., 2021).

The existing literature widely uses the *Centralized Training and Decentralized Execution* (CTDE) structure, in which agents learn decentralized policies in a centralized manner (Lowe et al., 2017). Within this framework, Value Decomposition Network (VDN) (Sunehag et al., 2017) and QMIX (Rashid et al., 2018) decompose joint value functions, with QMIX employing a monotonic mixing network. All these algorithms, including the Multi-Agent PPO (MAPPO) (Yu et al., 2022; Schulman et al., 2017) and Multi-Agent SAC (MASAC), are used in our experimental evaluation.

3.3. Limitations: Are traffic models ready to train RL algorithms?

We illustrate our position with a numerical simulation designed to replicate the real world and its complexity. However, the actual conditions in which AVs will be deployed are much more complex. Namely, at the level of:

- **Traffic flow**. We use Simulation of Urban MObility (SUMO) (Lopez et al., 2018), which applies the Intelligent Driver Model (IDM) (Treiber et al., 2000), an accurate, but not perfect, microscopic model of traffic. Real traffic is less predictable and noisier and admits rare events like accidents (Chen et al., 2018).
- **Demand patterns**. In this paper (except from Section 6.6), we assume a fixed commute pattern every day. Namely, each agent has a fixed origin from which it departs every day at the same time to a fixed destination. However, in real systems, humans change departure times, work irregularly, and occasionally stay at home or travel to different destinations (González et al., 2008; Horni et al., 2016; Bhat & Koppelman, 1999).
- **Route choices**. Humans are non-deterministic decision makers, making probabilistic choices, with a significant variation and heterogeneity in the choice probabilities (as in the Logit model (Ben-Akiva & Bierlaire, 1999)).
- **Action space**. In real networks, the number of feasible routes explodes quickly, reaching 10^{40} in many real-world examples (Frejinger et al., 2009).

Unfortunately, all of the above are only roughly approximated by state-of-the-art transport system models. Although

the setup in Figure 1 is simple, it already gives rise to issues (like increased human travel time), and these worsen as the system scales to bigger networks (replicated using the Ingolstadt, Saint-Arnoult networks) with irregular demand patterns and non-deterministic traffic flows (see Figure 4c). In these cases, we get similarly poor results: MARL fails to find the optimal solution, which supports our position.

4. Related work

4.1. Alternative solutions to MARL

While the nature of the problem and its formalization as an AEC game renders (MA)RL likely to be the tool of choice to solve fleet route choice problems we do not argue that it is the best and only solution. The classic traffic assignment solvers identify User Equilibrium (UE) and System Optimum (SO) using Operations Research (OR) methods, like Frank-Wolfe (Fukushima, 1984). However, these methods operate on macroscopic flows and convex, continuous travel-time functions (Kucharski & Drabicki, 2017), while we study an agent-based setting with a realistic, microscopic traffic flow model (SUMO).

Alternatively, game-theoretical equilibria solvers, like Gambit (Turocy, 2001), can be used. However, Gambit requires a precomputed payoff matrix that includes 2^{10} episode runs (for the setup discussed in Figure 1) to account for all possible joint action combinations, and the game's complexity grows exponentially. Genetic algorithms (Holland, 1992) represent another possible alternative to MARL, where a global solution encodes routing decisions for all AVs, making it computationally challenging as the number of vehicles and available routes increases. Hence, both solutions often struggle to adapt to the complex and dynamic nature of mixed traffic scenarios involving AVs and human drivers (Bamdad Mehrabani et al., 2024) and scale poorly in bigger networks. All these highlight the need for alternatives; MARL emerges as the most promising.

4.2. Multi-agent reinforcement learning

RL has been already applied to the route choice problem in works that assume a macroscopic setting, modeling vehicles as aggregate flows rather than individual agents. For example, (Thomasini et al., 2023) studied route choice in a centralized multi-agent setting using a macroscopic traffic simulator. Additionally, (Zhou et al., 2020) formulated route choice as a congestion game, transforming it into the Traffic Assignment Problem (TAP). They proposed an RL-based solution that converges almost surely to the optimal solution, aiming to minimize the total travel time in traffic networks. While macroscopic approaches explore overall traffic dynamics, they abstract away the heterogeneity and strategic behavior of individual agents.

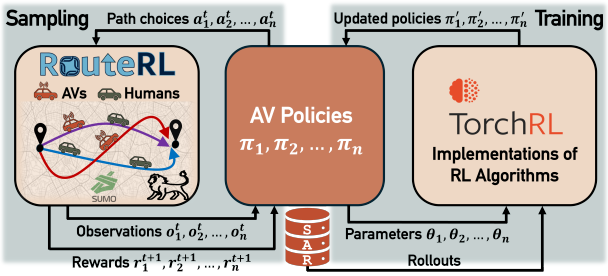


Figure 2. We model the routing "game" between humans and AVs on urban traffic networks using the **RouteRL** (Akman et al., 2025d) framework, which includes a custom *PettingZoo* (Terry et al., 2021) environment, communicating with *SUMO*, a microscopic traffic simulator, as it trains optimal routing policies with standard MARL algorithms via *torchRL* (Bou et al., 2023).

In contrast, microscopic approaches focus on individual vehicle decisions, making them more suitable for modeling AV–human interactions. For instance, (Tumer & Agogino, 2006) adopted a multi-agent approach where drivers use Q-learning for route selection using reward-shaping techniques to reduce traffic congestion. Additionally, (Ramos et al., 2018) proposed a regret-minimization approach that relies on external traffic data.

Some studies model route choice as a sequence of decisions made at each network node (Grunitzki et al., 2014; de O. Ramos & Grunitzki, 2015), allowing for dynamic adaptation to changing traffic conditions. In a different vein, (Lazar et al., 2021) consider a setting where human drivers act selfishly and AVs centrally controlled using deep RL decrease congestion by indirectly influencing human's routing decisions. Lastly, (Akman et al., 2025c) introduced AV-specific behavioral reward formulations in mixed-traffic environments which are later included in the RouteRL framework (Akman et al., 2025d) for simulating the collective route choices of both human drivers and AVs. To the best of our knowledge, no prior work has modeled AVs as independent RL agents in a microscopic, shared environment with human drivers, showing that AVs learning MARL routing strategies could exacerbate congestion.

5. Problem statement

We illustrate our position on a simple **T**wo-**R**oute (**Y**ield) network (`TRY`) (Figure 1(a, b)), which, while abstract, is carefully designed to capture potential issues of AV introduction in traffic networks. To assess the generalizability of our findings we also include two real-world traffic networks. In the `TRY` network, each agent selects from two possible routes (0 or 1) to reach the destination. Then, the reward (travel time) is collected from the environment to update choices for the next episode. Each episode is interpreted as a day on which the 22 drivers commute through the net-

work. For clarity, the setting is unrealistically static. Each human driver follows the same route every day, departing at the same time, and the travel times do not vary day-to-day (none of the above holds in real networks as discussed in Section 3.3). By design, the system has two Nash equilibria achievable by humans, one optimal (System Optimal - User Equilibrium, Figure 1a) and one suboptimal (Suboptimal - User equilibrium, Figure 1b). Humans *mutate* into AVs, which will use any MARL algorithms to find optimal policies. To simulate this scenario, we use the RouteRL (Akman et al., 2025d) framework (Figure 2), which models mixed-traffic scenarios, where AVs are simulated as RL agents and humans behave according to a given human-behavior model.

SUMO. An open-source, state-of-the-art, microscopic, agent-based traffic simulator used as the traffic environment in which each vehicle navigates the road network according to the IDM (Lopez et al., 2018) (see Appendix C).

Action. The action space is the set of available routes connecting an agent's origin and destination and is discrete with value two on the `TRY` network (See Figure 1(a, b)).

Reward. The reward $r_i^{t+1}(a_i^t, o_i^t)$ is the negative travel time of each agent i to reach from its origin to its destination, as calculated by SUMO.

Observation. We assume, plausibly for the future systems, that each AVs' observation o_i^t is composed of their departure time and the number of agents that departed before them and selected each alternative route. In the TRY network, this corresponds to the number of vehicles departed before agent i and chose routes 0 and 1.

Human agents. We follow the classical representation of human route-choice behavior from transportation research. Human drivers are rational decision-makers aiming to maximize their perceived utility (Cascetta, 2009) by selecting actions that minimize expected travel times. Their expectations are updated based on experienced travel times (from SUMO). In scenarios with adaptation (see Section 6.6), humans shift to an alternative route with 10% probability.

5.1. Human system and its equilibria.

We consider two plausible equilibria resulting from human collective decisions: first, when all humans select the shorter route ($[\{0\}^{22}]$, Figure 1a), and second when they all opt for the longer route ($[\{1\}^{22}]$, Figure 1b). Both meet Nash criteria for User Equilibrium (Wardrop, 1952) (common paraphrase of Nash equilibrium for the route choice context), with the former being also System Optimal (minimizing total system travel times (Merchant & Nemhauser, 1978)). None of the drivers is inclined to change their route, as it would reduce their individual rewards (traveling longer and arriving later). For stability, we fix the route for the first three agents to stabilize early system loading (see Appendix D). After 200 days of simulation, the system is stable (the next day the rational drivers will replicate today's choices), fair (travel times are equal among drivers), and either globally optimal (Figure 1a where total travel costs are minimized) or suboptimal (Figure 1b). The individual and system costs (travel times) are reported in Table 2.

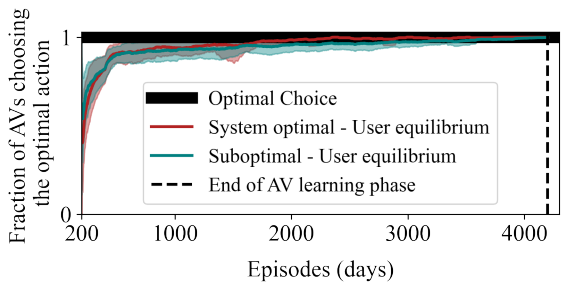


Figure 3. **Single AV** replicates human choices, and each tested RL algorithm finds the optimal solution to the binary choice problem (means, error bars computed across replications; see Appendix A).

6. Experimental support for the position

6.1. Single AV routing with RL.

In the equilibrated human system, we first replace a random human vehicle with a RL-controlled AV. Its traffic properties (reaction, acceptance gap, and other IDM parameters (Treiber et al., 2000)) remain intact. AVs are indistinguishable from humans by all but routing decisions: they may use any RL algorithm to converge to the optimal policy.

We demonstrate that RL finds the optimal solution, and AV agents follow the crowd (replicate the decision of the human driver they replaced). Unsurprisingly, since the problem is a trivial binary decision in a static environment, this is true for all suitable RL algorithms, including DQN, PPO, and SAC, as shown in Figure 3. In any equilibrated system, any RL algorithm training AVs with the same reward formulation, and action space as their human predecessors will replicate the optimal strategy, which can be derived directly from conditions of Nash equilibrium. Moreover, in a dense traffic environment, the impact of a single vehicle is unlikely to be noticed by other humans. The marginal cost of actions taken by a single AV to other humans will fall within what is known as the indifference band (Di et al., 2017), making them indistinguishable from the traffic stochastic noise (Neun et al., 2023). **Our position, that AVs will disequilibrate the traffic networks, does not hold for a system with a single AV**, which will converge to the solution of its human predecessor and will not impact other human drivers.

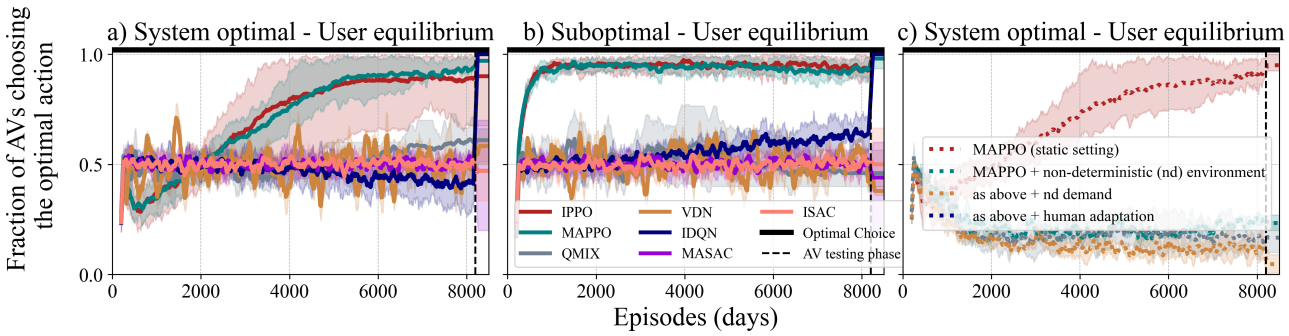


Figure 4. **(a, b): 8000 episodes** (that correspond to over **20 years**) of training for selected MARL algorithms. The trivial solution is found only by IPPO, MAPPO, and IDQN. Other algorithms fail to converge to the optimal policy. Convergence in the suboptimal equilibrium is achieved after 1000 days, but System optimal requires much more training, far exceeding the patience threshold for the remaining 12 human drivers in the system. The first 200 episodes represent the human learning phase, followed by the AV's training and testing phases. **(c):** The MAPPO solution (red), however, is not robust and falls to suboptimal as soon as non-determinism is introduced to the environment (green), demand (yellow), or human behavior (blue). The line represents the mean and standard deviation. See Appendix A for the experiments' setup.

6.2. Simultaneous learning of multiple AVs with MARL.

Already in our simple scenario, serious issues arise when multiple vehicles learn optimal routing policies simultaneously. As multiple AVs are increasingly introduced in cities worldwide, each will potentially be solving the same problem of identifying optimal routing policies to reach their destination. Such a multi-agent setting allows a significant alteration to the initial problem: the AVs may communicate (becoming the so-called CAVs) and share information.

Selfish AV behavior. First, we demonstrate the natural first stage of AV introduction: with no communication and reward formulation identical to the one of the selfish human drivers. Now, the environment has become non-stationary (Jiang et al., 2024), as the state transitions and rewards of an agent are influenced by the evolving policies of other agents. With as few as 10 AVs (where non-convergence appears after the introduction of just 4 AV agents discussed in Section 6.4 and Appendix I, K) our setting is sufficient to argue for our *position*.

Despite the small size of joint action space, some algorithms fail to find the optimal solutions after thousands of iterations. Others need hundreds of policy updates to find the trivial solution. Specifically, the trivial solutions are $[\{1\}^{22}]$ or $[\{0\}^{22}]$ depending on the human equilibrium from which we start, and the joint action space is 2^{10}.

In Figure 4(a, b), we present two arguments supporting our position. First is the class of algorithms that failed to converge after sufficiently long training (MASAC, ISAC). Specifically, the choices are far from optimal, fluctuating noisily around 0.5. The consequences of introducing MARL routing algorithms to AVs are negative to all parties: not only are the travel times longer for all agents (humans and AVs), but they also become variable, as we report in Table 2. IPPO, MAPPO, and IDQN, however, converged after long training. Theoretical explanations of the convergence difficulties are provided in Appendix G.

Cooperative AV behavior Next, we analyze the convergence of AV agents when their reward is the average travel time of all the AV agents in the system. When AVs aim to maximize this new reward, the system's equilibria remain unchanged. As shown in Figure 4(a, b), QMIX and VDN fail to converge to either equilibria even after extensive training, further supporting our position.

6.3. Training in virtual traffic environment.

Some MARL algorithms converge and successfully manage to identify optimal policies (the small joint action space of 1024 can be easily enumerated to identify optimal solutions). Eventually, the 10 AVs manage to stabilize their actions and reliably make optimal choices (see Figure 4(a, b)). After that, the negative impact on the system and the humans diminishes (see Table 2). Nonetheless, the learning process was lengthy, with thousands of episodes (days).

Training can be interpreted in two ways: within a virtual environment or in a real system. The former is neutral to the real system and its users since AVs can train their policies virtually and deploy them only after training is complete and the policies have converged to optimal solutions. Humans will not be affected, and iterations will remain only virtual. This, however, requires *virtual environments* suitable for training such a policy, which are not yet available, as we argued in Section 3.3. If so, the **learning period needs to be treated physically** and algorithmic iterations are not abstract episodes anymore, but physical days of real disturbances. Eventually, disturbances diminish (as for IPPO, MAPPO, IDQN), yet the negative impact on the system and

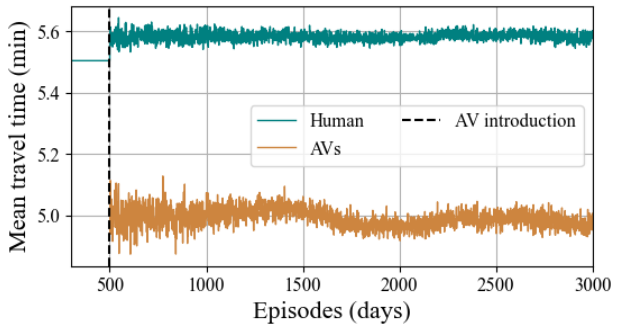


Figure 5. Introducing AVs on the **Ingolstadt** network (with 1000 agents and 1000 nodes) not only increased human travel times but introduced variability of travel times during training, which did not converge in **3000** days (MAPPO training, see Appendix E).

its users accumulates (to values presented in Section H). Alternatively, RL can be inaccurately trained on imperfect models and collected data and later fine-tuned in reality, as shown in (Nair et al., 2021). However, *sim2real* transfer presents significant challenges, as discussed in (Zhao et al., 2020). Real-world complexity exceeds any simulation's capabilities.

6.4. Critical fleet size analysis.

As demonstrated, the problem behind our position lies in multiple agents learning simultaneously. To determine how many simultaneously learning agents can negatively affect the system, we simulate scenarios with gradually increasing numbers of agents and report when convergence issues arise. In each simulation, the specific AVs are chosen at random, with the condition that no two AVs are consecutive (there is always a human agent between AVs), and the first three vehicles are never AVs. The algorithms start failing to converge even with 3-5 agents, as shown in Appendix I. **This reveals a systemic issue that becomes increasingly prevalent as the number of agents grows, which will pose a serious threat to traffic performance when more AVs participate in our daily commute.**

6.5. Privacy of personal data in centralized systems.

Another aspect of our position lies in the communication, and/or centralization (Schwarting et al., 2019), which makes AVs the **C**AVs. This enables better use of autonomous driving, presumably at the cost of sharing private information with others or with a central agent (Nayak et al., 2021). Can we trade our private destination and origin to resolve the non-stationarity issues? To some extent, centralization may speed up the convergence, yet nowhere close to solving the problem and leading to issues with a combinatorially growing search space (see Figure 11 in Appendix J).

Table 1. MARL convergence under varying complexity levels (✓ = deterministic, ✗ = non-deterministic).

Network	Traffic Flow	Demand	Adaptation	Convergence
TRY	✓	✓	✓	yes
TRY	✗	✓	✓	no
TRY	✗	✗	✓	no
TRY	✗	✗	✗	no
Ingolstadt	✓	✓	✓	no
Saint-Arnoult	✓	✓	✓	no

6.6. Scaling to real-world scenarios

To reinforce our results, we reproduce our experiments and investigate what happens when complexity is added to the abstract case. We gradually introduce real-world phenomena from Section 3.3 (see Figure 4c and Table 1), successively including non-determinism in the traffic flow model (green), demand patterns (implemented as random departure times, yellow), and human adaptation (blue).

As these complexities are introduced, the system becomes increasingly disequilibrated. In response, humans will naturally seek ways to improve their payoffs (arrive faster) (Watling & Cantarella, 2013). This behavior is similar to the process of finding the equilibrium (Bie & Lo, 2010), but even less predictable (He & Liu, 2012). This adds another source of non-stationarity to the system. We include a probabilistic adaptation formula (to the human decisions, see Appendix D), and the previously optimal system now shifts to the suboptimal state (Figure 4c).

Finally, we report the results on two real-world traffic networks: Ingolstadt and Saint-Arnoult. For Ingolstadt, we use realistic demand from the RESCO benchmark (Ault & Sharon, 2021), and for Saint-Arnoult we report results from the URB benchmark (Akman et al., 2025a) (see Appendices E, F). The Ingolstadt network comprises 1000 agents, of which 400 are AVs with varying origins and destinations, each selecting from four paths, rendering the joint action space 4^{400}. As shown in Figure 5, the average travel time of all the human agents in the system (pre-AV) is stable. However, after the introduction of AVs (episode 500), the previously stable system becomes unstable, exhibiting fluctuating travel times with values higher than before. A similar phenomenon is observed in the Saint-Arnoult network: when AVs are trained using MAPPO, the system-level average travel time displays variability as depicted in Figure 9.

In all the above-mentioned cases, the **central issue highlighted in our position remains present:** agents did not converge to the optimal solution even after many episodes (4c), and during MARL training, the system was disrupted (demonstrated as the variability of travel times on Figure 5).

Table 2. Average travel times (in seconds) for each subgroup (AVs, humans, and both), with standard deviations within each subgroup in parentheses. 'Human system' refers to rollouts up to the 200th episode, before the introduction of AVs. The remaining values (MARL, Centralized) are calculated from aggregated results during the testing phase and averaged across repeated experiment folds. The lowest travel times for each subgroup in each experimental setting are highlighted in **bold**.

		System Optimum & User Equilibrium			Suboptimal & User Equilibrium		
		AVs	Humans	All	AVs	Humans	All
Human system		-	53.1 (13.1)	53.1 (13.1)	-	65.9 (15.5)	65.9 (15.5)
MARL	IPPO	59.1 (13.3)	55.9 (21.1)	57.4 (18.2)	**69.9 (13.3)**	62.5 (16.4)	**65.9 (15.5)**
	MAPPO	57.4 (12.0)	51.2 (15.6)	54.0 (14.4)	**69.9 (13.3)**	62.6 (16.4)	**65.9 (15.5)**
	ISAC	72.1 (17.4)	71.5 (26.9)	71.8 (23.1)	82.0 (16.7)	61.0 (15.4)	70.6 (19.5)
	MASAC	69.3 (15.3)	70.1 (24.7)	69.7 (21.1)	84.2 (18.0)	60.7 (15.5)	71.3 (20.4)
	QMIX	61.4 (14.6)	55.9 (20.2)	58.4 (18.2)	84.5 (16.6)	60.1 (14.8)	71.2 (19.9)
	VDN	67.6 (14.8)	62.8 (23.0)	65.0 (19.8)	83.8 (18.1)	**60.0 (14.4)**	70.8 (20.2)
	IDQN	**56.7 (11.2)**	**50.6 (14.7)**	**53.4 (13.5)**	69.9 (13.4)	62.5 (16.4)	**65.9 (15.5)**
MARL (Adaptation)	IPPO	65.3 (12.6)	74.1 (30.4)	70.1 (24.4)	69.7 (13.4)	65.4 (18.9)	**67.4 (16.8)**
	MAPPO	65.6 (12.2)	78.0 (29.9)	72.3 (24.4)	69.8 (13.4)	65.5 (18.9)	**67.4 (16.7)**
	ISAC	70.2 (16.8)	66.4 (23.9)	68.1 (21.1)	79.4 (17.9)	63.5 (17.8)	70.7 (19.7)
	MASAC	71.6 (17.4)	68.8 (26.8)	70.0 (23.1)	84.4 (17.9)	62.9 (17.8)	72.7 (20.9)
	QMIX	65.4 (15.3)	58.4 (19.9)	61.6 (18.3)	77.9 (17.8)	63.9 (18.2)	70.3 (19.4)
	VDN	70.8 (16.2)	66.7 (23.6)	68.6 (20.7)	82.2 (19.2)	**62.8 (17.7)**	71.6 (20.8)
	IDQN	**62.4 (13.9)**	**55.3 (17.3)**	**58.5 (16.2)**	72.4 (15.5)	64.8 (18.8)	68.3 (17.8)
Centralized	IPPO	**64.7 (10.6)**	**83.9 (30.3)**	**75.2 (25.3)**	**41.9 (8.0)**	**37.5 (9.8)**	**39.5 (9.3)**
	MAPPO	**64.7 (10.6)**	**83.9 (30.3)**	**75.2 (25.3)**	69.9 (13.3)	62.5 (16.4)	65.9 (15.5)

7. Conclusion

AVs are operating in cities worldwide, and MARL algorithms can be applied to optimize their route choices. In practice, the state-of-the-art MARL algorithms employed in this paper need hundreds of episodes to converge to optimal policies, even in trivial cases. The problem is amplified when more realistic traffic dynamics are introduced in the simulations. The current state of research on human route-choice behavior lacks strict models, verified by extensive data, limiting the realism of virtual simulations. Realistic urban mobility simulators are lacking, and the training episodes would likely need to be deployed directly in real traffic systems, disrupting traffic networks. Needless to say, such a disruption should be avoided at all costs. We hope our contribution will spark discussions within the MARL community, encouraging collaboration between authorities and the ML community to regulate autonomous collective routing. **Such an experimental research program is needed to ensure we fully exploit the opportunities of the new technology (AVs) and algorithms to help us improve the traffic in future cities.**

8. Alternative Views

This section covers some alternative perspectives that could challenge our position.

Authorities should not monitor ML-based routing algorithms. One could argue that authorities do not need to monitor ML-based routing algorithms, since cities already manage infrastructure through traffic signals, congestion pricing, and physical designs. Furthermore, the competitive market incentivizes companies to develop efficient routing tools to offer superior services to their users. While these points are valid, history shows that markets may degenerate into vicious, degrading competition if left alone, especially when novel strategies, whose efficiency may involve dumping, introducing chaos, etc., become available to group players.

What if car companies would not apply ML for the routing policies of AVs? Artificial intelligence has already been integrated into many aspects of our daily lives. As discussed in Section 4, RL has already been applied to optimize vehicle routing. It is therefore natural to expect that such technologies will also be adopted by car companies.

What if reinforcement learning is not suitable for vehicle routing? RL, as discussed in Section 4, is already regarded as a viable solution for routing optimization. However, this paper does not claim that MARL is the only method car companies might adopt, but it argues that companies may deploy a range of algorithmic approaches—some of which could have even more detrimental effects on traffic flow. This is why such routing algorithms must be monitored.

What if the limitations of the virtual environment are

specific to this study and could be resolved with better design or tools? As discussed in Section 3.3, accurately representing dynamic, multi-agent systems like traffic networks requires highly complex simulation environments. In addition, modeling realistic traffic demand patterns involves acquiring and handling data about real-world driver behavior, which is often private and sensitive.

References

Akman, A. O. JanuX, 2025. URL https://github.com/COeXISTENCE-PROJECT/JanuX.

Akman, A. O., Psarou, A., Hoffmann, M., Łukasz Gorczyca, Łukasz Kowalski, Gora, P., Jamróz, G., and Kucharski, R. URB – Urban Routing Benchmark for RL-equipped connected autonomous vehicles. In *Advances in Neural Information Processing Systems*, volume 39, 2025a.

Akman, A. O., Psarou, A., Hoffmann, M., Łukasz Gorczyca, Łukasz Kowalski, Gora, P., Jamróz, G., and Kucharski, R. URB Networks, 2025b. URL https://www.kaggle.com/ds/7406751.

Akman, A. O., Psarou, A., Varga, Z. G., Jamróz, G., and Kucharski, R. Impact of collective behaviors of autonomous vehicles on urban traffic dynamics: A multi-agent reinforcement learning approach, 2025c. URL https://arxiv.org/abs/2509.22216.

Akman, A. O., Psarou, A., Łukasz Gorczyca, Varga, Z. G., Jamróz, G., and Kucharski, R. RouteRL: Multi-agent reinforcement learning framework for urban route choice with autonomous vehicles. *SoftwareX*, 31:102279, 2025d. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2025.102279. URL https://www.sciencedirect.com/science/article/pii/S2352711025002468.

Arnott, R., de Palma, A., and Lindsey, R. Departure time and route choice for the morning commute. *Transportation Research Part B: Methodological*, 24(3):209–228, 1990.

Ault, J. and Sharon, G. Reinforcement learning benchmarks for traffic signal control. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Bamdad Mehrabani, B., Erdmann, J., Sgambi, L., Seyedabrishami, S., and Snelder, M. A multiclass simulation-based dynamic traffic assignment model for mixed traffic flow of connected and autonomous vehicles and human-driven vehicles. *Transportmetrica A Transport Science*, 2024. ISSN 2324-9935. doi: https://doi.org/10.1080/23249935.2023.2257805. URL https://www.sciencedirect.com/science/article/pii/S2324993523003044.

Ben-Akiva, M. and Bierlaire, M. Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pp. 5–33. Springer, 1999.

Bhat, C. R. and Koppelman, F. S. *Activity-Based Modeling of Travel Demand*, pp. 35–61. Springer US, Boston, MA, 1999. ISBN 978-1-4615-5203-1. doi: 10.1007/978-1-4615-5203-1_3. URL https://doi.org/10.1007/978-1-4615-5203-1_3.

Bie, J. and Lo, H. K. Stability and attraction domains of traffic equilibria in a day-to-day dynamical system formulation. *Transportation Research Part B: Methodological*, 44(1):90–107, 2010.

Bou, A., Bettini, M., Dittert, S., Kumar, V., Sodhani, S., Yang, X., Fabritiis, G. D., and Moens, V. Torchrl: A data-driven decision-making library for pytorch, 2023.

Bovy, P. H. and Hoogendoorn-Lanser, S. Modelling route choice behaviour in multi-modal transport networks. *Transportation*, 32:341–368, 2005.

Cascetta, E. *Transportation System Analysis: Models and Applications*. 01 2009. ISBN SBN 978-0-387-75857-2.

Chen, P., Tong, R., Lu, G., and Wang, Y. Exploring travel time distribution and variability patterns using probe vehicle data: case study in beijing. *Journal of Advanced Transportation*, 2018(1):3747632, 2018.

Cheng, E. Robotaxis in 2025: Waymo plots global expansion as zoox, tesla roll to the starting line, 2025. URL https://www.cnbc.com/2025/12/16/waymo-amazon-zoox-tesla-robotaxi-expansion.html. Accessed: 2026-01-21.

Claus, C. and Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pp. 746–752, USA, 1998. American Association for Artificial Intelligence. ISBN 0262510987.

Correa, J. R., Schulz, A. S., and Stier-Moses, N. E. Selfish routing in capacitated networks. *Mathematics of Operations Research*, 29(4):961–976, 2004.

de O. Ramos, G. and Grunitzki, R. An improved learning automata approach for the route choice problem. In Koch, F., Meneguzzi, F., and Lakkaraju, K. (eds.), *Agent Technology for Intelligent Mobile Services and Smart Societies*, pp. 56–67, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg. ISBN 978-3-662-46241-6.

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H. S., Sun, M., and Whiteson, S. Is independent learning all you need in the starcraft multi-agent challenge? *CoRR*, abs/2011.09533, 2020. URL https://arxiv.org/abs/2011.09533.

Di, X., Liu, H. X., Zhu, S., and Levinson, D. M. Indifference bands for boundedly rational route switching. *Transportation*, 44:1169–1194, 2017.

Foerster, J. N., Nardelli, N., Farquhar, G., Torr, P. H. S., Kohli, P., and Whiteson, S. Stabilising experience replay for deep multi-agent reinforcement learning. *CoRR*, abs/1702.08887, 2017. URL http://arxiv.org/abs/1702.08887.

Frejinger, E., Bierlaire, M., and Ben-Akiva, M. Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, 43(10):984–994, 2009.

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

González, M. C., Hidalgo, C. A., and Barabási, A.-L. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008. ISSN 1476-4687. doi: 10.1038/nature06958. URL http://dx.doi.org/10.1038/nature06958.

Grunitzki, R., Ramos, G. d. O., and Bazzan, A. L. C. Individual versus difference rewards on reinforcement learning for route choice. In *2014 Brazilian Conference on Intelligent Systems*, pp. 253–258, 2014. doi: 10.1109/BRACIS.2014.53.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

He, X. and Liu, H. X. Modeling the day-to-day traffic evolution process after an unexpected network disruption. *Transportation Research Part B: Methodological*, 46(1):50–71, 2012.

Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA, USA, 1992. ISBN 0262082136.

Holzman, R. and Law-Yone, N. Strong equilibrium in congestion games. *Games and Economic Behavior*, 21(1-2):85–101, October 1997. doi: None. URL https://ideas.repec.org/a/eee/gamebe/v21y1997i1-2p85-101.html.

Horni, A., Nagel, K., and Axhausen, K. W. (eds.). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London, 2016. doi: 10.5334/baw. URL http://dx.doi.org/10.5334/baw. License: CC-BY 4.0.

Iqbal, S. and Sha, F. Actor-attention-critic for multi-agent reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2961–2970. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/iqbal19a.html.

Jamróz, G., Akman, A. O., Psarou, A., Varga, Z. G., and Kucharski, R. Social implications of coexistence of CAVs and human drivers in the context of route choice. *Scientific Reports*, 15(1):6768, 2025. doi: https://doi.org/10.1038/s41598-025-90783-w.

Jiang, H., Cui, Q., Xiong, Z., Fazel, M., and Du, S. S. A black-box approach for non-stationary multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/pdf?id=LWuYsSD94h.

Kopelias, P., Demiridi, E., Vogiatzis, K., Skabardonis, A., and Zafiropoulou, V. Connected & autonomous vehicles–environmental impacts–a review. *Science of the total environment*, 712:135237, 2020.

Kucharski, R. and Drabicki, A. Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows. *Journal of Advanced Transportation*, 2017(1):4629792, 2017. doi: 10.1155/2017/4629792.

Lazar, D. A., Bıyık, E., Sadigh, D., and Pedarsani, R. Learning how to dynamically route autonomous vehicles on shared roads, 2021. URL https://arxiv.org/abs/1909.03664.

Lobo, S. C., Neumeier, S., Fernandez, E. M. G., and Facchi, C. InTAS – The Ingolstadt Traffic Scenario for SUMO, 2020. URL https://arxiv.org/abs/2011.11995.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., and Wießner, E. Microscopic traffic simulation using SUMO. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. URL https://elib.dlr.de/124092/.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Lu, C., Bao, Q., Xia, S., et al. Centralized reinforcement learning for multi-agent cooperative environments. *Evol. Intel.*, 17(2):267–273, 2024. doi: 10.1007/s12065-022-00703-4. URL https://doi.org/10.1007/s12065-022-00703-4.

Ma, J. and Wu, F. Feudal multi-agent deep reinforcement learning for traffic signal control. In *Adaptive Agents and Multi-Agent Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:216340812.

Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. MAVEN: multi-agent variational exploration. *CoRR*, abs/1910.07483, 2019. URL http://arxiv.org/abs/1910.07483.

Mei, Y., Zhou, H., Lan, T., Venkataramani, G., and Wei, P. MAC-PO: Multi-agent experience replay via collective priority optimization, 2023. URL https://arxiv.org/abs/2302.10418.

Merchant, D. K. and Nemhauser, G. L. Optimality conditions for a dynamic traffic assignment model. *Transportation Science*, 12(3):200–207, 1978.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Nair, A., Gupta, A., Dalal, M., and Levine, S. AWAC: Accelerating online reinforcement learning with offline datasets, 2021. URL https://arxiv.org/abs/2006.09359.

Nayak, B. P., Hota, L., Kumar, A., Turuk, A. K., and Chong, P. H. Autonomous vehicles: Resource allocation, security, and data privacy. *IEEE Transactions on Green Communications and Networking*, 6(1):117–131, 2021.

Neun, M., Eichenberger, C., Martin, H., et al. Traffic4cast at NeurIPS 2022 – predict dynamics along graph edges from sparse node data: Whole city traffic and eta from stationary vehicle detectors. *arXiv preprint arXiv:2303.07758*, 2023.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks, 2021. URL https://arxiv.org/abs/2006.07869.

Rahmati, Y., Khajeh Hosseini, M., Talebpour, A., Swain, B., and Nelson, C. Influence of autonomous vehicles on car-following behavior of human drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2673:036119811986262, 07 2019. doi: 10.1177/0361198119862628.

Ramos, G., Bazzan, A., and da Silva, B. Analysing the impact of travel information for minimising the regret of route choice. *Transportation Research Part C: Emerging Technologies*, 88:257–271, 03 2018. doi: 10.1016/j.trc.2017.11.011.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J. N., and Whiteson, S. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. *CoRR*, abs/1803.11485, 2018. URL http://arxiv.org/abs/1803.11485.

Sachs, G. Partially autonomous cars forecast to comprise 10 percent of new vehicle sales by 2030, 2024. Accessed: 2025-01-27.

Schrab, K., Protzmann, R., and Radusch, I. A large-scale traffic scenario of berlin for evaluating smart mobility applications. In Nathanail, E. G., Gavanas, N., and Adamos, G. (eds.), *Smart Energy for Smart Transport*, Lecture Notes in Intelligent Transportation and Infrastructure. Springer, 2023. doi: 10.1007/978-3-031-23721-8_24. URL https://doi.org/10.1007/978-3-031-23721-8_24.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., and Rus, D. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116 (50):24972–24978, 2019.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V. F., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., and Graepel, T. Value-decomposition networks for cooperative multi-agent learning. *CoRR*, abs/1706.05296, 2017. URL http://arxiv.org/abs/1706.05296.

Taiebat, M., Brown, A. L., Safford, H. R., Qu, S., and Xu, M. A review on energy, environmental, and sustainability implications of connected and automated vehicles. *Environmental science & technology*, 52(20):11449–11465, 2018.

Talebpour, A. and Mahmassani, H. S. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation research part C: emerging technologies*, 71:143–163, 2016.

Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al. PettingZoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

Thomasini, L. A., Alegre, L. N., Ramos, G. O., and Bazzan, A. L. C. RouteChoiceEnv: a route choice library for multiagent reinforcement learning. In *Adaptive and Learning Agents Workshop at AAMAS*, 2023.

Treiber, M., Hennecke, A., and Helbing, D. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805–1824, August 2000. ISSN 1095-3787. doi: 10.1103/physreve.62.1805. URL http://dx.doi.org/10.1103/PhysRevE.62.1805.

Tumer, K. and Agogino, A. Agent reward shaping for alleviating traffic congestion. 01 2006.

Turocy, T. L. Gambit: Software tools for game theory, version 0.2007.01.30. Technical Report 01-01, Texas A&M University Department of Economics, 2001. URL http://www.gambit-project.org/doc/gambit01.pdf.

Wardrop, J. G. Road paper. some theoretical aspects of road traffic research. 1952. URL https://api.semanticscholar.org/CorpusID:131127018.

Watling, D. P. and Cantarella, G. E. Modelling sources of variation in transportation systems: theoretical foundations of day-to-day dynamic models. *Transportmetrica B: Transport Dynamics*, 1(1):3–32, 2013.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and WU, Y. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624. Curran Associates, Inc., 2022.

Zhao, W., Queralta, J. P., and Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744. IEEE, 2020.

Zhou, B., Song, Q., Zhao, Z., and Liu, T. A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game. *Applied Mathematics and Computation*, 371:124895, 2020. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2019.124895. URL https://www.sciencedirect.com/science/article/pii/S0096300319308872.

Zhou, W., Weng, J., Li, T., Fan, B., and Bian, Y. Modeling the road network capacity in a mixed HV and CAV environment. *Physica A: Statistical Mechanics and its Applications*, 636:129526, 2024.

A. Experiments

The experiments were conducted using the RouteRL framework (Akman et al., 2025d), employing version 0.0.1 for the experiments on the TRY network and version 1.0.0 for the experiments on the Ingolstadt network. RouteRL is released under the MIT License.

All experiments on the TRY network involve 200 policy updates, with each update consisting of 40 frames collected per agent (interpreted as episodes or days). Training is performed for 10 epochs, using a minibatch size of 2, and the Tanh activation function. The algorithm scripts are derived from the state-of-the-art (SOTA) implementations provided by the TorchRL library (Bou et al., 2023). All experiments were conducted using version 0.3.0 of the open-source TorchRL library to ensure reproducibility. In the experiment conducted on the Ingolstadt network, 4,000 frames were used with 800 policy updates. The AVs were introduced into the traffic network from episode 500. The training was performed using 1 epoch and a minibatch size of 32. All hyperparameters were consistent with those listed in Table 3.

Each experiment was replicated a different number of times, depending on the scenario. In the case of multiple simultaneously routing AVs (described in Section 6.2 and shown in Figure 4(a, b)), each algorithm experiment had 10 replications for each equilibrium. For the experiments investigating the minimum number of simultaneously routing AVs that can disrupt the traffic network (described in Section 6.4 and shown in Figure 10), each algorithm was replicated for five independent runs for each equilibrium. The experiments exploring the effects of non-determinism and centralization (described in Sections 6.6, 6.5 and shown in Figures 4c and 11) were conducted for 3 independent replications per setting. Hyperparameter optimization was performed for each algorithm and setting (e.g., centralization) separately. The experiment scripts used in this work are available at https://anonymous.4open.science/r/RouteRL_two_route_net-E7BB.

Table 3. Hyperparameters used in the experiments.

Hyperparameter	IDQN	VDN	QMIX	MASAC	ISAC	MAPPO	IPPO	IPPO/MAPPO - Centr
Learning rate	1e-4	1e-2	1e-3	1e-5	1e-5	1e-5	1e-5	1e-4
Memory size	1600	1600	1600	1600	1600	-	-	-
Max gradient norm	-	2	1.5	0.5	0.5	0.5	0.5	1.0
Tau	-	1e-2	1e-2	5e-3	5e-3	-	-	-
Gamma	0.9	0.85	0.95	0.98	0.99	0.99	0.99	0.85
Lambda	-	-	-	-	-	1	1	0.9
Clip epsilon	-	-	-	-	-	0.2	0.2	0.2
Entropy coefficient	-	-	-	-	-	1e-4	1e-4	1e-3

Hardware. Our experiments were carried out on our institution's computing nodes with resources allocated as listed in Table 4.

Table 4. Summary of the hardware used for the experiments.

Component	Specification
CPU	Intel(R) Xeon(R) Gold 5122 CPU, 3.60GHz
GPU	NVIDIA GeForce RTX 2080
RAM	40 GB allocated per job
Operating system	Ubuntu 24.04.1 LTS
Job scheduler	SLURM
SUMO version	1.18.0

Execution time. The experiment computation times depend on the MARL algorithm. We share the computation time of some representative cases in Table 5.

Table 5. Computation time of representative experiments in minutes.

TRAFFIC NETWORK	ALGORITHM	RUNTIME
TRY	IDQN	∼64
TRY	QMIX	∼80
TRY	VDN	∼80
TRY	IPPO	∼200
TRY	MAPPO	∼200
TRY	ISAC	∼220
TRY	MASAC	∼220
INGOLSTADT	IPPO	∼7200

Overall, the experiments described in the paper required approximately 2,000 hours of computation. Full research required up to twice as much computation in total, due to preliminary testing, errors, and our curiosity-driven exploration.

Plots. In all plots, except Figure 5, the lines represent the mean across multiple replications for each episode, and the error bars indicate the standard deviation values. In Figure 3, the line depicts the mean across replications, while the error band accounts for variations due to different agent start times and the use of various algorithms, as described in Section 6.1. All plotted data smoothed using a moving average method with a window size of value 10.

B. Expanded Discussion of the Call to Action

Certification of routing algorithms. For certifying the routing algorithms before deployment in public road networks, we are envisioning a procedure where fleet operators or car manufacturers would submit:

- The operational scope in which the routing policy is intended to operate (e.g., where it can be used, what kind of vehicles and traffic conditions it was designed for)
- The objectives and reward design used during training.
- Evidence of robustness across a representative set of traffic scenarios, incidents, and penetration rates.
- The quantifiable metrics we envision for certification include:
  - The ratio of the total system travel time under a proposed policy π to the total system travel time under the system optimum solution. Certification would require this ratio to remain below a specified maximum threshold.
  - Stability of the learned policy, by computing the day-to-day route distribution shift.
  - Computing fairness metrics including the standard deviation of travel times among humans and AVs, the maximum delays experienced by AVs and human drivers, and the total system travel time in a human-only system and in a mixed-autonomy scenario.

Certification would not require disclosure of the full proprietary source code but instead would rely on standardized documentation and benchmark-based evaluation.

Examples of privacy-preserving monitoring systems. To detect instability caused by AV routing behavior, we can employ subspace anomaly detection (**?**) by analyzing link-count columns using loop detectors, which are already a standard, widely used traffic-sensing technology present in cities worldwide (**?**). This method identifies when traffic patterns deviate from a normal low-dimensional subspace and enables regulators to detect oscillatory or "sloshing" behaviors observed in MARL routing without ever identifying individual vehicles.

To address inequity, such as the offloading of traffic into residential zones through density leakage, we can utilize physics-informed state estimation (**?**). By using sparse and anonymized sensor data, this approach enables regulators to reconstruct the full traffic density of a district. Inequity is then flagged if the reconstructed density or the delays in specific socio-economic zones disproportionately exceed baseline functional-class levels.

For real-time reactions and collective stability, Federated Learning (**?**) can be incorporated into the system to detect the onset of instability by aggregating local model updates from vehicles or edge units. This ensures the central monitor can learn and respond to emerging congestion while raw trajectory data never leaves the local agent, directly addressing the privacy concerns we discussed in Section 6.5. Ultimately, by focusing on the network's fluid dynamics rather than tracking specific agents, the city can maintain a technically robust safety layer that remains fully privacy-compliant.

Technical research directions We believe that research directions should focus on:

- **Safe MARL for transportation systems.** This includes developing policies with verifiable safety constraints, improving robustness under stochastic traffic conditions, and making learned policies more interpretable, so that they can be audited before deployment.
- **Accurate human routing models.** This can be achieved using Imitation Learning to build better human driver models, and hence, simulation-based evaluation will become more realistic.
- **Offline training and sim-to-real transfer.** Better traffic simulators are needed, methods that can handle uncertainty, and techniques that allow policies trained in simulation to transfer reliably to real traffic networks.

C. SUMO

SUMO is the traffic simulator used in this study and is licensed under the Eclipse Public License Version 2.0 (EPL V2). Throughout our experiments, SUMO operates under deterministic conditions, except Figure 4c, where we explicitly introduce non-determinism. Under deterministic settings, if all vehicles select the same routes across consecutive iterations, their travel times remain identical across runs. In the SUMO screenshots below, red vehicles represent human drivers, and yellow vehicles depict AVs. The SUMO network file used to simulate the `TRY` network is available in the anonymized repository at `https://anonymous.4open.science/r/RouteRL_two_route_net-E7BB`, under `RouteRL/network_and_config`, together with simulation videos in the `videos` directory.

(a) System optimal user equilibrium (see Figure 1a).

(b) Suboptimal user equilibrium (see Figure 1b).

Figure 6. SUMO screenshots illustrate the network under the two distinct equilibria.

D. Equilibria of the human system

As illustrated in Figure 1(a,b), the system initially consists solely of human drivers. We assume that the first three agents are human and are exogenously constrained in their route choices: they select route 0 in the system-optimal user equilibrium and route 1 in the suboptimal user equilibrium. Due to the network design, the system-optimal equilibrium occurs when all human agents choose route 0. In this configuration, total system travel time is minimized, and no agent can reduce their individual travel time by unilaterally deviating from this route while others' choices remain fixed. The resulting travel times for all agents under this equilibrium are reported in Table 6.

Conversely, the suboptimal user equilibrium arises when all agents choose route 1, with the first three human agents fixed to that route. Although this equilibrium is stable under unilateral deviations, it yields higher total travel time compared to the system-optimal configuration. The corresponding agent travel times are summarized in Table 6.

Below is the equation defining the route choice in the human adaptation case. Let $X \sim \text{Uniform}(0, 1)$ and let $\epsilon = 0.1$. The action $a \in \{0, 1\}$ is chosen according to

$$a = \begin{cases} 1 - a^*, & \text{if } X < \epsilon, \\ a^*, & \text{otherwise,} \end{cases}$$

Table 6. Individual travel times of human agents under the system-optimal and suboptimal equilibria (SUMO random seed = 23).

Human agents	System optimal- User equilibrium	Suboptimal - User equilibrium
Human agent 1	0.68	0.85
Human agent 2	0.68	0.87
Human agent 3	0.72	0.92
Human agent 4	0.75	0.95
Human agent 5	0.83	1.03
Human agent 6	0.85	1.03
Human agent 7	0.87	1.08
Human agent 8	0.90	1.13
Human agent 9	0.93	1.2
Human agent 10	0.98	1.25
Human agent 11	1.00	1.27
Human agent 12	1.00	1.28
Human agent 13	1.05	1.37
Human agent 14	1.08	1.4
Human agent 15	1.12	1.43
Human agent 16	1.18	1.52
Human agent 17	1.25	1.6
Human agent 18	1.27	1.62
Human agent 19	1.28	1.63
Human agent 20	0.48	0.68
Human agent 21	0.53	0.68
Human agent 22	0.65	0.8

where $a^* = 0$ for the system-optimal user equilibrium and $a^* = 1$ for the suboptimal user equilibrium.

E. Ingolstadt network experiment

The Ingolstadt network consists of 850 links, with demand data sourced from InTAS (Lobo et al., 2020) and used in the RESCO benchmark (Ma & Wu, 2020). This demand includes 1,000 agents. In our experiments, we consider that 400 human agents transition to using AVs in episode 500, each learning to select the optimal path from four available alternatives. These alternative paths are generated using `Janux` software (Akman, 2025), which is integrated into the RouteRL framework. As shown in Figure 5, the system remains stable when only humans are present, with a constant average travel time. However, starting at episode 500, the introduction of AVs increases variability in human drivers' travel times.

F. Saint-Arnoult network results

In this section, we include additional results originally reported in (Akman et al., 2025a). The authors simulated 100% market penetration of an AV fleet on the real-world traffic network of Saint-Arnoult (Figure 8) with a total of 222 agents. The results illustrated in Figure 9 show the case where AVs are trained with MAPPO and cooperative rewards to minimize the average daily CAV travel times. The resulting training instability observed consistently across seeded repetitions aligns with our findings and further reinforces our position.

G. Theoretical challenges of the MARL algorithms applied in this paper

This section outlines the theoretical barriers to the convergence of the MARL algorithms employed in this paper. Although empirical performance is often emphasized in MARL, for example in (de Witt et al., 2020; Mei et al., 2023), theoretical convergence guarantees are often lacking. Moreover, in recent benchmarks, agents are trained for millions of timesteps (de Witt et al., 2020; Papoudakis et al., 2021).

In IL methods, such as IDQN, each agent treats others as part of the environment. This breaks the Markov assumption and leads to non-stationarity, as the transition dynamics $P(s'|s, a_i)$ are no longer fixed, due to the evolving policies of other

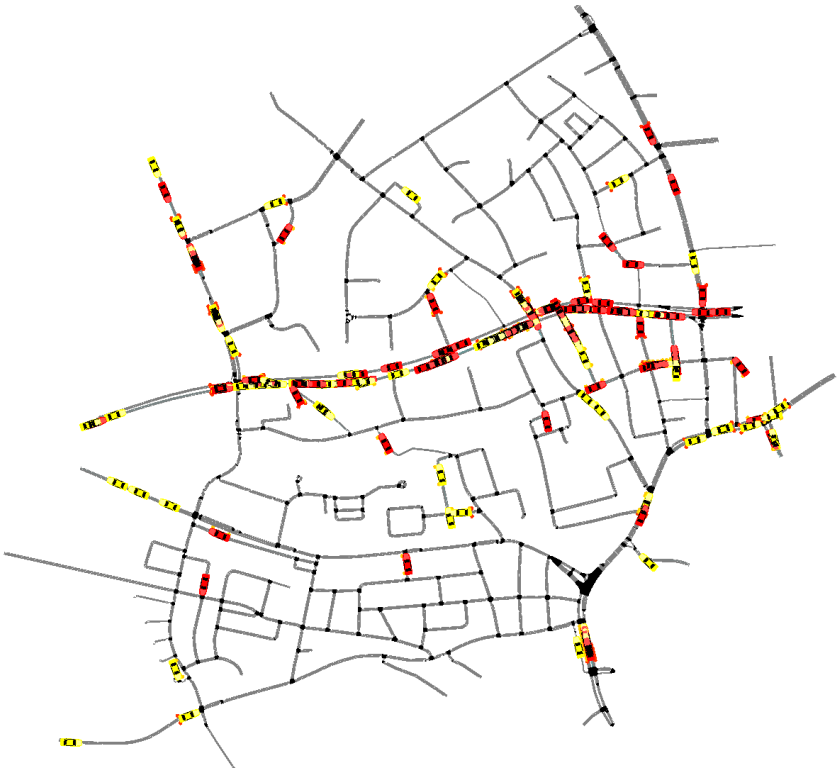


Figure 7. Sumo screenshot from the Ingolstadt network.

Figure 8. Saint-Arnoult traffic network from the URB dataset (Akman et al., 2025b), which consists of 1115 nodes (intersections) and 2397 links.

agents. As a result, even though IDQN is a baseline used in MARL problems that often works well in practice, it lacks theoretical convergence guarantees (Papoudakis et al., 2021; Foerster et al., 2017; Lowe et al., 2017).

Value factorization algorithms, such as VDN and QMIX assume that the joint action-value function $Q_{tot}(s, a)$ can be decomposed into individual functions $Q_i(s, a_i)$, either additively (VDN) or under monotonicity constraint (QMIX). This assumption is violated in many coordination tasks, especially those evolving non-monotonic utility landscapes (Mahajan et al., 2019). However, this property does not guarantee non-convergence. As far as the SAC algorithm is concerned, prior work has proposed a method to reduce the variance and address the credit assignment problem that arises in the multi-agent

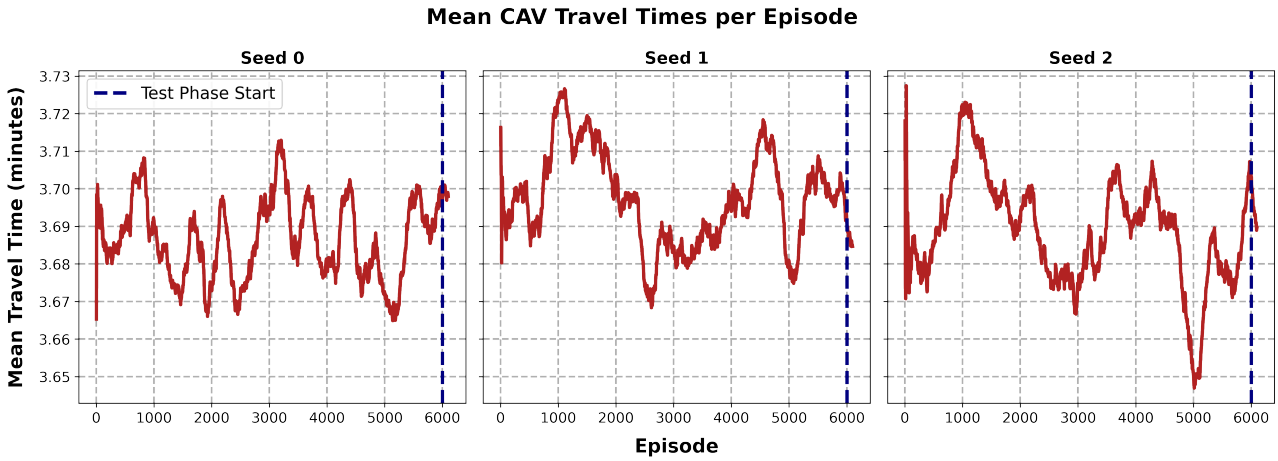


Figure 9. Introducing AVs on the Saint-Arnoult network (222 agents with trips between 215 unique origin-destination pairs) leads to variability in the system travel times during training with the MAPPO algorithm.

setting (Iqbal & Sha, 2019).

H. Cumulative time difference

The cumulative travel time difference denotes the average time loss experienced by an agent from mutation to the final episode, under the assumption that, without mutation, the agent would follow the policy observed in the last pre-mutation episode. Positive cumulative travel time differences, i.e., $c_t > 0$, mean that, on average, human agents after mutation experience longer travel times; thus, the impact of AV introduction on them is negative. This occurred in the system's optimal user equilibrium scenario for each algorithm tested. In the suboptimal user equilibrium, $c_t < 0$ for each algorithm. This means that human agents experience shorter travel times than before the introduction of AVs. This can be attributed to the fact that humans follow the priority route (route 1), and these values are negligible.

Table 7. **Cumulative travel time differences** (in hours) with standard deviations between experiment replications for each user equilibrium and algorithm. The lowest cumulative travel time differences are highlighted in **bold**.

ALGORITHM	SYSTEM OPTIMUM & USER EQUILIBRIUM	SUBOPTIMAL & USER EQUILIBRIUM
IPPO	12.87 (7.17)	-0.21 (0.09)
MAPPO	**11.80 (3.72)**	-0.19 (0.16)
ISAC	33.33 (0.19)	-4.77 (0.04)
MASAC	33.27 (0.39)	**-4.78 (0.04)**
QMIX	36.20 (12.66)	-3.95 (1.47)
VDN	32.88 (0.92)	-4.72 (0.07)
IDQN	35.76 (4.63)	-4.04 (0.34)

Let $\mathcal{H}$ be the set of all agents that do not mutate to AVs. Let N be the number of experiment replications and E the number of episodes. For each experiment $i = 1, \ldots, N$ and episode $e = 1, \ldots, E$, we annotate travel time by $t^i_{e,h}$. Let $\bar{t}^i_e$ describe the average travel time of agents from $\mathcal{H}$ for each episode e and experiment i:

$$\bar{t}^i_e = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} t^i_{e,h}.$$

For each MARL algorithm and user equilibrium, the cumulative travel time difference c_t is given by the following formula:

$$c_t = \sum_{i=1}^{N} \sum_{e=1}^{E} (\bar{t}^i_e - \bar{t}^i_{\text{lh}}),$$

where $\bar{t}^i_{\text{lh}}$ stands for the average travel time of agents $\mathcal{H}$ in the last episode before the mutation in the i-th experiment, i.e.

$$\bar{t}^i_{\text{lh}} = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} t^i_{\text{lh},h}.$$

where lh stands for the last episode before introducing AVs. For each algorithm and each equilibrium, we conducted 10 experiment replications. The set $\mathcal{H}$ contains 12 human agents. Mutation began in the 201st episode, and experiments continued until episode 8300. The values in Table 7 are cumulative travel time differences and the standard deviations over experiments.

I. Critical fleet analysis

Figure 10 shows that, even when a small number of AVs (2–5) learn optimal routing strategies simultaneously, they may fail to converge to the optimal solution after 1,000 days of training.

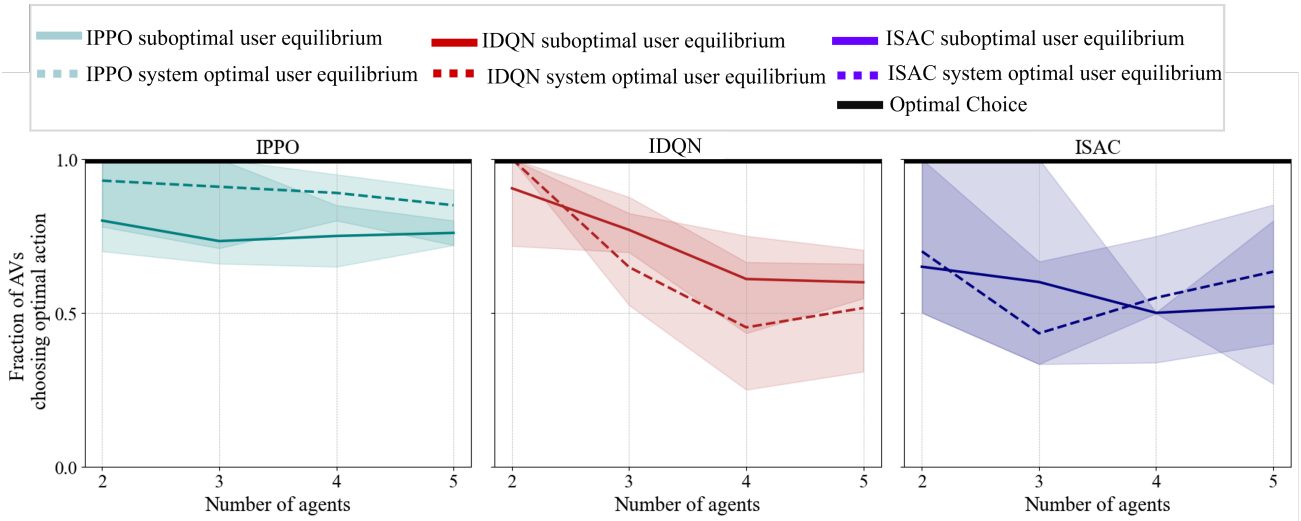


Figure 10. With an increasing number of agents, a fraction of those who learn the optimal policy decreases, achieving similar efficiency to random choice (equilibrium 0: system optimal - user equilibrium - Figure 1a, equilibrium 1: suboptimal user equilibrium - Figure 1b). This appears to happen with as few as four agents for some of the tested algorithms (ISAC, IDQN) that failed to converge in Figure 4.

J. Limitations of the centralized case

In the TRY network, the centralized variant of IPPO can accelerate convergence under the Suboptimal User Equilibrium, reducing the time required to reach the optimal solution, as demonstrated in Figure 11b. However, under the System Optimal User Equilibrium, the centralized approach converges to a suboptimal solution even after long training, showcasing its limited effectiveness in improving convergence (see Figure 11a).

Another limitation of centralized approaches is their limited scalability with respect to the number of agents and available actions (Sunehag et al., 2017). Specifically, in real-world traffic scenarios, such as the Berlin SUMO Traffic (BeST) scenario, introduced in (Schrab et al., 2023), there are more than two million daily vehicle trips distributed over across more than 70,000 distinct routes. At this scale, centralized control or learning becomes computationally infeasible, as the joint state-action space grows exponentially with the number of agents.

K. Results for smaller AV fleets

Figure 12 shows the outcomes of testing episodes under different AV population sizes, highlighting the fraction of AV agents that converge to the optimal solution. The figure reveals that multiple algorithms fail to reach the optimal policy.

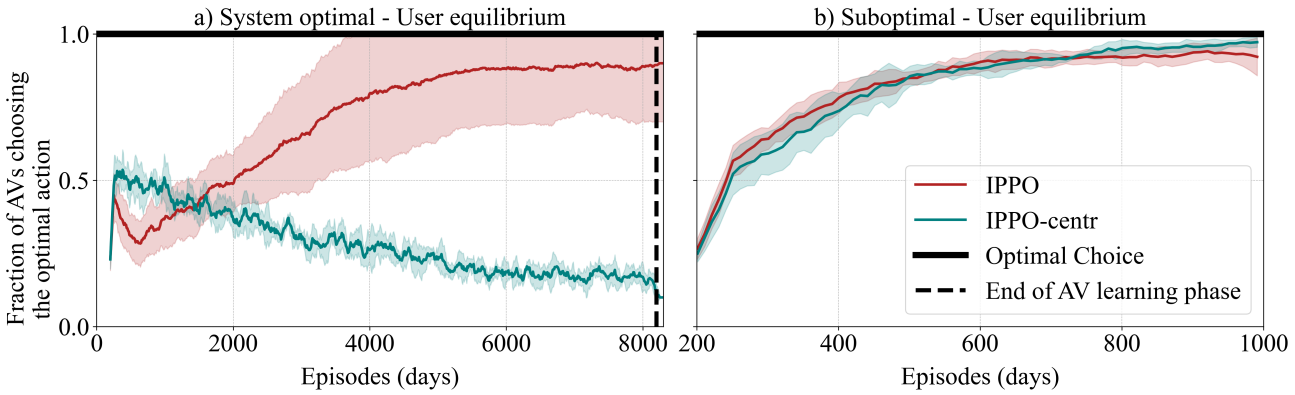
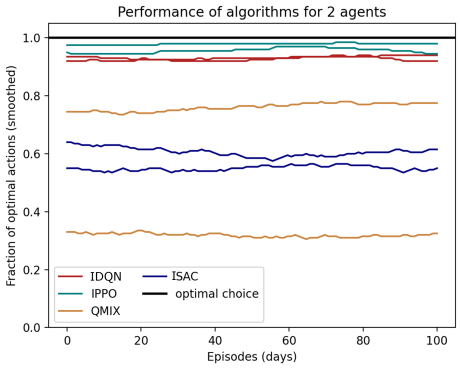
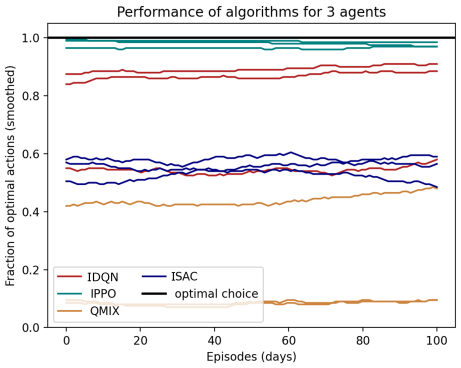


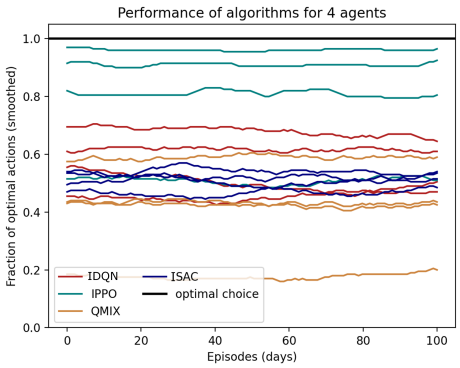
Figure 11. In the System Optimal User Equilibrium case (a), the centralized IPPO converges to the suboptimal solution, showing the limitations of the centralized case. This contrasts with the result shown in (b), where the centralized version of IPPO can accelerate convergence, reducing the time required to reach an optimal solution. However, this reduction is for a few days.



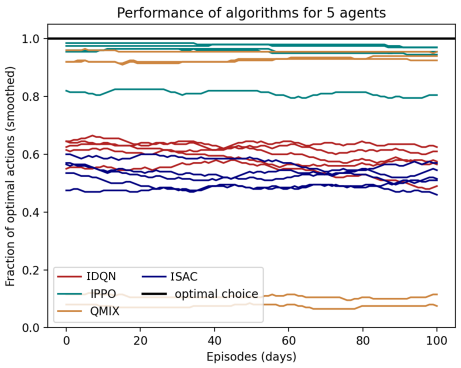
(a) Performance of 2 AV fleet



(b) Performance of 3 AV fleet



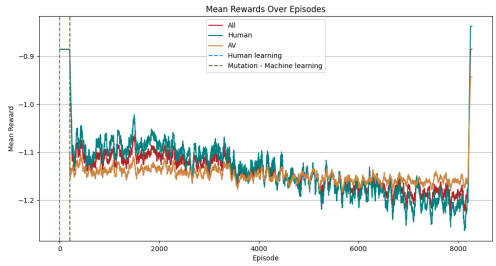
(c) Performance of 4 AV fleet



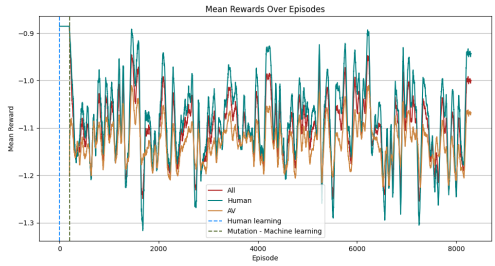
(d) Performance of 5 AV fleet

Figure 12. Results (last 100 episodes—testing phase) of the randomly sampled simulations for smaller fleet sizes. Several algorithms failed to find the optimal solutions and performed worse as the number of AVs in the system increased.

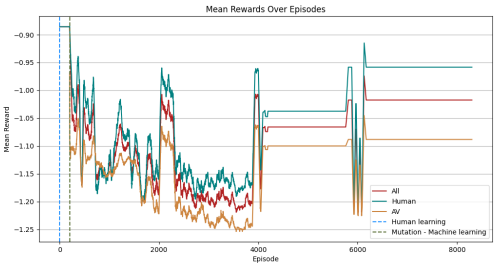
L. Mean rewards



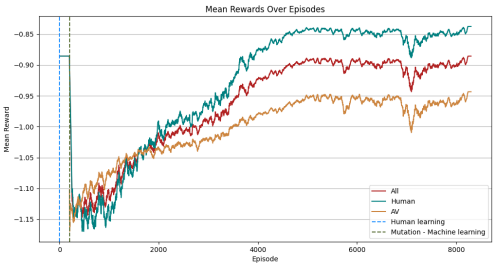
(a) Independent Deep Q-learning (IDQN)



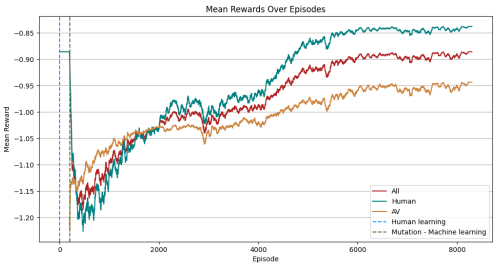
(b) Value-decomposition networks (VDN)



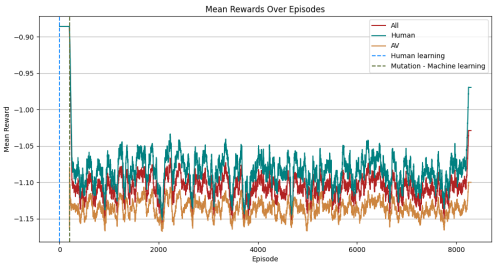
(c) Monotonic value function factorisation (QMIX)



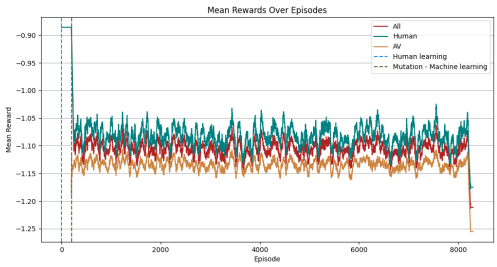
(d) Multi-agent proximal policy optimization (MAPPO)



(e) Independent proximal policy optimization (IPPO)

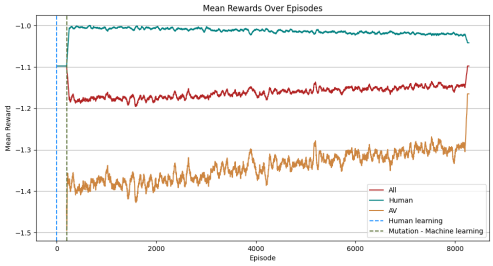


(f) Independent soft actor-critic (ISAC)

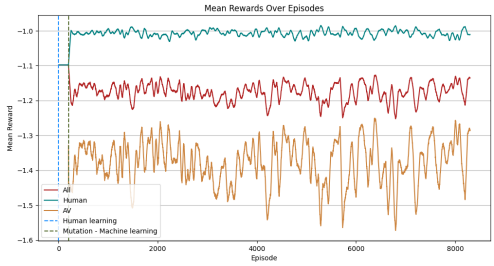


(g) Multi-agent soft actor-critic (MASAC)

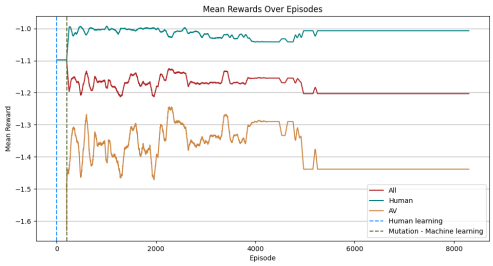
Figure 13. **Mean rewards** of AV agents, human agents, and both combined under the System Optimal User Equilibrium of one experiment seed. The initial 200 episodes of the simulation represent the human learning phase. The mutation event at episode 200 initiates the training of AV agents using different MARL algorithms and marks the end of the human learning phase. The testing phase corresponds to the last 100 episodes. Notably, IPPO, MAPPO, and IDQN algorithms converge after 8000 episodes to the lowest average reward, while other algorithms either stabilize at higher reward values or exhibit variability during training. The episodes in these plots do not correspond to policy updates (see Appendix Section B).



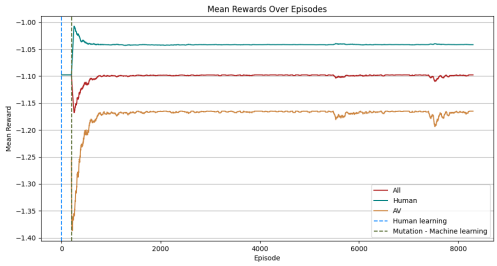
(a) IDQN



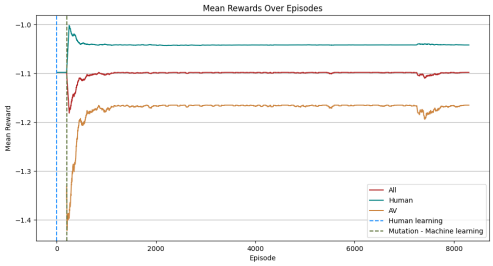
(b) VDN



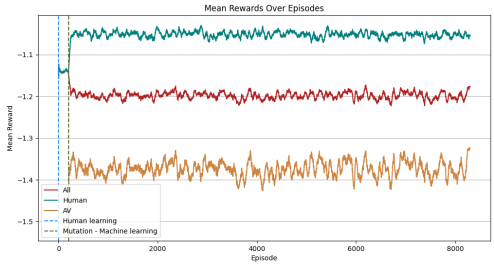
(c) QMIX



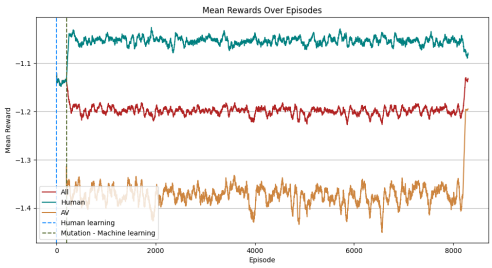
(d) MAPPO



(e) IPPO



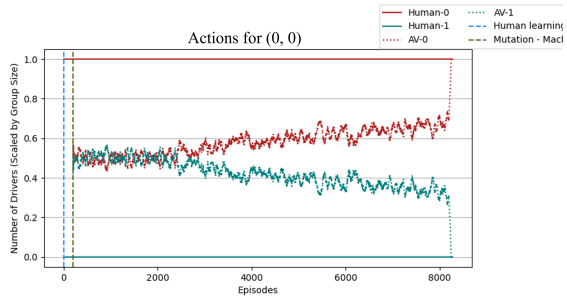
(f) ISAC



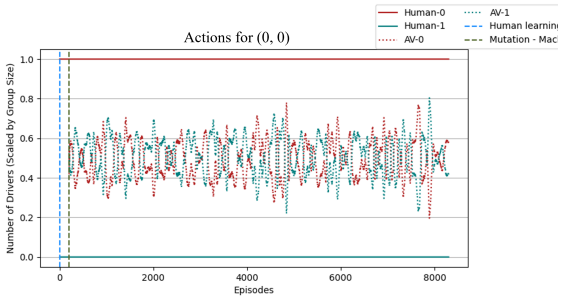
(g) MASAC

Figure 14. **Mean rewards** of AV agents, human agents, and both combined in the Suboptimal User Equilibrium under a specific experiment seed. IPPO, MAPPO, and IDQN achieve the lowest average rewards in the testing phase, and during training, their rewards are less variable than those of the other algorithms. The rewards in this scenario are higher than those in 13, as the system is under the Suboptimal User Equilibrium.

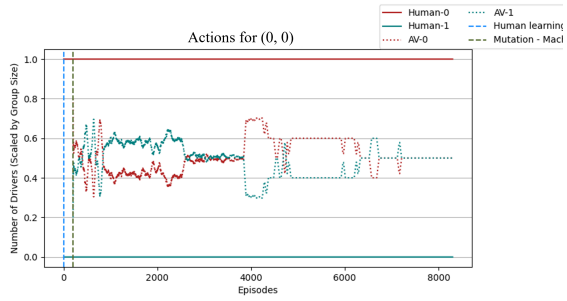
M. Action shifts



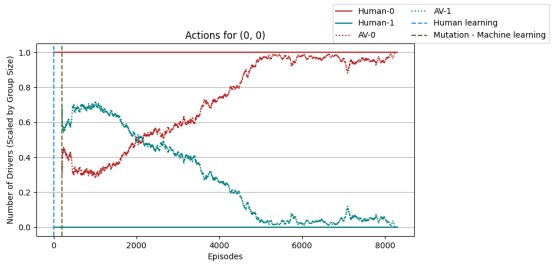
(a) IDQN



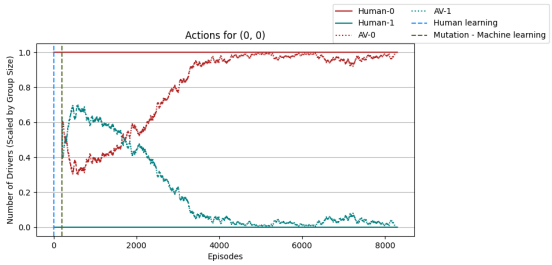
(b) VDN



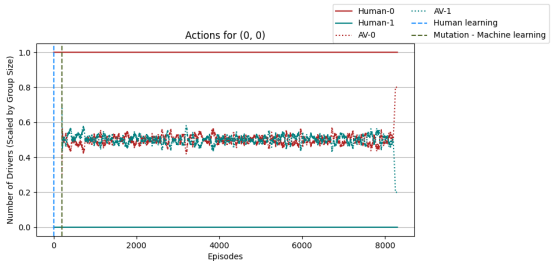
(c) QMIX



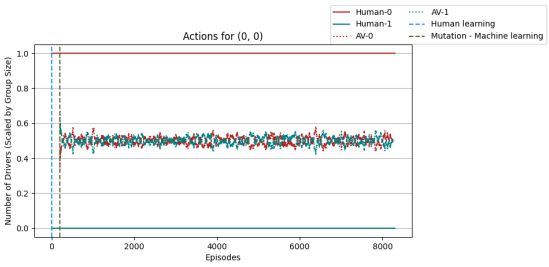
(d) MAPPO



(e) IPPO

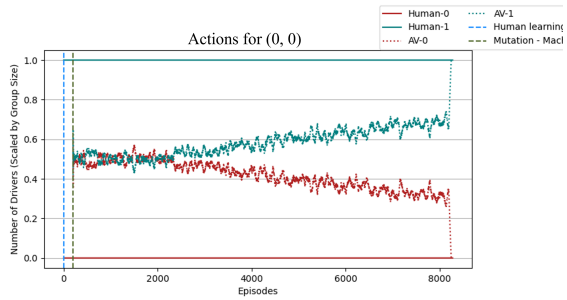


(f) ISAC

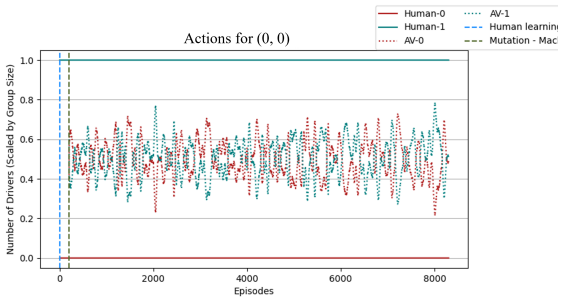


(g) MASAC

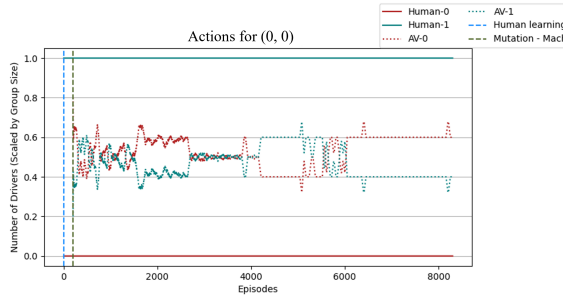
Figure 15. **Action shifts** of AV and human agents under the System Optimal User Equilibrium. The plots show the number of human agents and AVs that choose routes 1 and 0, respectively. Among the algorithms, IPPO converges the fastest to the optimal solution, where all AV agents select route 0. MAPPO and IDQN also converge to this solution, but require additional episodes. The remaining algorithms settle on suboptimal solutions.



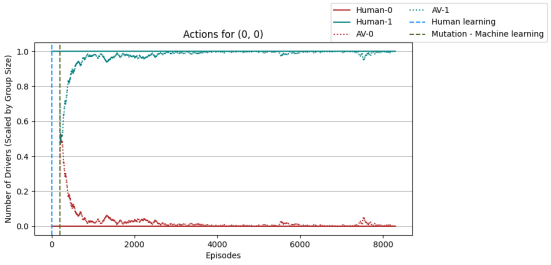
(a) IDQN



(b) VDN

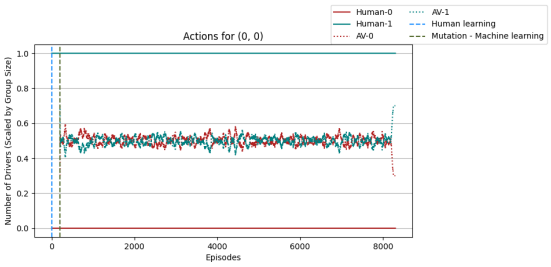


(c) QMIX

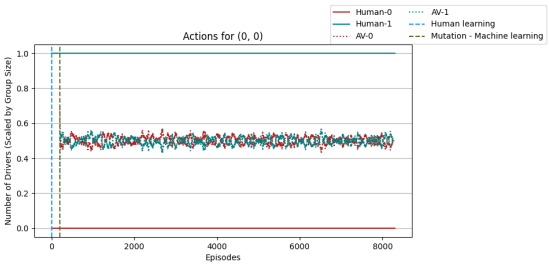


(d) MAPPO

(e) IPPO



(f) ISAC



(g) MASAC

Figure 16. **Action shifts** of AVs and human agents in the suboptimal user equilibrium. The plots illustrate the number of human agents and AVs selecting routes 0 and 1. Among algorithms, IPPO and MAPPO converge even faster to the optimal solution than in the system-optimal user equilibrium scenario, as this equilibrium is more stable. IDQN required longer training to converge to the optimal solution. The remaining algorithms converge to suboptimal solutions.