

Collective Autonomous Vehicle Routing Needs Social Awareness for Coexistence with Human Drivers

Anastasia Psarou*, Mikolaj Rams, Blazej Torbus, Rafal Kucharski

Faculty of Mathematics and Computer Science, Jagiellonian University
Kraków, Poland

Abstract

Collective Autonomous Vehicle (AV) route choices in real-world traffic networks can create externalities for Human-Driven Vehicles (HDVs), as routes that benefit AVs may increase travel times for other road users. To address this, we introduce a social component, the marginal cost, to the selfish travel time AV reward. Marginal cost captures an AV's impact on other road users by comparing their travel times with those in a counterfactual simulation in which the AV is removed.

We study AV route choice in mixed-autonomy traffic settings using Multi-Agent Reinforcement Learning (MARL), where AVs learn routing policies through repeated interactions with the traffic environment. Experiments on the real-world networks of Ingolstadt and Provins show that **incorporating marginal cost into the AV reward not only limits the increase in HDV travel times and shortens AV travel times relative to selfish AV routing, but also improves overall system performance**. In Ingolstadt, selfish routing of 40% AVs increases the travel times of other vehicles by 1.7%, and this drops to 1.1% when the social reward is included. Interestingly, not only does overall system performance improve, but the AVs are also better off when acting socially. This effect is consistent across tested cities, penetration rates, and replications, yet when the weight of the social component in the reward increases, the system fails to reach the desired performance.

For a more scalable approach to social AV routing in mixed-autonomy traffic, we compute the marginal cost only for the AVs whose route choices most affect the traffic system. This reduces the number of counterfactual simulations and achieves similar performance at a significantly lower computational cost.

Our results suggest that AV routing should go beyond selfish rewards and, for the benefit of all traffic participants, include a social component in its objective.

Keywords: Autonomous vehicles, urban route choice, multi-agent reinforcement learning, social awareness, reward shaping.

1 Introduction

Every day, human drivers commute through urban traffic networks, choosing among several possible routes the ones with the minimal expected perceived travel time [4, 38]. The growing presence of Autonomous Vehicles (AVs) in cities may introduce new dynamics, as AVs may deviate from the shortest routes provided by real-time navigation systems and can use a variety of algorithms to learn routing strategies, potentially achieving shorter travel times than Human-Driven Vehicles (HDVs) or disadvantaging them by consistently increasing their travel times [16, 28].

We consider a city that initially consists entirely of HDVs who learn routing strategies by updating their utility through repeated days of commuting [8]. We assume this repeated learning process brings the traffic system close to an equilibrium state [40]. Then, a subset of HDVs switches to traveling with AVs, which use Multi-Agent Reinforcement Learning (MARL) to learn routing policies. As learning progresses, AVs reduce their travel times and eventually achieve shorter average travel times than HDVs. However, we find that when AVs select their routes solely to minimize their individual travel times (selfish behavior), their routing policies can increase the travel times of HDVs relative to those experienced

before AVs were introduced in the system (see Figure 1A).

We therefore ask: **How can AV routing MARL algorithms be designed to reduce AV travel times while limiting the additional costs on other road users, given that urban road networks are shared public infrastructure?** To solve this problem, we augment the selfish travel-time AV reward with a component that accounts for each AV's impact on the system, the marginal cost, following the formulation introduced in [29]. We show that incorporating marginal cost into AV rewards can reduce the increase in HDV travel times and allow AVs to achieve shorter travel times than under selfish AV routing (see Figure 1B). We further find that these travel-time improvements depend on the relative weighting of the two reward components: assigning a greater weight to marginal cost can further improve HDV travel times but may increase AV and overall system travel times.

We further investigate how to make marginal cost computation scalable, since it requires an additional counterfactual simulation per AV each day of commute. In each counterfactual, we remove an AV from the traffic system and compare the remaining vehicles' travel times with those observed when the AV is present. For example, with 400 AVs and 10,000 training iterations (equivalent to commute days), computing marginal costs for all AVs would require four million

*Corresponding author: anastasia.psarou@uj.edu.pl

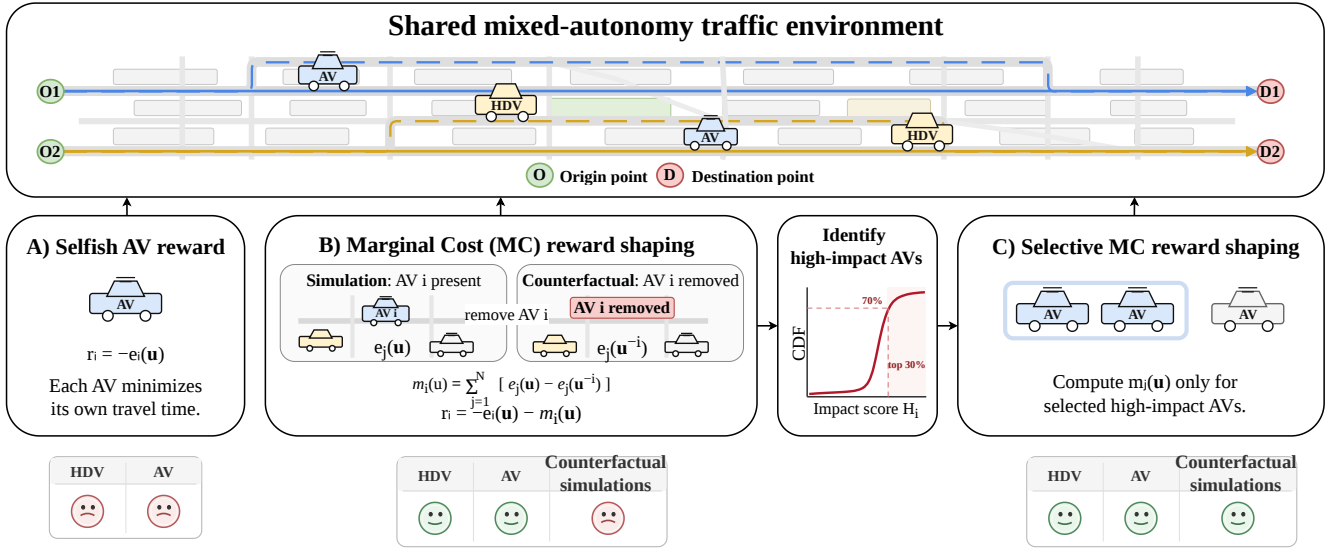


Figure 1: We study AV route choice in traffic environments shared with human drivers. (A) When AVs selfishly minimize their own travel times, their routing decisions can increase HDV travel times. (B) Incorporating our marginal cost into the AV reward allows AVs to account for the impact of their route choices on other road users, mitigating the increase in HDV travel times observed under selfish AV routing while also improving AV travel times on average. However, computing marginal cost requires additional counterfactual simulations for each AV, adding computational cost. (C) To reduce this burden, we introduce selective marginal cost reward shaping, which computes marginal cost rewards only for a subset of the most impactful AVs. This preserves some of the benefits for AVs and HDVs’ travel times while reducing the number of counterfactual simulations.

additional simulation runs. We therefore investigate whether marginal cost rewards can be computed and applied only to a fraction of the most impactful AVs on the traffic system (selective marginal cost reward shaping), rather than to all AVs (marginal cost reward shaping). We then evaluate whether selective marginal cost reward shaping can achieve system, AV, and HDV travel-time improvements comparable to those obtained with marginal cost reward shaping, relative to selfish AV routing (Figure 1C).

The contributions of this work include:

- We show that AV routing policies trained with MARL to minimize individual AV travel times can increase HDV travel times relative to those experienced before the introduction of AVs in the system, revealing a negative externality of selfish AV routing (see Figure 3A, D).
- We demonstrate that incorporating a social component, the marginal cost, into the AV reward can mitigate the increase in HDV travel times observed after AV introduction while also improving AV travel times compared to those achieved when AVs minimize their individual travel times. We further show that AV and HDV travel times depend on how much weight is assigned to the marginal cost. Higher weights can further improve HDV travel times but worsen AV and overall system performance (see Figures 3, 4).
- We show that computing marginal costs only for a fraction of the most impactful AVs, *selective marginal cost reward shaping*, is sufficient to retain HDV, AV, and system travel times comparable to those achieved when all AVs receive the marginal cost reward. This approach also reduces the number of counterfactual simulations (see Table 1). We identify impactful AVs from sampled joint actions that capture how total system travel time varies across alternative AV route choices.

1.1 Literature review

Route choice can be formulated as a congestion game in which travelers repeatedly select among alternative routes in a traffic network while sharing limited road capacity [15]. Assuming homogeneous, fully rational users, this game converges to the Wardrop or User Equilibrium (UE), the Nash equilibrium in the route choice context, where no traveler can improve their travel time by unilaterally switching routes [6, 39]. The System Optimum (SO) is the traffic allocation that minimizes total system travel time across the network [24], but achieving it may require some travelers to take routes that are less favorable for them to improve overall system performance.

Human route choice is commonly modeled using the utility-maximization decision rule [12, 16, 20]. The main component of the utility is the expected travel time for each route [7], which is updated after each trip. In random utility theory, utility additionally includes a random stochastic variable that captures heterogeneity in travelers’ preferences, imperfect perception of travel time and congestion, and fluctuations in driving conditions [8, 40]. This stochastic route choice behavior is assumed to converge to the Stochastic User Equilibrium (SUE), in which no driver *believes* they can improve their travel times by switching routes.

In a mixed-traffic environment where HDVs and AVs coexist, AV route assignment can be solved in several ways. MARL is a promising method [1, 2, 19, 36, 37], where AVs can learn routing policies through repeated interactions with the traffic environment. These policies can be optimized using different objectives, such as minimizing AVs’ individual travel time, the system’s travel time, or even maximizing HDV travel times [1, 2, 16]. Among these objectives, if AVs optimize only their individual travel times, their routing choices may increase other road users’ travel times because of limited road capacity. This motivates the need to search for AV routing objectives that improve AV travel times while limiting the additional

costs imposed on other road users.

In economics, marginal cost denotes the change in total cost resulting from an increase in the quantity produced [23]. In transportation research, marginal cost quantifies a vehicle’s impact on the system by measuring the additional travel time it imposes on other road users through its presence in the traffic system [25, 32]. This idea forms the basis of marginal cost pricing, in which travelers are charged according to the increase in latency their action or presence causes to the system [10, 26]. Marginal cost pricing has also been incorporated into reinforcement learning approaches for route choice. Ramos et al. [30], for example, incorporated marginal cost tolls into the reward signal for commuters using reinforcement learning for route choice, aiming to align individual routing decisions with system-level efficiency and minimize congestion. Inspired by the concept of marginal cost and following [29], we define the marginal cost of an AV as the difference in travel time for all vehicles in a system when the AV is present versus a counterfactual simulation in which that AV is removed.

Marginal cost relates to difference rewards, introduced in cooperative multi-agent settings to address credit assignment when agents optimize a shared global objective [11, 27, 37]. Difference rewards compare the global reward with a counterfactual in which an agent’s contribution is removed or replaced by a default action, providing a reward that reflects the agent’s individual contribution to the system. In traffic routing, Grunitzki et al. [13] studied route choice with reinforcement learning and compared agents trained with selfish travel time rewards and difference rewards, showing that the latter can achieve lower average system travel times.

Prior work introduced the marginal cost reward to help MARL routing AVs reach SO solutions faster than selfish routing AVs in a toy network where the UE and SO solutions coincided [29]. In this work, we examine whether marginal cost reward shaping can prevent selfish AV routing from reducing AV travel times at the expense of HDVs in real-world traffic networks. We additionally propose a method to decrease the counterfactual simulations needed to compute the marginal cost reward by rewarding a fraction of the most impactful AVs in the system.

In the literature, researchers have studied the identification of impactful agents to explain agents’ importance in multi-agent systems. For example, Heuillet et al., [14] used Shapley values to quantify each agent’s contribution to the global reward. Chen et al. [9] stated that an agent is impactful if the system total reward is largely changing when its action is randomized. Following a similar intuition, we consider an AV to be impactful when alternative route choices lead to substantial differences in total system travel time.

Research gap: To the best of our knowledge, little attention has been given to designing AV routing objectives that minimize AV travel times while mitigating travel time increases for other road users in mixed-autonomy traffic. Prior work has explored different AV routing objectives but has not explicitly investigated how to mitigate the travel time externalities these objectives may impose on HDVs or the system [1, 2, 16]. Similarly, existing work on difference rewards has primarily focused on cooperative settings, whereas we use the counterfactual term to account for the externalities that one agent population can impose on another while coexisting in the same environment [11, 13, 27, 37].

We address this gap by incorporating marginal cost into the AV routing reward and examining how the trade-off between individual AV travel times and their system impact during MARL training affects AV, HDV, and system travel times after training. Since AV routing objectives can shape how congestion costs are distributed among agents participating in this complex system. We further investigate selective marginal cost reward shaping, in which we apply marginal cost only to a fraction of the most impactful AVs. We examine whether we can achieve similar AV, HDV, and system travel times with fewer counterfactual simulations, providing a more scalable approach to social AV routing.

2 Preliminaries

2.1 Urban routing problem

We study the route choice problem in which, each day, a number of vehicles make one trip from an origin to a destination point in an urban traffic network at a specific departure time by selecting a route from a set of discrete alternatives. To simulate this problem, we use the *RouteRL* [3] framework integrated with *SUMO* [22], a continuous microscopic traffic simulator that enables modeling complex urban traffic networks and agent interactions.

Our urban environment initially consists only of HDVs, which choose routes using random utility theory [8]. After consecutive days of commuting, HDV route choices stabilize. Then, a number of HDVs switch to using AVs, which can use various MARL algorithms to learn routing strategies. AVs keep the departure time, origin, destination points, and action space (set of routes they can travel on) of their human predecessors. Human learning is thereafter disabled, and human route choices are sampled from the fixed learned distribution. This provides behaviorally plausible human route choices that AV routing policies can be trained on before deployment. AV observations include an estimate of each route’s travel time based on travel times of earlier-departing vehicles with the same origin and destination, together with the AV’s origin, destination, and departure time. The set of routes does not change during the simulation or from one day to the next, and both departure times and demand remain constant across simulation days.

2.2 Problem formulation

We abstract the daily (repeated) route choices made by humans (and AVs in the future) in capacitated networks (limited resources) as a repeated congestion game [15]. This can be formalized as a one-cycle Agent Environment Cycle (AEC) problem [35], defined in [2, 28] as a tuple:

$$\langle \mathcal{S}, \mathcal{N}, \mathcal{I}, \{\mathcal{U}_i\}_{i \in \mathcal{N}}, \{r_i\}_{i \in \mathcal{I}}, \{\Omega_i\}_{i \in \mathcal{I}}, v, \rangle$$

where $\mathcal{N} = \mathcal{I} \cup \mathcal{H}$ denotes the set of all agents, with $\mathcal{I}$ representing AV agents and $\mathcal{H}$ human agents. In each episode (day or iteration), each AV agent in $\mathcal{I}$ and each human agent in $\mathcal{H}$ selects a route from their respective action space, with agents acting sequentially according to v . The AEC game tuple includes:

- A finite set of states $\mathcal{S}$.
- A finite set of AV agents $\mathcal{I}$, a finite set of human agents $\mathcal{H}$.
- For each agent $i \in \mathcal{N}$:
 - The agent selects an action $u_i \in \mathcal{U}_i$, forming a joint action $\mathbf{u} \in \mathbf{U} \equiv \mathbf{U}^N$.

- For each agent $i \in \mathcal{I}$:
 - r_i is the immediate reward agent i aims to maximize, as defined in [2, 28].
 - Ω_i is the set of possible observations for agent i .
- v is the agent-selection mechanism that determines which agent acts next by sorting agents by their departure time from their origin.

In this problem, AVs learn routing policies $\pi_i(u_i | \omega_i)$ over consecutive days (episodes) simulated in SUMO to maximize their immediate rewards. A wide range of MARL algorithms can be used to learn such policies, which select routes based on each AV's current observation ω_i .

3 Method

3.1 Reward with marginal cost

We postulate that adding to the AV reward a component that captures the agent's impact on all other agents in the system, in our case, the marginal cost, can guide AVs towards route choices that account for the travel time increases they may impose on HDVs. We formulate the reward of an AV agent i as:

$$r_i(\mathbf{u}) = -e_i(\mathbf{u}) - \beta m_i(\mathbf{u}), \quad i \in \mathcal{I} \quad (1)$$

where $e_i(\mathbf{u})$ denotes the travel time experienced by individual AV i , under the joint action $\mathbf{u}$ as computed by SUMO, and $m_i(\mathbf{u})$ represents the marginal cost of AV i in the system. The parameter β weights the marginal cost component in the agent's reward. The marginal cost of an agent i is represented by:

$$m_i(\mathbf{u}) = \sum_{\substack{j=1 \\ j \neq i}}^N (e_j(\mathbf{u}) - e_j(\mathbf{u}^{-i})), \quad N = |\mathcal{H} \cup \mathcal{I}|, \quad (2)$$

and corresponds to the travel time differences of all the agents j in the system, when agent i is present, and in a counterfactual simulation where agent i is removed. $\mathbf{u}^{-i}$ denotes the joint action of the system where agent i is not present in the simulation.

When $\beta = 0$, AVs behave selfishly by minimizing their individual (selfish) travel time. Positive values of β introduce a trade-off between AVs routing selfishly and accounting for the impact of their route choices on the system.

3.2 Marginal cost table

To compute the marginal cost for each AV, we construct the marginal cost matrix $M(\mathbf{u})$, which records the impact of each AV on the travel times of all other agents in the system. Each entry $M_{i,j}(\mathbf{u})$ quantifies the impact of agent j on the travel time of agent i .

$$M_{i,j}(\mathbf{u}) = \begin{cases} e_i(\mathbf{u}) - e_i(\mathbf{u}^{-j}), & i \neq j, \\ 0, & i = j, \end{cases} \quad j \in \mathcal{I}. \quad (3)$$

The marginal cost of agent j , $m_j(\mathbf{u})$, is obtained by summing the entries in column j of the marginal cost matrix, capturing the total change in travel time experienced by all agents in the system due to the presence of AV j .

We build the marginal cost matrix $M(\mathbf{u})$ by rerunning the simulation for a specific joint action $\mathbf{u}$, once for each AV

Algorithm 1 AV training with marginal cost reward

Require: Human agents $\mathcal{H}$, AV agents $\mathcal{I}$, AV policies $\{\pi_i\}_{i \in \mathcal{I}}$, HDV route-choice distributions $\{P_h\}_{h \in \mathcal{H}}$, set of AVs for which we compute the marginal cost $\mathcal{I}_{\text{MC}} \subseteq \mathcal{I}$, coefficient β

```

1: for each episode (day) do
2:    $\mathbf{u} \leftarrow \emptyset$ 
   // Main simulation
3:   for each agent  $k \in \mathcal{H} \cup \mathcal{I}$  by  $v$  order do
4:     Advance the simulation to agent  $k$ 's decision time
5:     if  $k \in \mathcal{H}$  then  $\triangleright$  HDVs
6:       Sample route  $u_k \sim P_k$ 
7:     else  $\triangleright$  AVs
8:       Observe  $\omega_k$  and select route  $u_k \sim \pi_k(\cdot | \omega_k)$ 
9:     end if
10:    Agent  $k$  takes route  $u_k$ ,  $\mathbf{u}[k] \leftarrow u_k$ 
11:  end for
12:  Record travel times  $\{e_j(\mathbf{u})\}_{j \in \mathcal{H} \cup \mathcal{I}}$ 

   // Counterfactual simulations
13:  for each AV  $i \in \mathcal{I}_{\text{MC}}$  do
14:    for each agent  $k \in \mathcal{H} \cup \mathcal{I}$  in  $v$  order do
15:      Advance the simulation to agent  $k$ 's departure
time
16:      if  $k \neq i$  then
17:        Apply the stored route  $u_k$ 
18:      end if
19:    end for
20:    Record counterfactual travel times  $\{e_j(\mathbf{u}^{-i})\}_{j \neq i}$ 
21:     $m_i(\mathbf{u}) \leftarrow \sum_{j \neq i} (e_j(\mathbf{u}) - e_j(\mathbf{u}^{-i}))$ 
22:     $r_i(\mathbf{u}) \leftarrow -e_i(\mathbf{u}) - \beta m_i(\mathbf{u})$ 
23:  end for

   // AV learning
24:  for each AV  $i \in \mathcal{I}$  do
25:    Update policy  $\pi_i$  using  $r_i(\mathbf{u})$ 
26:  end for
27: end for

```

agent. In each additional simulation, we remove a single AV j , while keeping the routes and departure times of all other agents fixed. As a result, constructing the matrix for one episode (joint action) requires generating $|\mathcal{I}|$ additional and separate environment instances. Since these counterfactual simulations are independent, they can be executed in parallel to reduce computation time. Algorithm 1 details the marginal cost reward computation during AV training.

3.3 Selective marginal cost

To reduce the number of simulations needed to compute marginal cost counterfactuals for all AV agents in the system, we propose selective marginal cost reward shaping. In selective marginal cost reward shaping, we rank AVs by their impact on the system. We assess an AV's impact by how much system travel times differ when it chooses different routes. Then, we compute the marginal cost m_i only for a fraction of the most impactful AVs, forming the subset $\mathcal{I}_{\text{MC}} \subseteq \mathcal{I}$. The remaining AVs aim to minimize their own travel time. In

this way, selective marginal cost reward shaping concentrates the additional counterfactual computation on the AVs with the largest estimated system impact, allowing us to investigate whether similar AV, HDV, and system travel times can be achieved with fewer counterfactual simulations.

To identify impactful AV agents, we construct a dataset $\mathcal{D}$ of sampled joint action realizations. Each sample $d \in \mathcal{D}$ represents a day of commute, where $\mathbf{u}^{(d)} = (u_1^{(d)}, \dots, u_N^{(d)})$ is the vector of joint actions and the vector of realized travel times is $\mathbf{e}^{(d)} = (e_1(\mathbf{u}^{(d)}), \dots, e_N(\mathbf{u}^{(d)}))$. For each sample, we compute the total system travel time as:

$$T(\mathbf{u}^{(d)}) = \sum_{i=1}^N e_i(\mathbf{u}^{(d)}), \quad N = |\mathcal{H} \cup \mathcal{I}|. \quad (4)$$

For each AV agent $i \in \mathcal{I}$ and each available action $a \in \mathcal{U}_i$, we then extract the subset of samples in which agent i selects action a :

$$\mathcal{D}_{i,a} = \left\{ d \in \mathcal{D} \mid u_i^{(d)} = a \right\}. \quad (5)$$

Then, we compute $\bar{T}_{i,a}$ as the average total system travel time over all such samples where agent i chooses action a :

$$\bar{T}_{i,a} = \frac{1}{|\mathcal{D}_{i,a}|} \sum_{d \in \mathcal{D}_{i,a}} T(\mathbf{u}^{(d)}) \quad (6)$$

Then, for each action $a \in \mathcal{U}_i$ we collect the resulting action-conditioned system travel times $\bar{\mathbf{T}}_i = [\bar{T}_{i,a_1}, \bar{T}_{i,a_2}, \dots, \bar{T}_{i,a_{|\mathcal{U}_i|}}]$. We then quantify the historical impact of agent i , denoted as H_i , as the standard deviation of the average system travel times across each possible action.

$$H_i = \text{Std}_{a \in \mathcal{U}_i} (\bar{T}_{i,a}). \quad (7)$$

A high H_i value indicates that, among the samples in dataset $\mathcal{D}$, the system travel times differ substantially when agent i chooses different routes, while a low H_i value indicates small system travel times changes when agent i chooses different routes. By ranking AV agents by this historical impact score H_i , we can reward a fraction of the most impactful AV agents with the marginal cost reward formulation. Selective marginal cost reward shaping therefore allows us to examine whether similar AV, HDV, and system travel times can be achieved with fewer counterfactual simulations.

4 Results

We begin by introducing the experimental setting in Section 4.1. Next, in Section 4.2, we show that on two real-world traffic networks, selfish AV routing can disadvantage HDVs, whose travel times increase after AV introduction. We then demonstrate that rewarding all AVs with the marginal cost can mitigate the increase in HDV travel times and additionally improve AV travel times. Finally, in Section 4.3, we apply the marginal cost reward only to a fraction of the most impactful AVs, and show that rewarding just 40% of AVs can achieve AV, HDV, and system travel times comparable to those obtained when marginal cost is applied to all AVs, while reducing the number of additional simulation runs by 40%.

4.1 Experimental setting

We assess the performance of our marginal cost reward formulation through experiments in the real-world Ingolstadt and

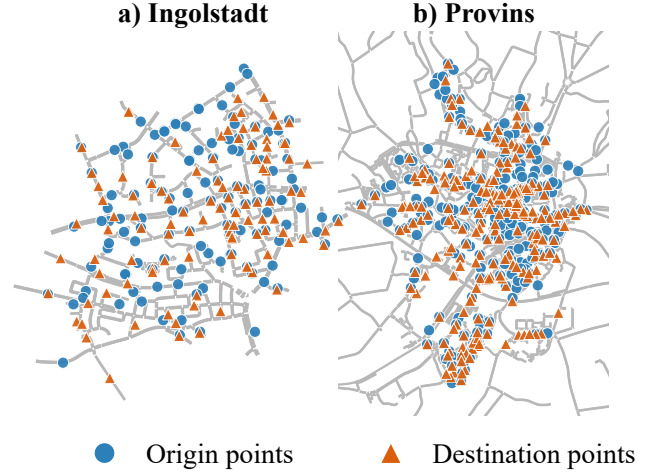


Figure 2: Traffic networks of Ingolstadt with 1035 agents (a) [21], and Provins (b) with 523 agents [2].

Provins traffic networks, shown in Figure 2. The Ingolstadt traffic network includes 850 nodes, 1426 links, 306 unique Origin-Destination (OD) pairs, and a demand of 1035 agents sourced from the InTAS traffic scenario [21] and used in the RESCO benchmark [5]. The Provins traffic network includes 2173 nodes, 4216 links, 517 unique OD pairs, and 523 traveling vehicles, sourced from the URB benchmark [2].

At the beginning of each experiment, the system contains only HDVs, which repeatedly choose routes among four options over 300 days of commute, as in [2, 3, 16]. On day 301, a specified fraction of HDVs is replaced by AVs that learn MARL routing policies using the REINFORCE algorithm [18, 41], where each agent is an independent learner [34]. We use REINFORCE in our experiments, but the proposed reward formulation is not specific to this algorithm and can be incorporated into any MARL algorithm. We train the AVs for 10,000 training iterations or days of commute in both the Ingolstadt and Provins traffic networks, followed by a 100-day testing phase in which the learned policies are no longer updated. All results reported in Sections 4.2 and 4.3 are computed over the 100-day testing phase. We report the REINFORCE algorithm hyperparameters in the Supplementary Material.

For selective marginal cost reward shaping, we identified the most impactful vehicles from the 800k joint action dataset $\mathcal{D}$. We sampled from enormous space of possible joint actions (for 1,000 vehicles where each vehicle can choose among four alternative routes, there are 4^{1000} possible joint actions) a diverse dataset $\mathcal{D}$ using several sampling methods, including random, balanced, Dirichlet [33], Maximin distance sampling [17], and simplex-lattice design [31] (see Section 3.3).

4.2 Performance of the marginal cost

We demonstrate in Figure 3(a, d) that in both the Ingolstadt and Provins real-world traffic networks, humans consistently experience longer travel times across all examined AV penetration rates when AVs selfishly minimize their individual travel times ($\beta = 0$). The negative relative travel-time change indicate that after AV introduction, HDV travel times increase. AVs, on the other hand, experience shorter travel times compared to those achieved by the human drivers they replaced at episode 300 (see Figure 3(b, e) for $\beta = 0$). When marginal cost is incorporated into the AV reward ($\beta = 0.1$ for Ingolstadt and

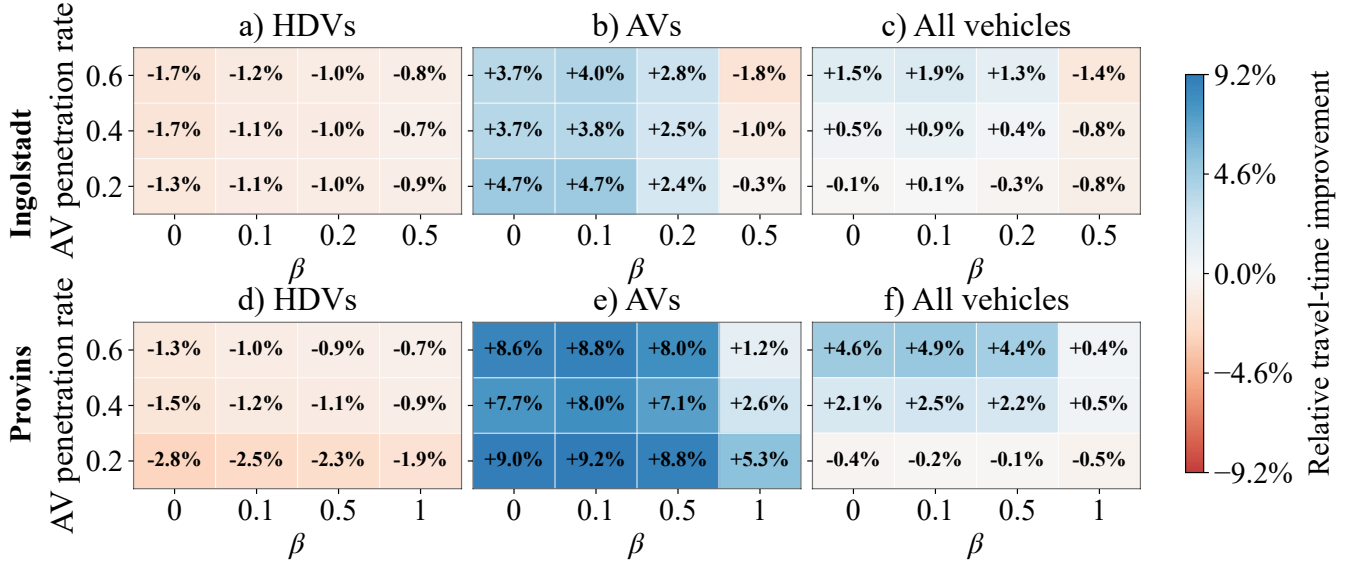


Figure 3: Average agent relative travel time change, computed as the difference between travel time before AV introduction and travel time during the testing phase, normalized by travel time before AV introduction. Negative values indicates the vehicle is disadvantaged after AV introduction, and a positive indicate the vehicle is better off after AV introduction. Values are averaged across three experimental seed replications, with standard deviations reported in the Supplementary Material.

Provins), the HDV relative travel time change increases relative to $\beta = 0$, indicating that HDVs experience shorter travel times than under selfish AV routing. At the same time, AVs experience slightly shorter travel times on average, showing that **accounting for the system does not necessarily come at the expense of AV travel times.**

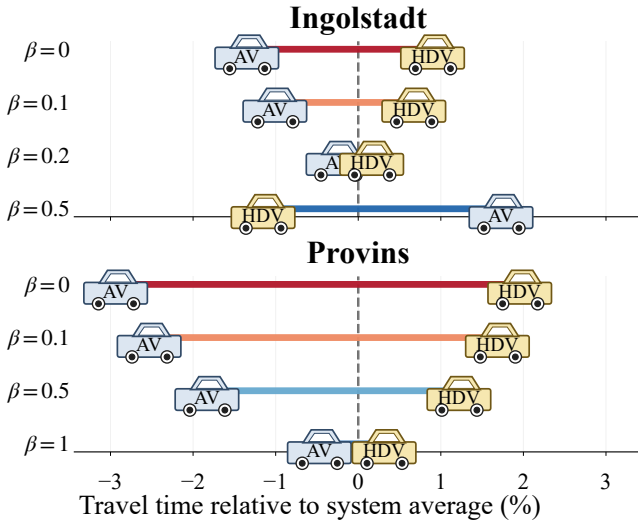


Figure 4: Change in average AV, and HDV travel times at 40% AV penetration in Ingolstadt at various weights of social component β in the reward. Negative values indicate shorter travel times compared to the system average.

We then observe that as β increases, e.g., $\beta = 0.2, 0.5$ in Ingolstadt, and $\beta = 0.5, 1$ in Provins, the relative travel-time change on HDVs increases, indicating that HDVs experience shorter travel times. However, this comes at the cost of longer AV travel times. Increasing β beyond a certain value worsens both AV and overall system travel times (see Ingolstadt, $\beta = 0.5$, Provins $\beta = 1$). **This indicates that placing too much weight on the social component can disadvantage**

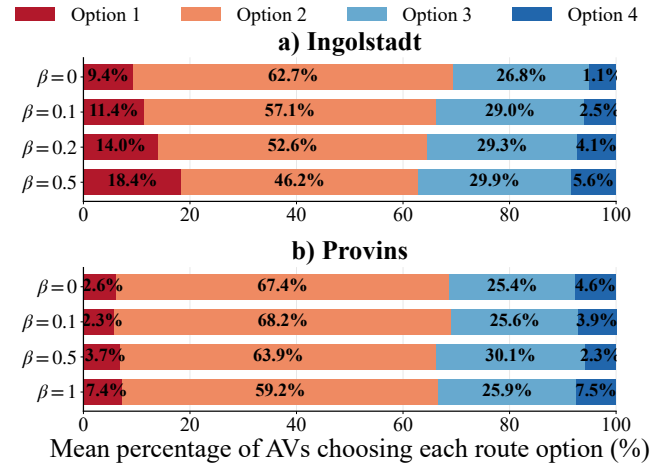


Figure 5: Mean percentage of AVs choosing routes from respective groups averaged over the 100 days of the testing phase on the Ingolstadt (a) and Provins (b) networks for a 40% AV penetration rate.

the system and individual AVs.

Figure 4 shows that for $\beta = 0.1, 0.2$ in Ingolstadt, the gap between HDV and AV average travel times relative to the system average narrows, indicating that AVs, by accounting for their impact on the system, can reduce the travel-time gap between the two groups. At $\beta = 0.5$, however, the HDV ratio decreases while the AV ratio increases, indicating that AVs placing greater weight on system performance can benefit HDVs relative to the system average while making AVs worse off. In Provins, increasing β from 0.1 to 1 progressively narrows the travel-time gap between AVs and HDVs, with the two groups approaching the system average at $\beta = 1$.

Differences in HDV, AV, and system travel times across values of β are accompanied by changes in AV route choices, as shown in Figure 5. As the weight on marginal cost increases, the mean percentage of AVs choosing routes from respective groups shifts from the one observed under selfish routing. In

Table 1: Selective marginal cost reward shaping, specifically when 40% or 50% of the most impactful AVs are rewarded with the marginal cost, can achieve relative travel-time changes (see Figure 3) close to those obtained when all AVs receive the marginal cost reward, while reducing the number of counterfactual simulations. The results are reported for the Ingolstadt traffic network with 40% AV penetration. The AV subgroups column reports the relative travel-time change of selfish AVs (Selfish) and AVs receiving the marginal cost reward (MC). MC denotes marginal cost. Values are reported as mean $\pm$ standard deviation across experimental replications.

Behavior	MC AVs	Counterfactual simulations	HDDVs	AVs	AV subgroups	All
$\beta = 0.0$						
Selfish	–	–	-1.66 ± 0.19	$+3.74 \pm 0.75$	–	$+0.50 \pm 0.41$
$\beta = 0.1$						
MC	100%	4,140k	-1.06 ± 0.17	$+3.79 \pm 0.71$	MC: $+3.79 \pm 0.71$	$+0.88 \pm 0.39$
Selective MC	50%	2,070k	-1.17 ± 0.18	$+3.78 \pm 0.72$	MC: $+3.51 \pm 0.74$ Selfish: $+4.05 \pm 0.70$	$+0.81 \pm 0.39$
Selective MC	40%	1,660k	-1.25 ± 0.17	$+3.78 \pm 0.73$	MC: $+3.30 \pm 0.75$ Selfish: $+4.10 \pm 0.72$	$+0.76 \pm 0.38$
Selective MC	10%	420k	-1.50 ± 0.19	$+3.78 \pm 0.74$	MC: -0.18 ± 0.70 Selfish: $+4.21 \pm 0.74$	$+0.61 \pm 0.40$
Selective MC (random)	10%	420k	-1.60 ± 0.18	$+3.75 \pm 0.72$	MC: $+3.45 \pm 0.70$ Selfish: $+3.78 \pm 0.72$	$+0.54 \pm 0.39$

Ingolstadt, for example, the share of AVs choosing routes from the 2nd group decreases from $\beta = 0$ to $\beta = 0.5$, while the shares choosing Routes from groups 1, 3, and 4 increase. Here, Options 1-4 correspond to route options or groups rather than physical routes. In Ingolstadt, each of the 306 unique OD pairs has four candidate routes, which correspond to 1224 distinct route alternatives.

4.3 Performance of the selective marginal cost

We evaluate selective marginal cost reward shaping on the Ingolstadt traffic network at $\beta = 0.1$, where marginal cost improves travel times for AVs, HDDVs, and the system compared with selfish AV routing. Our results show that applying this reward only to a subset of AVs yields relative travel-time changes (see Figure 3) for HDDVs, AVs, and the system that are close to those obtained when all AVs receive the marginal cost reward, while requiring fewer counterfactual simulations.

Figure 6 shows the cumulative distribution of the historical impact (H_i see Equation 7) of each of the 1035 vehicles on the Ingolstadt traffic network. Most vehicles have impact scores concentrated around 30. A smaller fraction of vehicles show lower or higher impact, indicating that vehicles differ in how strongly their route choices are associated with total system travel time.

Table 1 reports the relative travel-time change for HDDVs, AVs, and all vehicles in the system, as shown in Figure 3, when a fraction of the most impactful AVs receive the marginal cost reward. Specifically, when marginal cost is applied to the 50% and 40% most impactful AVs, HDDV travel times improve by 0.5% and 0.4% compared to selfish AV routing, respectively, around 0.1% and 0.2% below the improvement obtained when marginal cost is applied to all AVs. We then tested a smaller fraction, specifically the 10% most impactful AVs, and found that HDDV travel times still improve by 0.16% compared to selfish AV routing, although this is about 0.45% below the case when marginal cost is applied to all AVs. Thus, **applying selective marginal cost to either 40% or 50% of the most impactful AVs could approximate the results of applying**

marginal cost to all AVs, retaining approximately 68% and 82% of the improvement in HDDV relative travel-time change over selfish AV routing, respectively, while reducing the number of counterfactual simulations.

To see if rewarding a fraction of the most impactful AVs based on H_i is better than choosing a random fraction of AVs in the system. In the *Selective MC (random)* row of Table 1, we compare relative travel-time changes for HDDVs, AVs, and all vehicles in the system when we apply marginal cost to a randomly selected fraction of AVs instead of the most impactful ones. We find a 0.1% improvement in relative travel-time change for HDDVs when we apply marginal cost to the 10% most impactful AVs instead of a random 10% of AVs in the system.

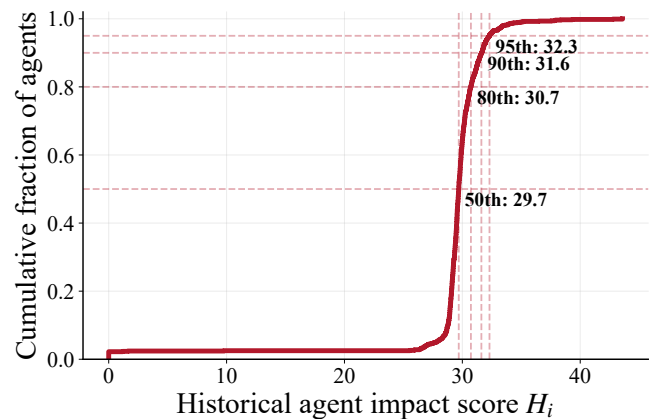


Figure 6: Empirical cumulative distribution of historical agent impact scores H_i retrieved from dataset $\mathcal{D}$ of 800 thousand joint actions sampled in Ingolstadt. Dashed lines indicate the 50th, 80th, 90th, and 95th percentiles.

For the Ingolstadt network at a 40% AV penetration rate (around 414 AVs), 10k training iterations require about 4,140k counterfactual simulations. When we compute the marginal cost reward for 40% of the AVs, this reduces to about 1,660k

counterfactual simulations. We then need to account for the additional 800k joint-action samples of the dataset $\mathcal{D}$ used to identify the AVs with the largest system impact (see Figure 6), for a total of approximately 2,456k additional simulation runs. Thus, selective marginal cost reward shaping can reduce the additional simulation runs from 4 to 2.4 million, about a 40% reduction, while achieving an average HDV relative travel-time change of about 0.2 percentage points below that obtained when marginal cost reward shaping is applied to all AVs.

In the *AV subgroups* column of Table 1, we compare relative travel-time changes of AVs receiving the selfish reward (*Selfish*) and those receiving the marginal cost reward (*MC*). When the 50% or 40% most impactful AVs receive the marginal cost, MC AVs experience smaller relative travel-time change than selfish AVs. This difference is more pronounced when marginal cost is applied to the 10% most impactful AVs, where MC AVs experience a relative travel-time change of -0.18% , while selfish AVs experience a $+4.21\%$ change. Thus, AVs that account for their impact on the system may be individually worse off, experiencing smaller relative travel time changes than selfish AVs.

5 Conclusion

AVs will increasingly operate in complex, capacitated urban traffic networks, where road users compete for limited road capacity, and each user’s route choices can affect others’ travel times. AVs may use a range of algorithms, with MARL being a promising one, to learn routing strategies that may differ from, and be more efficient than, those typically made by HDVs, potentially resulting in shorter AV travel times. Understanding how MARL AV routing can influence traffic performance is crucial to understanding how this emerging technology, AVs, may shape the performance of future urban transportation systems.

In this work, we study social AV route choice in mixed-autonomy traffic environments shared with HDVs, where AVs use MARL to learn routing policies. **To reduce the travel time costs AV routing imposes on other road users, we incorporate marginal cost into the AV reward.** Marginal cost is a counterfactual defined as the difference in travel times for all vehicles in the system when a given AV is present versus absent from the simulation. This formulation lets us include in MARL training the trade-off between minimizing an AV’s own travel time and reducing its impact on the system.

Our results show that incorporating marginal cost into the AV reward can mitigate increases in HDV travel times while also reducing AV travel times relative to selfish AV routing, indicating that **AVs that account for their impact on the system can improve travel times for both HDVs, AVs, and the system.** However, placing greater weight on marginal cost can further improve HDV travel times at the expense of AVs and worsen total system performance. For example, at 40% AV penetration in Ingolstadt, increasing β from 0 to 0.1 improves HDV, AV, and system travel time relative to the travel times before AVs were introduced by 0.6, 0.1, and 0.4 percentage points, respectively. Increasing β to 0.2 yields an additional 0.1 improvement in travel time change for HDVs but reduces the relative travel time change for AVs and the system by 1.3 and 0.5 percentage points, respectively, relative to $\beta = 0.1$.

Since computing the marginal cost requires an additional

counterfactual simulation for each AV, we propose selective marginal cost reward shaping to reduce this computational burden. We find that assigning the marginal cost reward to a fraction of the most impactful AVs yields improvement in relative travel-time change for HDVs, AVs, and the system that are close to those obtained when all AVs receive the marginal cost reward. In the Ingolstadt traffic network, rewarding the 40% most impactful AVs reduces the number of additional simulations by around 40%, while the HDV relative travel time change is around 0.2 percentage points lower than when all AVs receive the marginal cost reward.

Overall, our findings show that AV routing algorithms can influence not only AV travel times but also how travel time costs are distributed across road users. AV routing policies that optimize only individual AV travel times may systematically disadvantage human drivers. By showing that marginal cost reward shaping can reduce these effects, our results highlight the importance of designing AV routing algorithms that account for AVs’ impact on shared transportation systems.

References

- [1] Ahmet Onur Akman, Anastasia Psarou, Zoltán György Varga, Grzegorz Jamróz, and Rafał Kucharski. Impact of collective behaviors of autonomous vehicles on urban traffic dynamics: A multi-agent reinforcement learning approach. In *Seventeenth European Workshop on Reinforcement Learning*, 2024. URL <https://openreview.net/forum?id=88zP8xh5D2>.
- [2] Ahmet Onur Akman, Anastasia Psarou, Michał Hoffmann, Łukasz Gorczyca, Łukasz Kowalski, Paweł Gora, Grzegorz Jamróz, and Rafał Kucharski. Urb – urban routing benchmark for rl-equipped connected autonomous vehicles. In *Advances in Neural Information Processing Systems*, 2025.
- [3] Ahmet Onur Akman, Anastasia Psarou, Łukasz Gorczyca, Zoltán György Varga, Grzegorz Jamróz, and Rafał Kucharski. Routerl: Multi-agent reinforcement learning framework for urban route choice with autonomous vehicles. *SoftwareX*, 31:102279, 2025. ISSN 2352-7110. doi: <https://doi.org/10.1016/j.softx.2025.102279>. URL <https://www.sciencedirect.com/science/article/pii/S2352711025002468>.
- [4] W. Brian Arthur. Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, 1994. ISSN 00028282. URL <http://www.jstor.org/stable/2117868>.
- [5] James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, December 2021.
- [6] Martin J. Beckmann, C. B. McGuire, C. B. Winsten, and T. C. Koopmans. *Studies in the Economics of Transportation*. Yale University Press, New Haven, 1956. doi: 10.2307/2227641.

- [7] Enide A. I. Bogers, Michel Bierlaire, and Serge P. Hoogendoorn. Modeling learning in route choice. *Transportation Research Record*, 2014(1):1–8, 2007. doi: 10.3141/2014-01.
- [8] Ennio Cascetta. *Transportation Systems Analysis: Models and Applications*, volume 29 of *Springer Optimization and Its Applications*. Springer, 2009. ISBN 978-0-387-75857-2. doi: 10.1007/978-0-387-75857-2. URL <https://link.springer.com/book/10.1007/978-0-387-75857-2>.
- [9] Jianming Chen, Yawen Wang, Junjie Wang, Xiaofei Xie, Jun Hu, Qing Wang, and Fanjiang Xu. Understanding individual agent importance in multi-agent system via counterfactual reasoning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’25/IAAI’25/EAAI’25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i15.33733. URL <https://doi.org/10.1609/aaai.v39i15.33733>.
- [10] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. How much can taxes help selfish routing? *Journal of Computer and System Sciences*, 72(3):444–467, 2006. ISSN 0022-0000. doi: <https://doi.org/10.1016/j.jcss.2005.09.010>. URL <https://www.sciencedirect.com/science/article/pii/S0022000005001121>. Network Algorithms 2005.
- [11] Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS ’14, page 165–172, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381.
- [12] Lawrence Christopher Duncan, Thomas Kjær Rasmussen, David Paul Watling, Laurent Cazor, and Otto Anker Nielsen. Identification of local detouredness and correlation effects in a bounded route choice model: solution & estimation in large-scale networks. *Transportmetrica A Transport Science*, 2025. ISSN 2324-9935. doi: <https://doi.org/10.1080/23249935.2025.2543467>. URL <https://www.sciencedirect.com/science/article/pii/S2324993525001022>.
- [13] Rafael Grunitzki, Guilherme de Oliveira Ramos, and Ana L C Bazzan. Individual versus difference rewards on reinforcement learning for route choice. In *2014 Brazilian Conference on Intelligent Systems*, pages 179–184. IEEE, 2014.
- [14] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Collective explainable ai: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *Comp. Intell. Mag.*, 17(1):59–71, February 2022. ISSN 1556-603X. doi: 10.1109/MCI.2021.3129959. URL <https://doi.org/10.1109/MCI.2021.3129959>.
- [15] Ron Holzman and Nissan Law-Yone. Strong equilibrium in congestion games. *Games and economic behavior*, 21(1-2):85–101, 1997.
- [16] Grzegorz Jamróz, Ahmet Onur Akman, Anastasia Psarou, Zoltán György Varga, and Rafał Kucharski. Social implications of coexistence of cavs and human drivers in the context of route choice. *Scientific Reports*, 15(1):6768, 2025.
- [17] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990. ISSN 0378-3758. doi: [https://doi.org/10.1016/0378-3758\(90\)90122-B](https://doi.org/10.1016/0378-3758(90)90122-B). URL <https://www.sciencedirect.com/science/article/pii/037837589090122B>.
- [18] Dong Ki Kim, Miao Liu, Matthew D Riemer, Chuangchuan Sun, Marwa Abdulhai, Golnaz Habibi, Sebastian Lopez-Cot, Gerald Tesauero, and Jonathan How. A policy gradient algorithm for learning to learn in multiagent reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5541–5550. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/kim21g.html>.
- [19] Daniel A. Lazar, Erdem Bıyık, Dorsa Sadigh, and Ramtin Pedarsani. Learning how to dynamically route autonomous vehicles on shared roads, 2021. URL <https://arxiv.org/abs/1909.03664>.
- [20] Dawei Li, Siqi Feng, Yuchen Song, Xinjun Lai, and Shlomo Bekhor. Asymmetric closed-form route choice models: Formulations and comparative applications. *Transportation Research Part A: Policy and Practice*, 171:103627, 2023. ISSN 0965-8564. doi: <https://doi.org/10.1016/j.tra.2023.103627>. URL <https://www.sciencedirect.com/science/article/pii/S0965856423000472>.
- [21] Silas C. Lobo, Stefan Neumeier, Evelio M. G. Fernandez, and Christian Facchi. Intas – the ingolstadt traffic scenario for sumo, 2020. URL <https://arxiv.org/abs/2011.11995>.
- [22] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. URL <https://elib.dlr.de/124092/>.
- [23] A. Mas-Colell, M.D. Whinston, and J.R. Green. *Microeconomic Theory*. Oxford student edition. Oxford University Press, 1995. ISBN 9780195073409.

URL <https://books.google.pl/books?id=KGtegVXqD8wC>.

- [24] Deepak K Merchant and George L Nemhauser. Optimal-ity conditions for a dynamic traffic assignment model. *Transportation Science*, 12(3):200–207, 1978.
- [25] Srinivas Peeta and Hani S. Mahmassani. System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Annals of Operations Research*, 60:81–113, 1995. doi: 10.1007/BF02031941.
- [26] A. C. Pigou. The economics of welfare. *The Economic Journal*, 43(170):329–330, 06 1933. ISSN 0013-0133. doi: 10.2307/2224491.
- [27] Scott Proper and Kagan Tumer. Modeling difference rewards for multiagent learning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3, AAMAS '12*, page 1397–1398, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0981738133.
- [28] Anastasia Psarou, Ahmet Onur Akman, Łukasz Gorczyca, Michał Hoffmann, Grzegorz Jamróz, and Rafał Kucharski. Position: Collaboration between the city and the machine learning community is crucial to efficient autonomous vehicles routing. In *AI4GOOD Workshop at ICML 2026*, 2026. URL <https://openreview.net/forum?id=24C18nSM4d>.
- [29] Anastasia Psarou, Łukasz Gorczyca, Dominik Gaweł, and Rafał Kucharski. Autonomous vehicles need social awareness to find optima in multi-agent reinforcement learning routing games. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '26*, page 3417–3419, Richland, SC, 2026. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400723179. doi: 10.65109/FTHN6981. URL <https://doi.org/10.65109/FTHN6981>.
- [30] Gabriel de O. Ramos, Roxana Rădulescu, Ann Nowé, and Anderson R. Tavares. Toll-based learning for minimising congestion under heterogeneous preferences. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, page 1098–1106, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.
- [31] Henry Scheffé. Experiments with mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):344–360, 07 1958. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1958.tb00299.x. URL <https://doi.org/10.1111/j.2517-6161.1958.tb00299.x>.
- [32] Yosef Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Englewood Cliffs, 1985.
- [33] Chenxiao Song and Reiichiro Kawai. Sampling and change of measure for monte carlo integration on simplices. *Journal of Scientific Computing*, 98:64, 2024. doi: 10.1007/s10915-024-02461-0.
- [34] Ming Tan. *Multi-agent reinforcement learning: independent vs. cooperative agents*, page 487–494. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1558604952.
- [35] J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis Santos, Rodrigo Perez, Caroline Horsch, Clemens Dieffendahl, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. Pettingzoo: Gym for multi-agent reinforcement learning, 2021. URL <https://arxiv.org/abs/2009.14471>.
- [36] Luiz A. Thomasini, Lucas N. Alegre, Gabriel O. Ramos, and Ana L. C. Bazzan. Routechoiceenv: a route choice library for multiagent reinforcement learning. In *Adaptive and Learning Agents Workshop at AAMAS*, 2023.
- [37] Kagan Tumer and Adrian Agogino. Agent reward shaping for alleviating traffic congestion. 01 2006.
- [38] J G WARDROP. Road paper. some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1(3):325–362, May 1952. ISSN 1753-7789. doi: 10.1680/ipeds.1952.11259. URL <http://dx.doi.org/10.1680/ipeds.1952.11259>.
- [39] J. G. Wardrop. Road paper. some theoretical aspects of road traffic research. 1952. URL <https://api.semanticscholar.org/CorpusID:131127018>.
- [40] David Paul Watling, Thomas Kjær Rasmussen, Carlo Giacomo Prato, and Otto Anker Nielsen. Stochastic user equilibrium with a bounded choice model. *Transportation Research Part B: Methodological*, 114:254–280, 2018. ISSN 0191-2615. doi: <https://doi.org/10.1016/j.trb.2018.05.004>. URL <https://www.sciencedirect.com/science/article/pii/S0191261517304083>.
- [41] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL <https://doi.org/10.1007/BF00992696>.