Multi-Agent Reinforcement Learning for Order Assignment and Payment Setting on Food-Delivery Platforms: The Implicit Algorithmic Biases

Zijian Zhao[1] and Sen Li[1]

[1]Department of Civil and Environmental Engineering, Hong Kong University of Science and Technology, Hong Kong, China

Abstract

This paper examines discriminatory order-assignment and payment-setting strategies for on-demand food-delivery platforms. We consider a platform that maximizes its profits by strategically bundling orders, assigning them to couriers, and setting personalized payments to couriers based on individual behavioral data accrued from past interactions with the platform. A novel multi-action, multi-agent deep reinforcement learning framework is proposed, where a Double Deep Q-Network is employed to develop discrete order-assignment strategies, and a Proximal Policy Optimization is utilized to determine continuous payment decisions. Our proposed method is validated through a case study using real-world food-delivery data from Hong Kong. Surprisingly, we find that couriers with higher reservation values and, consequently, higher order rejection rates actually receive more orders during peak hours but earn lower wages. The reasons for these counterintuitive results are identified, which expose implicit biases within the discriminatory algorithms employed by profit-maximizing platforms and underscore potential areas for regulatory intervention.

1 Introduction

Recent advancements in data accumulation and reinforcement learning (RL) algorithms have revolutionized the operational dynamics of on-demand food delivery platforms such as FoodPanda, DoorDash, and UberEats. These platforms now leverage sophisticated, opaque RL systems to make personalized order assignment and payment-setting decisions tailored to individual couriers, based on historical work-related data like order acceptance and rejection patterns. While these algorithmic enhancements improve labor management and customer experiences, they raise significant ethical concerns. Differentiated order allocation and payment settings can result in discriminatory profiling, where couriers in the same location and time may face unequal opportunities and payments for identical work, challenging the principle of "equal pay for equal work." Moreover, platforms can exploit these algorithms to identify couriers more reliant on the platform, subtly coercing them into longer hours for minimal pay, exacerbating tensions with gig workers and threatening the sustainability of the sharing economy.

Despite the growing importance of these issues, research on algorithmic discrimination in food delivery platforms remains limited. Most studies, such as (Zhang et al., 2023; Raman et al., 2021), focus on zone-level disparities rather than individual-level discrimination, leaving the interplay between order assignment and payment-setting underexplored. To address this gap, we examine algorithmic labor management in a food delivery market where a platform coordinates customers, couriers, and restaurants via decisions like order bundling, assignment, and personalized payments, using a Markov Decision Process (MDP) tailored to each courier's historical data. This is, to our knowledge, the first study to jointly analyze order bundling, discriminatory order assignment, and personalized payment-setting.

We propose a novel multi-action, multi-agent deep RL framework, integrating Double Deep Q-Network (DDQN) for discrete order assignments and Proximal Policy Optimization (PPO) for continuous payment decisions. The algorithmic framework is validated using real-world food-delivery data from Hong Kong. Surprisingly, couriers with higher reservation values and order rejection rates receive more orders during peak hours, contrary to the expectation that they would be disadvantaged due to higher incentivization costs. Even more counterintuitively, despite receiving more orders and higher payments per trip time, these couriers earn less during peak hours than those with lower reservation values and fewer orders. This arises from the platform's subtle strategy of assigning longer-distance trips to couriers with lower reservation values, indirectly penalizing those with higher reservation values. These findings expose implicit biases in the platform's profit-driven algorithms and highlight areas for potential government regulation.

2 Methodology

This paper examines a platform managing on-demand food-delivery services by coordinating customers, restaurants, and couriers, shown as Fig. 1. At each time step, the platform decides on order bundling, assignment, and payment based on couriers' historical behavioral data (e.g., order acceptance/rejection patterns) and their spatial relationship with orders. Couriers, treated as independent contractors, strategically accept or reject orders based on payment, order attributes, and personal reservation values, where higher values reflect greater selectivity due to less reliance on the platform. Based on this, we formulate an MDP model with courier details (e.g., location, onboard orders) and order details (e.g., origin-destination) as states, order assignment and payment as actions, and total profit as the reward. To find the optimal strategy for the MDP, a DDQN is first employed for order assignment, followed by the use of PPO to determine the payment for each courier-order pair.

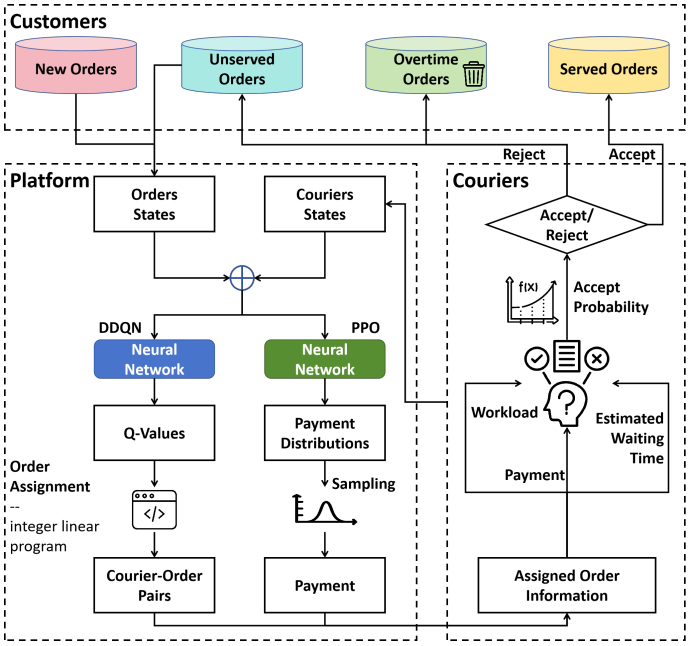


Figure 1: Diagram for the proposed multi-agent reinforcement learning algorithm.

First, we explicitly represent the policy as a joint vector comprising two sub-policies, each corresponding to the discrete and continuous decisions, respectively. Specifically, we define the policy as $\pi = [\pi^A, \pi^P]$, where π^A pertains to order assignment, mapping the system state $s_{i,t}$ to the order assignment decision $\kappa_{i,t}$, and π^P relates to payment-setting, mapping the system state $s_{i,t}$ to the payment-setting decision $p_{i,t}$. In our multi-agent RL scenario, we follow the independent assumption applied in many works (Feng et al., 2022; Wang et al., 2023; Tan, 1993): (i) each agent's reward and state transition probability depend solely on its own states and actions, independent of others; and (ii) all the agents follow a shared policy π. This implies that the global Q-value and V-value can be expressed as the aggregation of the values from each agent. *Due to page limitations, we will only describe the aspects that differ from the standard versions of DDQN and PPO below:*

1) **DDQN for Order Assignment:** When training DDQN, we view PPO as a part of environment or state and only focus on how to determine the optimal order assignment decisions under these fixed payment-setting decisions. In this way, we can define the Q-value for DDQN as:

$$Q_{\pi^A}^{DDQN}(s_{i,t}, \kappa_{i,t}|\pi^P) = \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in K_t} \gamma^\tau \cdot r_{i,t+\tau}(s_{i,t+\tau}, a_{i,t+\tau}) \mid s_{i,t}, \kappa_{i,t}\right] \tag{1}$$

where the Q-value is defined for the courier under a given payment policy π^P. Then the aggregated Q-value for the entire platform is the summation of individual Q-values of each courier. However, since the actions of different agents in our order assignment task are interrelated, as one order can be assigned to a specific courier, we use integer linear program (ILP) to maximize the global Q-value, following Feng et al. (2022).

2) **PPO for Payment Setting:** Next, we consider the use of PPO to derive the optimal payment-setting decisions for the platform. Under a given action of the DDQN (which determines the

courier-order pairs we need to set payment) and the given policy of order assignment, we can then define the V-value of PPO for each individual courier as:

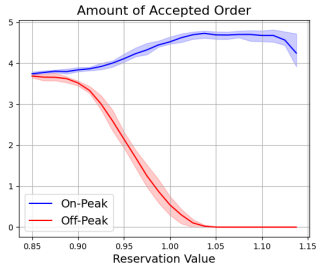
$$V_{\pi^P}^{PPO}(s_{i,t}|\kappa_{i,t},\pi^A) = \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in K_t}\gamma^\tau \cdot r_{i,t+\tau}(s_{i,t+\tau},[\pi^A(s_{i,t+\tau}),\pi^P(s_{i,t+\tau})]) \mid s_{i,t},\kappa_{i,t}\right] \quad (2)$$

where the V-value is defined for the courier under a given order assignment decision. Therefore, it is parameterized by the order assignment decision at time t, denoted as $\kappa_{i,t}$, and the given policy π^A that determines order assignment decisions in the future. In this case, the aggregated V-value for the entire platform is the summation of individual V-values of each courier.

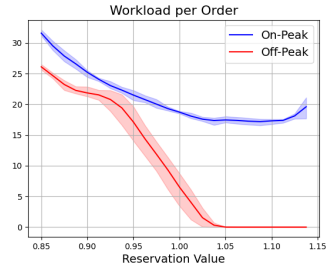
An interesting observation is made: the V-value defined in (2) for PPO is essentially equivalent to the Q-value of DDQN presented in (1), provided the policies in both scenarios are aligned. This equivalence allows us to directly utilize DDQN as the critic in the PPO framework. This approach enhances stability compared to training a critic for PPO independently. When the policy of DDQN changes, the V-value for PPO also adjusts accordingly. If the critic cannot keep pace with this change, the update direction for the actor may be incorrect, increasing the risk of model collapse.

3 Numerical Experiments

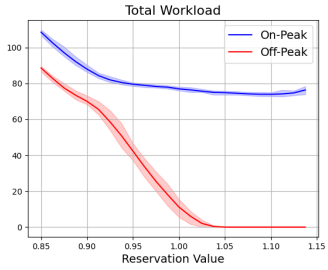
To validate our proposed reinforcement learning framework, we applied the model and solution algorithm to a real-world case study in Hong Kong, utilizing actual food delivery data. The experiment is illustrated in Fig. 2. Due to page limitations, we primarily analyze the results for the peak period. Intuitively, couriers with higher reservation values might be disadvantaged in order allocation because the platform would need to offer higher rates to incentivize these couriers to accept orders. Contrary to this expectation, surprisingly, Figure 2a reveals that during the peak hour, the number of orders assigned to couriers is an increasing function of the courier's reservation value. This suggests that couriers with higher reservation values actually receive more orders during the peak hours, which contradicts our initial intuition. Furthermore, what's even more surprising is that although couriers with higher reservation values receive more order assignments and benefit from higher payments per unit of trip time, their earnings during the peak hours are actually lower than those of couriers with lower reservation values who receive fewer orders, as evidenced by Figure 2f.



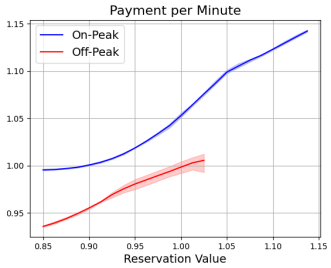
(a) Amount of orders accepted by the courier per hour



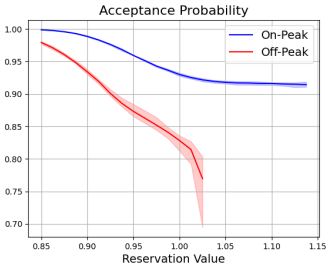
(b) Trip time (minutes) per order assigned to the courier



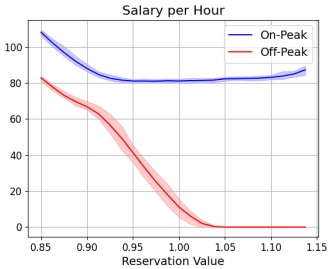
(c) Total trip time (minutes) of all order assigned to the courier per hour



(d) Payment per minute for the orders assigned to the courier



(e) Order acceptance probability



(f) Salary per hour for the courier

Figure 2: Experiment results under difference reservation values of the couriers.

To explain the counter-intuitive observation that couriers with higher reservation values have a higher number of orders yet lower hourly earnings, we note that their earnings per trip must be lower. This paradox arises even though the payment per unit time is higher for couriers with higher reservation values. since the payment per trip is calculated as the payment per unit time multiplied by the trip time, this suggests that couriers with higher reservation values are more likely to receive orders with shorter distances. To validate this hypothesis, we examined the average trip time and the total trip time served by couriers as a function of their reservation values. Figures 2b and 2c clearly show that the average trip distance is a sharply decreasing function of the couriers' reservation values. This reveals an interesting strategy by the platform to differentiate order allocation among couriers: rather than discriminatory order allocation based on volume, the platform implements a more subtle discriminatory strategy where longer distance trips are preferentially allocated to couriers with lower reservation values, which opaquely penalizes couriers with higher reservation values, causing them to deliver more orders but earn less. This strategy is particularly rational for the platform during peak times, when the demand is high and supply is limited. Long-distance orders are economically more valuable and generate greater revenue for the platform, making it crucial that these orders are accepted as a priority. Therefore, couriers with lower reservation values, who are more likely to accept orders, are strategically prioritized for these long-distance orders to minimize the risk of losing high-value customers. It is important to note that such implicit bias in order allocation is not explicitly designed by the platform but is deeply embedded in the machine learning algorithms used. This phenomenon, often referred to in existing literature as "algorithmic labor manipulation" (Yang et al., 2024; Liu et al., 2024), highlights the complex ways in which data-driven decision-making can inadvertently lead to biased or unexpected outcomes in labor markets.

4 Conclusions

This paper presents a novel reinforcement learning framework to assess algorithmic labor management in on-demand food delivery platforms. We analyze a profit-driven platform employing discriminatory order assignments and personalized payments, optimized via a multi-action, multi-agent approach using DDQN for discrete order strategies and PPO for continuous payment decisions. Validated with real-world Hong Kong data, our study reveals that in peak hour, the platform subtly allocates longer-distance trips to couriers with lower reservation values, causing those with higher reservation values to receive more orders but earn less due to shorter trips. Future research will explore courier scheduling and regulatory policies to address ethical concerns in algorithmic labor management.

References

Feng, S., Duan, P., Ke, J., & Yang, H. (2022). Coordinating ride-sourcing and public transport services with a reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, *138*, 103611.

Liu, Y., Zheng, Y., Zhang, S., & Liu, L. T. (2024). Evaluating fairness in black-box algorithmic markets: A case study of ride sharing in chicago. *arXiv preprint arXiv:2407.20522*.

Raman, N., Shah, S., & Dickerson, J. (2021). Data-driven methods for balancing fairness and efficiency in ride-pooling. *arXiv preprint arXiv:2110.03524*.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning* (pp. 330–337).

Wang, D., Wang, Q., Yin, Y., & Cheng, T. (2023). Optimization of ride-sharing with passenger transfer via deep reinforcement learning. *Transportation Research Part E: Logistics and Transportation Review*, *172*, 103080.

Yang, Y., Umboh, S. W., & Ramezani, M. (2024). Freelance drivers with a decline choice: Dispatch menus in on-demand mobility services for assortment optimization. *Transportation Research Part B: Methodological*, *190*, 103082.

Zhang, X., Varakantham, P., & Jiang, H. (2023). Future aware pricing and matching for sustainable on-demand ride pooling. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 14628–14636).