Adaptive Regulation via Dual-Layer Evolution (ARDE): A Multi-Agent Approach to Balancing Efficiency, Fairness, and Diversity in Crowdsourced Platforms

XuWen Zhang[1,2,3], **Xiao Xue**[1,2,3*], **Xia Xie**[4], **Qun Ma**[1,2,3]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Tianjin Key Laboratory of Healthy Habitat and Smart Technology,Tianjin University, Tianjin, China
[3] Laboratory of Computation and Analytics of Complex Management Systems, Tianjin University, Tianjin, China
[4] School of Computer Science and Technology,Hainan University, Hainan, China
1023244032@tju.edu.cn, jzxuexiao@tju.edu.cn*, shelicy@hainanu.edu.cn, 1023244018@tju.edu.cn

Abstract

Crowdsourced delivery platforms (e.g., Meituan, Uber Eats, DoorDash) have become vital infrastructure in urban logistics, yet their competitive order-grabbing mechanisms often lead to strategy homogenization, inefficiency, and income inequality. This paper presents ARDE (Adaptive Regulation via Dual-layer Evolution), an evolutionary governance framework that integrates individual reinforcement learning with adaptive platform-level regulation. The outer agent dynamically generates governance signals based on system diagnostics (strategy entropy, Gini coefficient, completion rate), while inner agents employ Diffusion Q-Learning guided by a language-model-driven reward shaping module to promote fairness and strategy diversity. Experiments on real-world datasets show that ARDE achieves stable diversity (0.997 ± 0.184), reduces inequality, and maintains high efficiency. Further comparison (ARDE-PPO vs. MAPPO) confirms that its advantages stem from explicit hierarchical governance rather than algorithmic coincidence. Overall, ARDE offers a scalable and interpretable paradigm for reconciling individual rationality with collective welfare in gig economies and other multi-agent socio-technical systems.

Introduction

Crowdsourced delivery platforms (e.g., Meituan, Uber Eats, DoorDash) have become indispensable infrastructure for urban logistics, orchestrating millions of daily orders through decentralized, incentive-driven mechanisms(Alnaggar et al., 2021). Unlike centralized dispatching systems, riders independently select orders based on expected payoffs and spatial proximity(Shiri et al., 2025). However, this self-interested competition frequently escalates into algorithmic rat races—a phenomenon characterized by strategic homogenization, pronounced income inequality, and efficiency losses of up to 20% during peak hours (Sun, 2019). Existing interventions—such as short-term bonuses, rule-based constraints, and conventional multi-agent reinforcement learning (MARL) approaches—have proven insufficient to mitigate this dynamic degradation(Xue et al., 2022). Fundamentally, this challenge poses a critical question for AI governance: how can large-scale multi-agent systems be effectively regulated to align individual rationality with collective welfare in continuously evolving environments?

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Despite the increasing prominence of this challenge in real-world contexts, existing approaches have largely failed to mitigate the systemic degradation of crowdsourced delivery platforms. In the field of multi-agent reinforcement learning (MARL), algorithms such as PPO (Yu et al., 2022), QMIX (Hady et al., 2025), and MAPPO optimize individual rewards but inadvertently intensify policy convergence (Liu et al., 2023), resulting in phenomena like riders clustering around high-value orders and, consequently, a substantial decline in overall system efficiency.Game-theoretic frameworks—including the price of anarchy and evolutionary game theory—offer valuable theoretical insights into the origins of collective inefficiency but lack actionable implementations for dynamic, learning-based AI ecosystems (Pandit et al., 2019).Similarly, research from social sciences on platform governance and algorithmic fairness has proposed static pricing schemes or fairness constraints; however, such approaches fail to adapt to real-time rider interactions, do not effectively reduce income inequality—as evidenced by persistently high Gini coefficients—and overlook the issue of overwork (Rudnik et al., 2023).Consequently, developing methods to steer large populations of self-interested agents toward globally cooperative equilibria under limited intervention and incomplete information represents one of the most pressing and complex open problems at the intersection of AI and socio-technical system research.

To address these limitations, we propose Adaptive Regulation via Dual-layer Evolution (ARDE) — a novel governance framework that models the platform as a learning-enabled Outer Agent guiding self-interested Inner Agents toward globally cooperative equilibria. ARDE achieves a dynamic balance between efficiency and fairness through adaptive policy shaping, rule generation, and information diffusion.

- Diffusion Q-Learning: Through an experience diffusion mechanism among neighboring policies, this approach enhances policy exploration capabilities, preventing collapse into singular strategies and extreme involution.
- Large Language Model-Constrained Reward Shaping (LLM-Constrained Reward Shaping): The outer-layer platform agent, based on current policy structures and fairness metrics, leverages large language models (e.g., GPT-4o) to generate interpretable incentive rules, guiding group evolution through reward shaping.

- Multi-Objective Optimization Mechanism: Incorporating individual, system fairness (e.g., Gini coefficient), and policy diversity (e.g., policy entropy) into a unified optimization objective, the platform adjusts reward structures to steer individual evolutionary trajectories, achieving effective coordination between self-interested behaviors and societal goals(Xue et al., 2024).

To evaluate the effectiveness of ARDE, we develop a high-fidelity, order-driven simulation environment grounded in the publicly available LaDe-D dataset, capturing the long-term evolutionary dynamics of delivery riders under diverse strategy-guidance mechanisms. The LaDe-D dataset contains rich multidimensional information—such as order acceptance records, delivery trajectories, location timestamps, and incentive schemes—allowing accurate reconstruction of strategic decision-making and behavioral feedback within platform ecosystems.Experimental results show that ARDE substantially outperforms all baseline methods across multiple dimensions: strategy diversity (0.997 vs. baseline average 0.769, a +29.6% improvement), fairness (Gini coefficient change rate of -1.3%, the only method achieving negative growth), and system efficiency (order completion rate of 0.411 and revenue of 6.86 CNY/hour).These findings demonstrate ARDE's capacity to proactively steer group-level evolutionary trajectories, establishing a scalable pathway toward an integrated "algorithm–mechanism–society" paradigm for intelligent platform governance.

Related Work

The aggregation of individually rational behaviors leading to collectively suboptimal outcomes has emerged as a central concern in research on the platform economy(Xiao et al., 2023). Existing approaches, however, encounter three fundamental limitations in effectively addressing this challenge:

Individual Optimization Orientation

Traditional approaches primarily emphasize self-interested optimization at the individual level, without incorporating mechanisms for collective coordination (Gavrilets et al., 2021). In the fields of group intelligence and platform governance, decentralized mechanisms are widely adopted to improve task allocation and route optimization (Zhao et al., 2022). For example, platforms such as Meituan and Didi employ algorithms based on ant colony optimization to enhance order-assignment efficiency, while recent studies have introduced behavioral perturbations to strengthen system resilience (Lunansky et al., 2024). However, these approaches commonly assume homogeneous rider behavior(Kang et al., 2024), neglecting the inherent heterogeneity in rider preferences and geographic conditions—ultimately leading to policy convergence and reduced strategic diversity.

Static Constraint Dilemma

Existing mechanism-design approaches face significant challenges in adapting to dynamically evolving environments (Saleh et al., 2024). While mechanism design and game theory provide a solid theoretical foundation for platform governance (Xu et al., 2024), their practical applicability remains limited. The Price of Anarchy framework demonstrates that individually rational decisions can lead to systemic inefficiency(Zhong et al., 2024), whereas evolutionary game theory further reveals that stable equilibria may become locked in long-term suboptimal states. In real-world crowdsourcing platforms, incentive schemes such as fixed subsidies or mandatory performance quotas (Wang et al., 2016) can temporarily improve income distribution but fail to accommodate the continual evolution of agent behavior and environmental dynamics(Lu et al., 2021).

Strategy Convergence Issue

Multi-agent systems are inherently susceptible to local optima, resulting in efficiency losses and exacerbated income inequality. Multi-agent reinforcement learning (MARL) has been extensively employed in crowdsourcing scenarios to optimize agent-level decision policies. Classical algorithms such as PPO, QMIX, and MAPPO aim to achieve individual or cooperative optimality primarily through local policy updates or value-function decomposition (Liu et al., 2023). Although decentralized MARL methods offer improved scalability, they often exhibit limited coordination and structural rigidity in highly heterogeneous environments with competing interests(Xue et al., 2023a). Moreover, most existing algorithms remain narrowly focused on maximizing individual returns, making it difficult for the overall system to escape non-cooperative equilibria and achieve collective efficiency(Xue et al., 2019).

In summary, existing research has yet to effectively reconcile dynamic adaptability, strategy diversity, and group fairness—highlighting the need for a novel governance mechanism capable of aligning individual rationality with collective optimality. To this end, we propose the Adaptive Regulation via Dual-layer Evolution (ARDE) algorithm, an innovative framework designed to overcome these limitations.

Problem Formulation

To systematically characterize the phenomenon of group degradation driven by individual rationality, this paper models the evolutionary trap where individual rationality leads to collective irrationality:

System State and Strategy Space Definition

To analyze the phenomenon of group-level degradation, we construct a multi-agent delivery system. At time step t, the system state S_t is defined as the set of individual states of all couriers:

$$S_t = \{s_i^t\}_{i=1}^N; \tag{1}$$

where N denotes the total number of couriers, and each individual state s_i^t consists of the geographical location (x_i^t, y_i^t), the current strategy label $\pi_i^t \in \mathcal{P}$, the historical reward trajectory $\{R_i^{t-k}\}_{k=1}^K$, and the order completion status O_i^t.

The set of strategies available to couriers, $\mathcal{P}$, includes the following types:

$$\mathcal{P} = \{\pi_1, \pi_2, \pi_3, \pi_4\}; \tag{2}$$

where:

- π_1 (reward_first): A reward-oriented strategy that prioritizes selecting orders with the highest reward per unit time.
- π_2 (balanced): A balanced strategy that considers distance, reward, and time jointly when selecting orders.
- π_3 (speed_max): A speed-oriented strategy that prioritizes orders with shorter delivery distances.
- π_4 (short_first): A short-order strategy that prioritizes orders with the shortest expected delivery time.

Individual Decision-Making Mechanisms and Group Degradation Phenomena

Each courier agent's core objective is to maximize their earnings per unit time (reward per hour). However, their decision-making relies solely on local perspectives and short-term returns(Yu et al., 2024), lacking awareness of global effects, which may lead to a degradation of overall system efficiency(Yu et al., 2025).

At each decision step t, a courier faces an order selection set $\mathcal{A}_t$, aiming to choose the optimal order $a_t^* \in \mathcal{A}_t$ that maximizes individual earnings per unit time:

$$a_t^* = \arg\max_{a \in \mathcal{A}_t} \mathrm{E}\left[\frac{r(a)}{T(a)} \,\middle|\, s_t \right]; \tag{3}$$

where $r(a)$ denotes the expected reward of order a (including subsidies), $T(a)$ is the expected time to complete order a, and s_t represents the current observable state, encompassing the courier's location, available order information, and historical earnings.

In the absence of global regulation and coordination mechanisms, individual courier behavior tends to exhibit "strategy concentration" or "strategy convergence," meaning multiple couriers compete for the same high-value orders, resulting in a loss of strategy diversity. We measure the diversity of strategies within the population using the entropy of the strategy distribution $\mathcal{H}(\pi)$:

$$\mathcal{H}(\pi) = -\sum_a \pi(a) \log \pi(a); \tag{4}$$

Strategy homogenization leads couriers to cluster around similar orders, neglecting long-tail orders, which causes local congestion and reduces overall system efficiency. The system efficiency is measured by the average reward per hour:

$$\eta = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{T_i}; \tag{5}$$

$$G = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |R_i - R_j|}{2N \sum_{i=1}^{N} R_i} \tag{6}$$

Furthermore, competition for high-quality orders exacerbates income inequality. We measure income fairness using the Gini coefficient defined in Equation (6).

Group Objective Function and Evolutionary Traps

Unlike the individual agents, whose primary objective is to maximize their own earnings, the platform regulator aims to ensure the long-term sustainability of the entire system while balancing efficiency, fairness, and strategy diversity. Therefore, the collective objective function should account for three core dimensions: system efficiency, income fairness, and evolutionary stability, and is defined as:

$$\mathcal{L}_{\text{system}} = \eta \; - \; \lambda_1 G \; - \; \lambda_2 \mathcal{H}^{-1}(\pi); \tag{7}$$

By modeling phenomena such as strategy concentration, efficiency degradation, and income inequality, we outline the conflict between individual rationality and collective optimality. In this context, the goal of this work is to introduce a dynamic regulatory mechanism that allows individuals, while pursuing their own optimal outcomes, to be naturally guided towards a global optimum, thereby achieving a synergistic balance among fairness, diversity, and efficiency.

Methodology

The core of the ARDE framework lies in modeling the platform as an Outer Agent and individual couriers as Inner Agents,thereby establishing a two-level collaborative learning architecture,An illustration of the two-level learning architecture is shown below:

- **Inner Agent Level:**Each courier is represented as an autonomous Q-learning agent whose objective is to maximize long-term cumulative earnings (e.g., hourly rewards).
- **Outer Agent Level:**The platform functions as a global regulator that adaptively adjusts system dynamics according to the aggregated strategy distribution and utility metrics—such as strategy entropy and the Gini coefficient. By leveraging diversity incentives, reward shaping, and diffusion mechanisms, the Outer Agent directs the evolutionary process toward collective rationality and global optimality.

The integration of large language models (LLMs) into ARDE is motivated by two key objectives: dynamic adaptivity and decision interpretability. Crowdsourced delivery environments are highly dynamic and stochastic, where static, rule-based governance quickly becomes obsolete. Leveraging their reasoning and generation capabilities, LLMs translate real-time system diagnostics—such as policy entropy, fairness, and completion rates—into context-aware governance rules, enabling truly adaptive regulation.Moreover, transparency and auditability are essential for trustworthy platform governance. LLM-generated natural language directives serve as an interpretability bridge between algorithmic control and human oversight, enhancing clarity and trust.

In a traditional Q-learning framework, each individual updates their strategy solely based on their own earnings, which easily leads to strategy concentration. In multi-agent environments, traditional Q-learning agents learn only from their immediate rewards and local states:

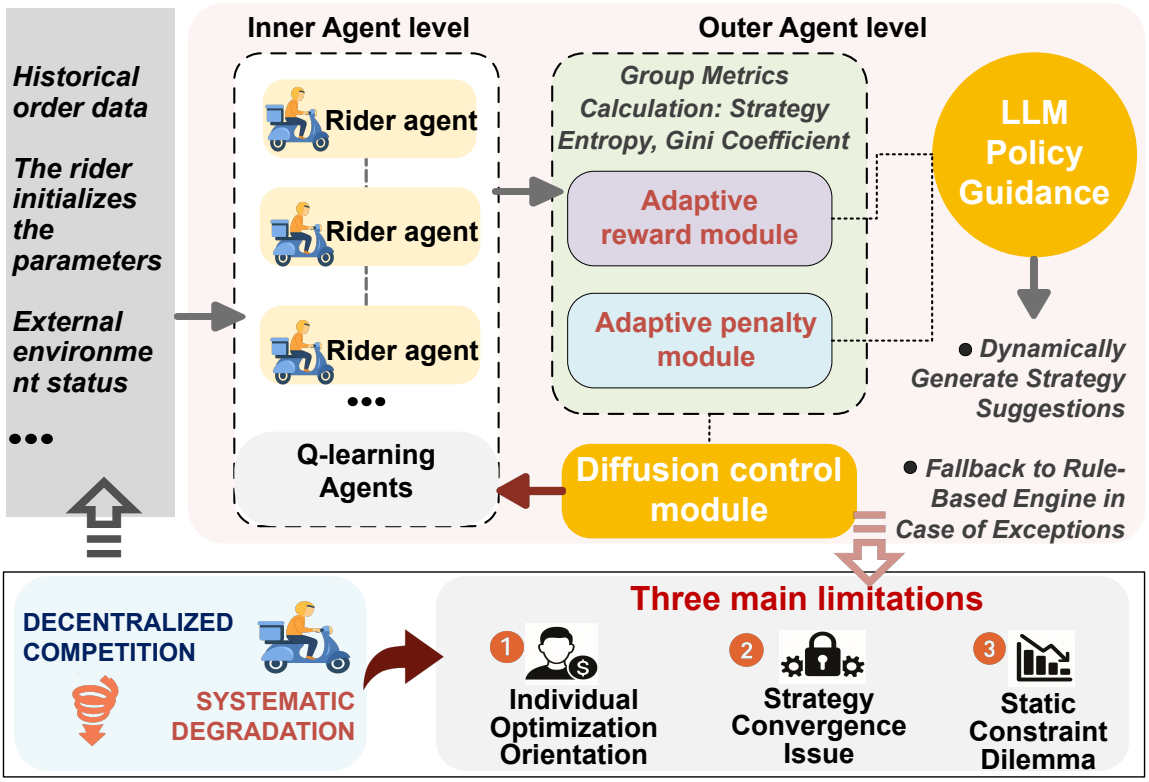

Figure 1: Architecture of ARDE: A Two-Level Adaptive Governance Framework for Crowdsourced Delivery Platforms.

$$Q_i(s,a) \leftarrow Q_i(s,a) + \alpha \left[r_i + \gamma \max_{a'} Q_i(s',a') - Q_i(s,a) \right]; \tag{8}$$

Under this mechanism, if the environment or reward structure exhibits certain preferences, all agents tend to converge toward a single dominant strategy. This results in a collapse of the strategy space, reducing system-level innovation and robustness.

State Representation and Observation Space

ARDE employs a fine-grained state representation mechanism to capture the complex dynamics of the delivery environment. Each agent's state is constructed from a three-dimensional observation vector:

$$s_i^{(t)} = [r_i^{(t)}, \pi_i^{(t)}, i \bmod 24]; \tag{9}$$

where $r_i^{(t)}$ represents the agent's total reward, $\pi_i^{(t)}$ denotes the current strategy identifier, and $i \bmod 24$ captures the temporal context based on agent index (simplified hour simulation). The reward component is discretized into bins of 50 CNY intervals to reduce state space dimensionality:

$$r_{\text{bin}}^{(t)} = \lfloor r_i^{(t)} / 50 \rfloor; \tag{10}$$

The state space is further compressed through a hash-based mapping mechanism that converts the continuous observation space into a finite set of discrete states, enabling efficient Q-table storage and lookup operations.

Adaptive Reward Shaping

To promote information sharing and the evolution of strategy diversity among individuals, ARDE introduces an adaptive reward shaping mechanism. This mechanism diffuses the global diversity objective into the local learning process of each agent through multi-dimensional reward terms, establishing a transmission pathway of "global objectives → local incentives → individual behaviors → collective evolution." Traditional reinforcement learning considers only individual immediate rewards, whereas ARDE transforms system-level goals such as diversity, fairness, and efficiency into perceivable reward signals for individuals via a multi-dimensional reward function, thereby coupling global objectives with local learning.

The reward function in ARDE consists of five dimensions, each addressing different aspects of system optimization. The base reward term preserves the original earnings incentive for couriers, ensuring that individual motivation remains intact: $r_{i,t}^{\text{base}} = r_{i,t}$. The diversity reward term leverages the entropy of the strategy distribution to encourage couriers to select distinct strategies and avoid strategy concentration: $r_{i,t}^{\text{diversity}} = \lambda_{\text{div}}^{(t)} H(\mathbf{p}_t)$. The rarity reward term encourages couriers to select strategies that are currently underutilized, promoting exploration of the strategy space: $r_{i,t}^{\text{rarity}} = \lambda_{\text{rarity}}^{(t)} \cdot (1 - p_{k_i,t})$. The efficiency reward term is based on the number of completed orders, incentivizing couriers to improve their working efficiency: $r_{i,t}^{\text{efficiency}} = \lambda_{\text{efficiency}}^{(t)} \cdot$

$\text{completed\_orders}_i^{(t)}$. Finally, the polarization penalty term is activated when a particular strategy's proportion becomes excessively high, preventing over-concentration of strategies: $r_{i,t}^{\text{penalty}} = -\lambda_{\text{pen}}^{(t)} \cdot I\big[\max_k p_{k,t} > \theta^{(t)}\big]$.

The total reward function of ARDE integrates all five dimensions into a balanced incentive system:

$$r_{i,t}^{\text{ARDE}} = r_{i,t}^{\text{base}} + r_{i,t}^{\text{diversity}} + r_{i,t}^{\text{rarity}} + r_{i,t}^{\text{efficiency}} + r_{i,t}^{\text{penalty}}; \tag{11}$$

Adaptive Parameter Adjustment Mechanism

ARDE's Outer Agent is equipped with an adaptive parameter adjustment capability, enabling dynamic optimization of reward weights based on the real-time state of the system, thereby achieving intelligent platform regulation. The diversity weight is dynamically adjusted according to the current strategy entropy and Gini coefficient:

$$\lambda_{\text{div}}^{(t)} = \min\left(0.5,\ \lambda_{\text{base}}(1 + H_t \cdot 0.2)(1 + G_t \cdot 0.4)\right); \tag{12}$$

When strategy entropy is low, the diversity weight is increased to encourage dispersed strategies; when the Gini coefficient is high, the diversity incentive is strengthened to improve income distribution. The upper bound of 0.5 is set to avoid excessive intervention in individual decision-making.

The polarization penalty is dynamically adjusted based on strategy concentration and the degree of income inequality:

$$\lambda_{\text{pen}}^{(t)} = \lambda_{\text{base}} + 4 \max_k p_{k,t} + 2G_t; \tag{13}$$

When a particular strategy's proportion is excessively high or income inequality is severe, the penalty intensity is increased to prevent the system from falling into a polarized state. A base penalty of $\lambda_{\text{base}} = 3.0$ ensures fundamental regulatory effectiveness.

The exploration rate is dynamically adapted according to strategy diversity and evolutionary time:

$$\epsilon^{(t)} = \epsilon_{\text{base}}(1 - H_t/2.0)\max\left(0.1, 1 - t/30\right); \tag{14}$$

When strategy diversity is low, the exploration rate is increased to encourage strategy innovation; as time progresses, the exploration rate gradually decreases to promote strategy convergence. The base exploration rate $\epsilon_{\text{base}} = 0.08$ maintains a proper exploration-exploitation balance. This adaptive adjustment mechanism ensures that the system can sustain optimal regulatory effectiveness under varying states.

Q-Learning Update Mechanism

ARDE adopts the standard Q-Learning update mechanism but couples global objectives with local learning through dynamic reward shaping by the Outer Agent, thereby addressing the conflict between individual optimality and collective optimality in traditional multi-agent reinforcement learning. Each rider's Q-table is updated following the standard Q-Learning rule, but using ARDE's multi-dimensional reward:

where the learning rate $\alpha = 0.08$ ensures stable Q-value updates and avoids excessive oscillations; the discount factor $\gamma = 0.97$ emphasizes long-term returns, suitable for scenarios like food delivery that require long-term planning; the exploration rate $\epsilon = 0.12$ balances exploration and exploitation, promoting strategy diversity; the state space $|\mathcal{S}| = 300$ captures complex environmental states; and the action space $|\mathcal{A}| = 4$ corresponds to four strategy choices, maintaining decision simplicity.

$$\begin{aligned} Q_i(s,a) \leftarrow Q_i(s,a) + \alpha\big[r_{i,t}^{\text{ARDE}} \\ +\gamma \max_{a'} Q_i(s',a') - Q_i(s,a)\big] \end{aligned} \tag{15}$$

Explicit Representation of Platform Rules via Large Language Models

The Large Language Model (LLM) analyzes the current system state and generates natural language guidance rules that inform the Outer Agent's parameter adjustment strategy. Specifically, the LLM, based on the real-time values of strategy entropy and the Gini coefficient, produces interpretable guidance such as "recommend increasing strategy diversity to avoid strategy concentration" or "recommend optimizing income distribution to reduce income disparity." The Outer Agent then translates the LLM's guidance into concrete parameter adjustments:

$$(\lambda_{\text{div}}^{(t+1)}, \lambda_{\text{pen}}^{(t+1)}, \theta^{(t+1)}) = f_{\text{Outer}}(\text{LLM}(\mathbf{S}_t)); \tag{16}$$

where f_{Outer} denotes the parameter adjustment function that converts the LLM's natural language recommendations into numerical control parameters.

ARDE establishes a complete "state-guidance-parameter-reward-action-state" feedback loop, where the system state $\mathbf{S}_t$ is first processed by the LLM module to generate high-level guidance Guidance_t. This guidance is then mapped by the outer-layer function f_{Outer} into updated platform parameters $\text{Parameters}_{\text{t+1}}$, which are used to shape individual rewards $\{r_{i,t+1}^{\text{ARDE}}\}$. These adapted rewards guide agent decisions, resulting in a new system state $\mathbf{S}_{t+1}$ for the next iteration. ensuring that platform governance rules can adaptively evolve with the system. This mechanism achieves a coordinated optimization of efficiency, fairness, and diversity, providing a novel technological pathway for the intelligent governance of crowdsourced delivery platforms.

The complete pseudocode of the ARDE algorithm is shown below:

Experiments

This study utilizes a real-world dataset from a food delivery platform, containing data from 350 couriers over a 30-day period. A five-stage experimental framework was designed, incorporating seven comparative studies to comprehensively evaluate the effectiveness of the ARDE algorithm. Specifically, the experimental setup includes:

- E1:Strategy free-evolution validation (no regulation), simulating strategy degradation driven by self-interest;
- E2:Real-strategy replication, assessing whether the current platform has already fallen into a non-cooperative dilemma;
- E3:Regulatory mechanism comparison, contrasting the effectiveness of rule-based control, reward shaping, and ARDE;

Algorithm 1: ARDE: Dual-Layer Adaptive Regulation

Input: Dataset D, Agents A, Strategy set S, Horizon T
Parameters: Learning rate α, Discount factor γ, Exploration rate ϵ
Output: Optimized strategy distribution Π^*, Metrics M

Initialize Q_i for all agents $a_i \in A$
Initialize Outer Agent parameters λ, β
for each time step $t \in T$ **do**
  $D_t \leftarrow$ Orders of day t
  stats $\leftarrow$ SimulateDay(D_t, A)
  $H \leftarrow$ Entropy(strategy_counts(stats))
  $G \leftarrow$ Gini(stats.rewards)
  $\lambda \leftarrow f_\lambda(H, G)$; $\beta \leftarrow f_\beta(H, G)$
  guidance $\leftarrow$ LLM(H, G)
  for each agent $a_i \in A$ **do**
    $s \leftarrow$ EncodeState(stats[i])
    $a \leftarrow \epsilon$-greedy(Q_i, s)
    $r \leftarrow$ RewardShape(stats[i], λ, β)
    $Q_i[s, a] \leftarrow Q_i[s, a] + \alpha[r + \gamma \max_{a'} Q_i[s', a'] - Q_i[s, a]]$
    $\pi_i \leftarrow$ MapActionToStrategy(a)
  end for
  Record(M, H, G, guidance)
end for
return Π^*, M

- E4:Multi-objective evaluation analysis, providing a cross-sectional comparison across efficiency, fairness, and strategy diversity.
- E5:A state-of-the-art centralized training–decentralized execution (CTDE) method.

To evaluate system performance, three key metrics are employed: strategy entropy (capturing strategy diversity), Gini coefficient (reflecting income fairness), and order completion rate (indicating overall system efficiency). The comparison of strategy entropy and Gini coefficient across all experimental methods (with error bands representing ±1 standard deviation) illustrates the balanced performance of ARDE in terms of diversity and fairness, as shown in Figure 2.

Strategy Diversity Performance

As shown in Table 1, different algorithms exhibit significant variations in strategy diversity. The ARDE algorithm maintains a consistently high strategy entropy (0.997 ± 0.184) throughout the entire experimental period, achieving an improvement of $+11.0\%$ over the initial state and a $+29.6\%$ increase in diversity compared to baseline methods. In contrast, the Q-learning algorithm demonstrates severe strategy polarization, with an average entropy of only 0.365 ± 0.045, achieving a marginal increase of $+2.2\%$, while the overall diversity level remains significantly low. Notably, the Reward Shaping method reaches the highest strategy entropy (1.029 ± 0.170) and the largest change rate ($+36.4\%$), yet its relatively large standard deviation indicates insufficient strategy stability. Both Rule-based and Real Strategy approaches exhibit pronounced strategy convergence, decreasing by -29.5% and -28.3%, respectively, validating the theoretical hypothesis that individual rationality can lead to collective suboptimality.

Fairness Improvement Effect

The fairness analysis reveals the groundbreaking performance of the ARDE algorithm. ARDE is the only method achieving a negative change in the Gini coefficient, with a variation rate of -1.3%, indicating that the algorithm effectively improves income distribution fairness. In contrast, all other methods lead to varying degrees of increased inequality: Q-learning ($+27.9\%$), Real Strategy ($+53.6\%$), Rule-based ($+36.3\%$), Static Rule ($+20.9\%$), Reward Shaping ($+1.2\%$), and QMix ($+0.2\%$). The average Gini coefficient under ARDE is 0.339 ± 0.028, slightly higher than that of Reward Shaping (0.329 ± 0.027). However, considering its simultaneous and significant enhancement of strategy diversity, this trade-off highlights a key advantage of the ARDE design.

System Efficiency and Adaptability

The system efficiency analysis indicates that the ARDE algorithm maintains stable system efficiency while achieving high strategy diversity and improved fairness. ARDE achieves an order completion rate of 0.411 and an average hourly income of 6.86 CNY, comparable to the Rule-based approach (completion rate 0.418, income 6.92 CNY/hour), demonstrating its competitiveness in terms of efficiency. Notably, the Q-learning method performs the worst in terms of efficiency, with a completion rate of only 0.189 and an average income of 3.86 CNY/hour, validating the theoretical assumption that purely individual rational optimization can lead to systemic efficiency collapse. ARDE also exhibits the highest strategy switching rate (0.694), indicating strong adaptability and learning capacity, whereas Q-learning shows the lowest switching rate (0.147), reflecting its rigid strategies and lack of adaptability.

E5: ARDE-PPO vs. MAPPO Fair Comparison

To ensure a fair and consistent evaluation between ARDE and mainstream MARL methods, we designed two additional experiments: E5-1 and E5-2. E5-1 preserves ARDE's dual-layer governance structure but replaces the inner-layer learner with PPO, allowing a direct comparison with E5-2 (MAPPO) under identical learning conditions. E5-1 (ARDE-PPO) achieves a negative Gini change (1.1%), whereas E5-2 (MAPPO) exhibits a positive change (+2.1%), a difference of 3.2 percentage points. This finding demonstrates that ARDE's fairness improvement stems from its explicit hierarchical governance mechanism rather than the specific choice of the inner learning algorithm. Notably, ARDE (E3-3 and E5-1) remains the only framework achieving negative Gini growth, validating the robustness of its governance design.

Discussion and Theoretical Validation

The experimental results strongly support the theoretical proposition that individual rationality can lead to collective suboptimality. While Q-learning optimizes individual

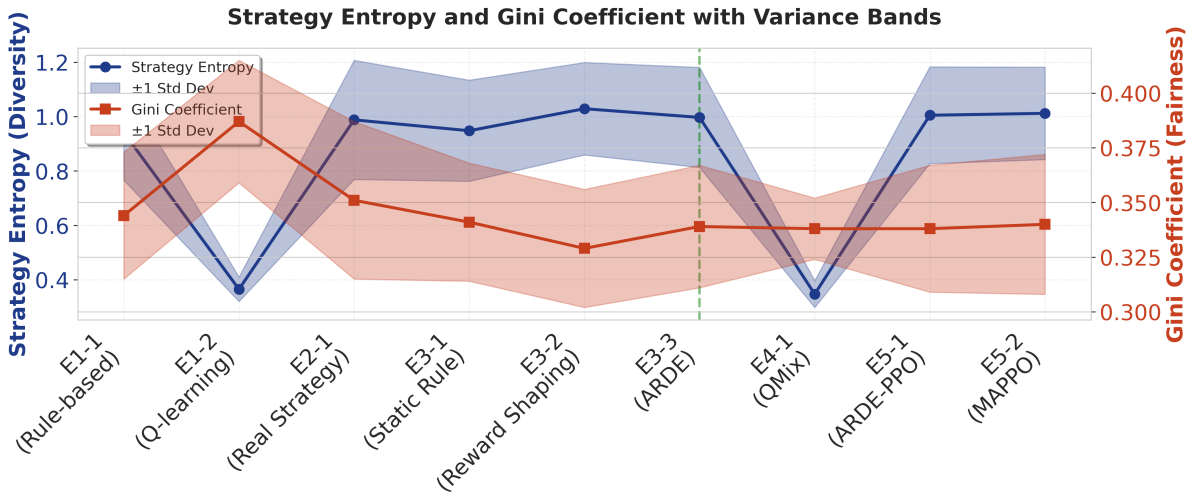

Figure 2: Strategy entropy (diversity) and Gini coefficient (fairness) across experiments, with error bands showing ±1 standard deviation. ARDE achieves high diversity (0.997) while maintaining relatively low inequality (0.339).

Method	Strategy Entropy	Gini Coefficient	Completion Rate	Average Revenue (CNY/h)	Entropy Change (%)	Gini Change (%)	System Efficiency
E1-1 (Rule-based)	0.941±0.175	0.344±0.029	0.418	6.92	-29.5	+36.3	0.494
E1-2 (Q-learning)	0.365±0.045	0.387±0.028	0.189	3.86	+2.2	+27.9	0.147
E2-1 (Real Strategy)	0.988±0.219	0.351±0.036	0.413	6.85	-28.3	+53.6	0.621
E3-1 (Static Rule)	0.948±0.186	0.341±0.027	0.417	6.89	-5.0	+20.9	0.692
E3-2 (Reward Shaping)	1.029±0.170	0.329±0.027	0.411	6.91	+36.4	+1.2	0.677
E3-3 (ARDE)	**0.997±0.184**	**0.339±0.028**	**0.411**	**6.86**	**+11.0**	**-1.3**	**0.694**
E4-1 (QMix)	0.347±0.049	0.338±0.014	4.723	4.20	+31.8	+0.2	0.000
E5-1 (ARDE-PPO)	1.005±0.178	0.338±0.029	0.409	6.88	+12.5	**-1.1**	0.691
E5-2 (MAPPO)	1.012±0.170	0.340±0.032	0.407	6.95	+15.2	+2.1	0.697

Table 1: Comprehensive performance comparison of all experimental methods on the LaDe-D dataset.

payoffs, it induces severe strategy homogenization (entropy 0.365) and worsening income inequality (+27.9%), with a dramatic drop in efficiency (completion rate 0.189). The Real Strategy setting further highlights the limitations of existing platform governance (Gini change +53.6%). In contrast, ARDE's dual-layer reinforcement learning framework, combined with LLM-driven constraint generation, successfully resolves this dilemma. The outer agent performs global governance and adaptive reward modulation, while inner agents handle tactical decision-making—jointly balancing efficiency, fairness, and diversity. The Diffusion Q-learning mechanism promotes cooperative exploration, avoiding premature convergence, and LLM-based guidance introduces interpretability and adaptive policy shaping.

Comparative experiments with state-of-the-art MARL methods (QMix and MAPPO) further demonstrate ARDE's capability to achieve a more balanced optimization across fairness, efficiency, and strategy diversity.Even when sharing the same inner learner (PPO), ARDE-PPO (E5-1) achieves a fairer outcome (Gini 1.1%) than MAPPO (+2.1%), reinforcing that fairness gains originate from the hierarchical governance design rather than algorithmic coincidence.

Conclusion

This study introduces ARDE (Adaptive Regulation via Dual-layer Evolution), an evolutionary governance framework that reconciles individual rationality with collective welfare in crowdsourced delivery platforms. By combining a dual-layer reinforcement learning architecture with language model–driven policy generation, ARDE enables adaptive, interpretable, and fairness-oriented governance in dynamic multi-agent ecosystems. Experiments on real-world datasets show that ARDE consistently enhances strategy diversity, mitigates income inequality, and sustains high operational efficiency, outperforming both rule-based mechanisms and state-of-the-art MARL baselines.Beyond empirical validation, ARDE establishes a theoretical foundation for algorithmic governance, demonstrating that hierarchical regulation can systematically resolve the enduring tension between individual rationality and collective welfare.Looking ahead, ARDE offers a scalable paradigm for intelligent and equitable governance, with potential extensions to cross-platform coordination, multi-level policy adaptation, and ethically aligned incentive design.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62472306, No. 62441221, No. 62206116), Tianjin University's 2024 Special Project on Disciplinary Development (No. XKJS-2024-5-9), the Tianjin University Talent Innovation Reward Program for Literature Science Graduate Student (C1-2022-010), and the Henan Province Key Research and Development Program (No. 251111210500).

References

Alnaggar, A.; Gzara, F.; and Bookbinder, J. H. 2021. Crowdsourced delivery: A review of platforms and academic literature. *Omega*, 98: 102139.

Sun, P. 2019. Your order, their labor: An exploration of algorithms and laboring on food delivery platforms in China. *Chinese Journal of Communication*, 12(3): 308–323.

Shiri, A.; Yarahmadi, A.; and Keivanpour, S. 2025. Real-time matching and dispatching for urban freight transportation: A hierarchical reinforcement learning through actor-critic and H3 spatial partitioning. *IEEE Transactions on Intelligent Transportation Systems*.

Xue, X.; Li, G.; Zhou, D.; et al. 2022. Research roadmap of service ecosystems: A crowd intelligence perspective. International Journal of Crowd Science, 6(4): 195–222.

Hady, M. A.; Hu, S.; Pratama, M.; et al. 2025. Multi-Agent Reinforcement Learning for Resources Allocation Optimization: A Survey. *arXiv preprint* arXiv:2504.21048.

Yu, C.; Velu, A.; Vinitsky, E.; et al. 2022. The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624.

Pandit, V. N.; Datar, M.; Hanawal, M. K.; and Moharir, S. 2019. Pricing in ride sharing platforms: Static vs dynamic strategies. *Proceedings of the 11th International Conference on Communication Systems & Networks (COMSNETS)*: 208–215.

Rudnik, J.; and Brewer, R. 2023. Care and coordination in algorithmic systems: An economies of worth approach. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*: 627–638.

Xue, X.; Yu, X.; Zhou, D.; et al. 2024. Computational experiments for complex social systems: Integrated design of experiment system. *IEEE/CAA Journal of Automatica Sinica*, 11(5): 1175–1189.

Xiao, X.; Xiang-Ning, Y.; De-Yu, Z.; et al. 2023. Computational experiments: Past, present and perspective. Acta Automatica Sinica, 49(2): 246–271.

Gavrilets, S.; and Shrestha, M. D. 2021. Evolving institutions for collective action by selective imitation and self-interested design. *Evolution and Human Behavior*, 42(1): 1–11.

Zhao, X.; Ai, P.; Lai, F.; et al. 2022. Task management in decentralized autonomous organization. *Journal of Operations Management*, 68(6–7): 649–674.

Lunansky, G.; Bonanno, G. A.; Blanken, T. F.; et al. 2024. Bouncing back from life's perturbations: Formalizing psychological resilience from a complex systems perspective. *Psychological Review*.

Kang, H.; Li, Z.; Shen, Y.; et al. 2024. From eligibility to suitability: Regulation and restriction of reputation-based access system on free-riding behavior in spatial public goods game. *Chaos, Solitons & Fractals*, 180: 114547.

Saleh, Z.; Al Hanbali, A.; and Baubaid, A. 2024. Enhancing courier scheduling in crowdsourced last-mile delivery through dynamic shift extensions: A deep reinforcement learning approach. *arXiv preprint* arXiv:2402.09961.

Xu, Z.; and Zheng, S. 2024. An evolutionary game-theoretic analysis of the "multi-agent co-governance" system of unfair competition on internet platforms. *PLOS ONE*, 19(6): e0304445.

Zhong, R.; Liang, Q.; Xu, D.; et al. 2024. Day-to-day road pricing and network performance analysis via output stability. *IEEE Transactions on Intelligent Transportation Systems*, 25(11): 16336–16353.

Wang, Y.; Cai, Z.; Yin, G.; et al. 2016. An incentive mechanism with privacy protection in mobile crowdsourcing systems. *Computer Networks*, 102: 157–171.

Lu, M.; Chen, S.; Xue, X.; et al. 2021. Computational experiments for complex social systems—Part II: The evaluation of computational models. IEEE Transactions on Computational Social Systems, 9(4): 1224–1236.

Liu, S.; Zhou, Y.; Song, J.; Zheng, T.; Chen, K.; Zhu, T.; Feng, Z.; and Song, M. 2023. Contrastive identity-aware learning for multi-agent value decomposition. *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)*: 11595–11603.

Xue, X.; Yu, X.; Zhou, D.; et al. 2023. Computational experiments for complex social systems—Part III: The docking of domain models. IEEE Transactions on Computational Social Systems, 11(2): 1766–1780.

Xue, X.; Guo, Y.; Chen, S.; et al. 2019. Analysis and controlling of manufacturing service ecosystem: A research framework based on the parallel system theory. *IEEE Transactions on Services Computing*, 14(6): 1598–1611.

Yu, X.; Xue, X.; Zhou, D.; et al. 2024. Beyond traditional metrics: The power of value entropy in multidimensional evaluation of the service ecosystem. *Proceedings of the 2024 IEEE International Conference on Web Services (ICWS)*: 611–621.

Yu, X.; Xue, X.; Zhou, D.; et al. 2025. Unlocking complexity: Harnessing value entropy for advanced multidimensional utility evaluation in service ecosystems. IEEE Transactions on Services Computing.