Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/BA08.20250076
Published: December 20, 2025

A Multi-Agent Reinforcement Learning Approach for Recovering Evolutionarily Stable Strategies in Evolutionary Games Using Proximal Policy Optimization

Mahamudul Hassan Siddique and Fahimul Haque
Department of Industrial and Production Engineering
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh
hassansiddque632@gmail.com, fahimulhaq2001@gmail.com.

Abstract

Evolutionary Game Theory (EGT) solutions become analytically intractable as game complexity increases, thereby limiting its applicability to complex, multi-agent systems. In this paper, we present a new computational framework that redefines evolutionary games as stateless multi-agent reinforcement learning (MARL) problems using independent Proximal Policy Optimization (PPO) agents. The proposed framework can effectively approximate evolutionarily stable strategies (ESS) without requiring closed-form replicator dynamics and extends naturally to multi-player, multi-strategy games. Based on a three-player e-commerce game case study, it is shown that the method aligns with two analytically derived equilibria, E_4 (1,0,1) and E_3 (0,0,1), with learned strategy probabilities greater than 0.999, thereby demonstrating near-perfect agreement. The thorough sensitivity analysis identifies important parametric thresholds that trigger changes in strategic behavior, explaining why the approach is effective in revealing equilibrium boundaries. Its key contributions are as follows: (1) scaling of MARL-based solutions for ESS approximation in standard evolutionary games; (2) empirical validation of theoretical equilibria through decentralized learning; and (3) a reproducible computational framework that complements traditional analytical EGT. The limitations include the use of a one-shot, stateless game model and validation restricted to two equilibrium states. This paper highlights the potential of MARL as a practical and generalizable tool for studying complex strategic interactions in contexts where traditional solution methods are impractical.

Keywords

Evolutionary Game Theory, Proximal Policy Optimization, Reinforcement Learning, Advantage Actor Critic (A3C), E-Commerce

1. Introduction

Evolutionary Game Theory (EGT) has been widely used to analyze strategic interactions in e-commerce, particularly in addressing counterfeit goods (Le, 2023). Replicator dynamics is used for the identification of evolutionarily stable strategies (ESS) in such games. However, analytical solutions become intractable as the number of players or strategies increases, limiting the applicability of EGT in complex scenarios (Blanc & Hansen, 2021).

To overcome these challenges, a Multi-Agent Reinforcement Learning (MARL) approach based on Proximal Policy Optimization (PPO) is adopted. Each participant is represented as an independent agent that interacts with others in a shared environment. Through repeated interactions, strategies are adapted and convergence toward equilibrium behaviors analogous to ESS is achieved without explicit analytical derivation. This approach is inherently scalable and adaptable to games with a large number of players and strategies, while also providing empirical validation of equilibrium outcomes. It allows decentralized learning by enabling each agent to optimize its policy based on observed

payoffs. By integrating the theoretical foundation of EGT with the computational capability of MARL, complex strategic interactions in e-commerce can be effectively analyzed, offering a practical alternative where classical analytical methods are difficult and cumbersome to implement.

1.1 Research Objectives

1) A multi-agent PPO framework is developed to approximate theoretical equilibrium strategies in an evolutionary game.
2) Convergence of learned strategies is validated by comparing the converged probabilities with specific parameter values given in the paper for two theoretical equilibrium points and stability test of the framework is also done.
3) Sensitivity analysis of price, cost parameters, alpha/beta/gamma, supervision cost, platform revenue, customer reward, and word-of-mouth loss was conducted to locate strategy-shift boundaries.

2. Literature Review

Recent progressions in multi-agent reinforcement learning (MARL) provide a principled mechanism to obtain strategies yielding a high payoff while maintaining stability under population dynamics, making it well aligned with the aims of evolutionary game theory (EGT). Policy-gradient algorithms with value baselines, namely Advantage Actor-Critic (A2C/A3C) and Proximal Policy Optimization (PPO), are the ingredients needed: they support on-policy updates with variance reduction, monotonic improvement heuristics, and architectures allowing centralized training while supporting decentralized execution. The structural issues discussed in modern surveys include credit assignment, non-stationarity, as well as partial observability, and they identify actor-critic and PPO variants as the most reliable for coordinated behaviour in multi-agent games (Canese et al. 2021; Wong et al. 2024; Zhang et al. 2021). These algorithms, through the EGT lens, realize selection-like forces through advantage estimates and policy updates, allowing groups of policies to approach equilibria rather than over-fitting to fixed opponents (Omidshafiei et al. 2019).

In PPO-like schemes, stability can be ensured by limiting the change in each policy update. Experimental results show that small, finely tuned policy steps coupled with strong critics can outperform more complicated off-policy stacks in coordination problems (Wong et al. 2024). Recent studies also extend the application of PPO to MARL by controlling the likelihood-ratio at the activation level, improving gradient reliability when agents co-adapt and thereby reducing training pathologies associated with non-stationary opponents (Jia et al. 2024). Complementary mechanisms focus on the critic. Centralized or information-sharing critics give each agent a learned approximation of the joint state, countering spurious attributions and enabling coherent collective behaviour; recurrent designs, in particular, capture long-horizon dependencies and hidden states caused by partial observability (Kargar et al. 2024; Gabler et al. 2024). The combined effect of these architectures reflects the EGT intuition that knowledge of the population state, embedded in the critic, improves the quality of selection pressure encoded by policy gradients.

Advantage Actor-Critic is also an important workhorse system in environments where decentralized execution is dominant. Natural-gradient and compatible-function-approximation variants preserve desirable convergence guarantees when extending from single-agent to multi-agent settings and explain the impact of advantage estimation, baselines, and compatible critics on the stability of joint learning (Trivedi and Hemachandra 2023). Actor-critic families have proved to be effective in safety-critical and interaction-dense applied contexts. An example is cooperative lane-change decision-making, which uses multi-agent RL blending global reward shaping with look-ahead search to enhance coordination without requiring centralized control at execution time (Zhou et al. 2022). Hierarchical hybrids combining actor-critic with value-based updates offer another way toward efficient exploration and robust coordination under partial observability and sparse feedback, which aligns with the structured interactions and multi-level selection emphasized by EGT (Gabler et al. 2024).

Another key EGT property is robustness against invasion: strategies should remain effective when the composition of opponents shifts. Empirical game-theoretic analysis (EGTA) tools such as AlphaRank substitute brittle head-to-head win-rate statistics with evolutionary ranking based on Markov chains over strategy profiles, providing a population-based measure of strategic strength that is well suited to evaluating MARL policy sets (Omidshafiei et al. 2019). In high-stakes and high-complexity settings, training large groups of agents has revealed that policy-gradient systems can discover sophisticated coordination strategies and counter-strategies that remain effective under extensive self-play, a property that mirrors evolutionary stability at scale (Vinyals et al. 2019). These results confirm the methodological consistency between PPO/A2C and EGT: as long as policy gradients are regularized and guided by joint-state critics, they tend to produce populations of strategies with strong resistance to exploitation and collapse.

The population-level consequences of stochastic learning are also vital. Evidence suggests that intrinsic fluctuations in reinforcement learning can promote cooperative behaviour in social dilemmas by preventing premature convergence to defective equilibria, thereby enlarging the basin of attraction for cooperative strategy profiles (Barfuss and Meylahn 2023). This phenomenon aligns with PPO's clipped updates and actor-critic variance-reduction: modest, noise-tempered policy adjustments allow sufficient exploration near payoff ridges for cooperative conventions to emerge (Barfuss and Meylahn 2023; Wong et al. 2024). This effect is further strengthened by credit-assignment refinements in multi-agent environments, which allocate advantage to contributory actions rather than to free-riding actions, a prerequisite for stable cooperation in repeated interactions (Huang et al. 2022; Zhang et al. 2021).

Overall, the existing literature converges on a consistent recipe for EGT-oriented multi-agent optimization with actor-critic and PPO: adopt centralized or information-sharing critics-often recurrent-to address non-stationarity and capture joint dynamics; use clipped or likelihood-controlled policy updates to stabilize improvement under co-adaptation; incorporate reward shaping or multi-level credit assignment to align individual gradients with population-level objectives; and evaluate learned policy sets with evolutionary metrics such as AlphaRank to verify robustness against invasion and cycling (Omidshafiei et al. 2019; Jia et al. 2024; Kargar et al. 2024; Wong et al. 2024). This toolbox is directly compatible with feasibility-first constrained optimization: differentiable penalties and constraint-aware critics provide the selection pressures, PPO/A2C ensure stable adaptation, and EGTA verifies that "optimal" means strategically resilient in the evolutionary sense (Zhang et al. 2021; Canese et al. 2021).

3. Methods

3.1 Tools and Libraries Used

Python was used to conduct all experiments; NumPy was used to assist with numerical computation, whereas PyTorch was used to build and train actor-critic models based on PPO agents. The libraries provided vectorized operations, the automatic differentiation, as well as the optimization to enable scalable simulation and analysis of the e-commerce game.

3.2 E-Commerce Evolutionary Game Model

The e-commerce game in this research is taken from Guo et al. (2021) in which interactions between sellers, customers and e-commerce platform were analyzed using evolutionary game theory (EGT) and replicator dynamics. The same model is used here to allow comparison between analytical and those solutions obtained by Multi Agent Reinforcement Learning framework (MARL).
The game has three strategic players: The Seller who decides to supply either a genuine (1) or fake (0) product, the Customer decides to accept (1) or reject back (0) the product, and the Platform decides whether to adopt the slack (0) or strict (1) supervision arrangements.

The payoffs for each player are defined by a set of economic and regulatory factors specified in the original study:

1) The Seller's payoff is influenced by the product's selling price P, manufacturing cost C_m, customer return compensation B, fines imposed by the platform F, the fraction α of value perceived for fake products, the cost reduction factor β for fake product, and the word of mouth loss, W_m.
2) The Customer's payoff is determined by the utility from consuming a genuine product R_c, the cost of returning an unsatisfactory product C_c, monetary compensation I provided in the case of returns, and the perception factor γ that reduces the value of counterfeit goods.
3) The Platform's payoff is shaped by profit or transaction revenue R_p, the cost of strict supervision C_p, and fines collected from counterfeit sellers F.

3.3 Expected Payoff Equations for Each Player

3.3.1 Seller (Merchant)

x = probability the seller sells genuine commodities
y = probability the platform adopts strict supervision
z = probability the customer accepts the commodity

Expected revenue when selling genuine commodities (Em1):

$$E_{m_1} = yz(P - C_m) + (1 - y)z(P - C_m) + y(1 - z)(-B) + (1 - y)(1 - z)(-B)$$

Expected revenue when selling fake commodities (Em2):

$$E_{m_2} = yz(\alpha P - C_m - F) + (1 - y)z(\alpha P - \beta C_m - W_m) + y(1 - z)(-F - B) + (1 - y)(1 - z)(-B - F - I)$$

Overall expected revenue:

$$\hat{E}_m = x * E_{m_1} + (1 - x) * E_{m_2}$$

3.3.2 E-commerce Platform

C_p = the strict supervision cost of the e-commerce platform
R_p = the e-commerce platform's revenue after the seller successfully selling a commodity
F = economic penalty that the e-commerce platform imposes on the seller after finding a fake commodity

Expected revenue under strict supervision (Ep1):

$$E_{p_1} = xz(R_p - C_p) + x(1 - z)(-C_p) + (1 - x)z(F + R_p - 2C_p) + (1 - x)(1 - z)(F - 2C_p)$$

Expected revenue under slack supervision (Ep2):

$$E_{p_2} = xzR_p + x(1 - z)(-C_p) + (1 - x)zR_p + (1 - x)(1 - z)(F - C_p)$$

Overall expected revenue:

$$\hat{E}_p = y * E_{p_1} + (1 - y) * E_{p_2}$$

3.3.3 Customer (Consumer)

R_c = the utility of the customer brought by a genuine commodity
γ = the ratio of the utility of the customer brought by a fake commodity to the utility by a genuine one ($\gamma \in (0, 1)$)
I = the economic compensation paid by the seller to the customer
C_c = returning cost of the customer

Expected utility when accepting (Ec1):

$$E_{c_1} = xyR_c + x(1 - y)R_c + (1 - x)yR_c + (1 - x)(1 - y)\gamma R_c$$

Expected utility when returning (Ec2):

$$E_{c_2} = xy(-C_c) + x(1 - y)(-C_c) + (1 - x)y(-C_c) + (1 - x)(1 - y)(I - C_c)$$

Overall expected utility:

$$\hat{E}_c = z * E_{c_1} + (1 - z) * E_{c_2}$$

In the original paper, up to eight equilibrium points were found for different regimes of the parameters. However, only two cases, $E_4(1,0,1)$ and $E_3(0,0,1)$ had explicit parameter values. These two cases were chosen to enable direct comparison of MARL results with theoretical results and also to perform MARL training.
For the context of $E_4(1,0,1)$ where the seller creates a genuine product, the customer has confidence in the receipt of the product and the platform adopts slack supervision, the parameters' values are $P = 200, C_m, = 120, B = 10, W_m = 50, C_p = 20, R_p = 10, F = 100, R_c = 200, I = 10, C_c = 10, \alpha = 0.8, \beta = 0.3, \gamma = 0.5$. These values are given for the equilibrium point $E_4(1,0,1)$ described in the original paper.

For $E_3(0,0,1)$ in which the seller sells the fake product, the customer is the trader, and the platform uses slack supervision, the parameter setting values were $P = 200, C_m = 120, B = 10, W_m = 30, C_p = 100, R_p = 10, F = 100, R_c = 200, I = 10, C_c = 10, \alpha = 0.8, \beta = 0.3, \gamma = 0.5$. These values are given for the equilibrium point $E_3(0,0,1)$ described in the original paper.
The experiments that were run for this study did not include the other six equilibria provided by the original paper. This was because the parameter settings for other six equilibria were not mentioned. By focusing on $\boldsymbol{E_4(1,0,1)}$ and $\boldsymbol{E_3(0,0,1)}$ the ability of the multi-agent PPO framework to converge to the equilibrium profiles predicted by EGT under the given parameters was validated.

3.4 Multi-Agent PPO Framework

Approximation of the equilibrium involving the e-commerce game specified in previous section is made a multi-agent self-play game where each agent, i.e., the seller, the customer, and the platform, is modeled as an independent agent. As in evolutionary game theory, each player then aims to maximize its expected payoff and so do the agents in the MARL model, who encourages them to modify their strategy (set of policies) to maximize future expected payoffs, and the policy mix eventually converges to a stationary mixed-strategy equilibrium.

Contrary to traditional problems of reinforcement-learning where decisions occur sequentially, every episode entails one joint action of all three agents and then the episode is concluded. Payoffs to each agent are a one-step form; the discount factor is set at γ = 1.0 and, therefore, every episode's payoff is given same importance to help learn each agent. The game itself is regarded as stateless, with no other players in the evolutionary games knowing the strategy of other players.

Each agent is equipped with an actor–critic neural network. The actor outputs logits over the agent's two actions, defining a stochastic policy $\pi_\theta(a|s)$, while the critic estimates a scalar baseline $V_\varphi(s)$ for the constant state. Although no temporal evolution is present, the critic is used to reduce variance in policy-gradient estimation and to improve the approximation of the expected reward.

During each episode, actions are sampled from the agents' policies, and the joint action is passed to the environment to compute payoffs r^i for each agent i. The advantage is calculated as

$$\hat{A}^i = r^i - V_\varphi^i(s)$$

which guides the policy toward actions that outperform the baseline and effectively maximizes the agent's expected reward, analogous to the payoff-maximization process in evolutionary games.
Policy parameters are updated using the clipped surrogate objective of Proximal Policy Optimization (PPO) (Schulman, 2017):

$$L_{CLIP}(\theta) = E\big(\min(r(\theta)\hat{A}, clip(r(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}\big),$$

where

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{old}(a|s)}$$

Clipping is applied to reduce large policy changes and increase the stability of training. By minimizing the mean squared error between the observed reward and the estimated baseline the critic is optimized. Through repeated self-play, strategies are iteratively adjusted in response to the actions of others. In this repeated one-shot static game, the stationary joint policy achieved after training corresponds to a mixed-strategy Nash equilibrium and, if resistant to unilateral deviations, an evolutionarily stable strategy (ESS). Using this framework, we are able to recover equilibrium strategy of each agent constructively through training on episodes therefore maximizing the payoff of agents of MARL similar to players on an evolutionary game achieve equilibrium that maximizes expected payoffs.

3.5 Model Architecture and Training Parameters

Multi-agent PPO was implemented in PyTorch. All three agents (seller, customer, and platform) were assigned a separate actor–critic neural network. The network architecture, shown in Figure 1, consists of fully connected neural network with ReLU activation. A constant state representation of 1.0 is fed into the input layer because the game is

stateless. The actor head outputs the logits for the two available actions (seller: fake/genuine; customer: return/accept; platform: slack/strict), which are normalized using a softmax layer to produce the stochastic policy, $\hat{p}\theta(a|s)$. The critic head produces a scalar value estimate $V\theta(s)$ of the constant state, serving as a baseline for reducing variance in policy-gradient updates.

The PPO hyperparameters were as follows: discount factor $\gamma = 1.0$ (appropriate for a one-shot game), surrogate objective clipping parameter $\varepsilon = 0.2$, and Adam optimizer learning rate 3×10^{-4}. Training consisted of 3000 episodes, with 100 environment samples collected per episode to update the policy. All agents received a total of 3000 training episodes.

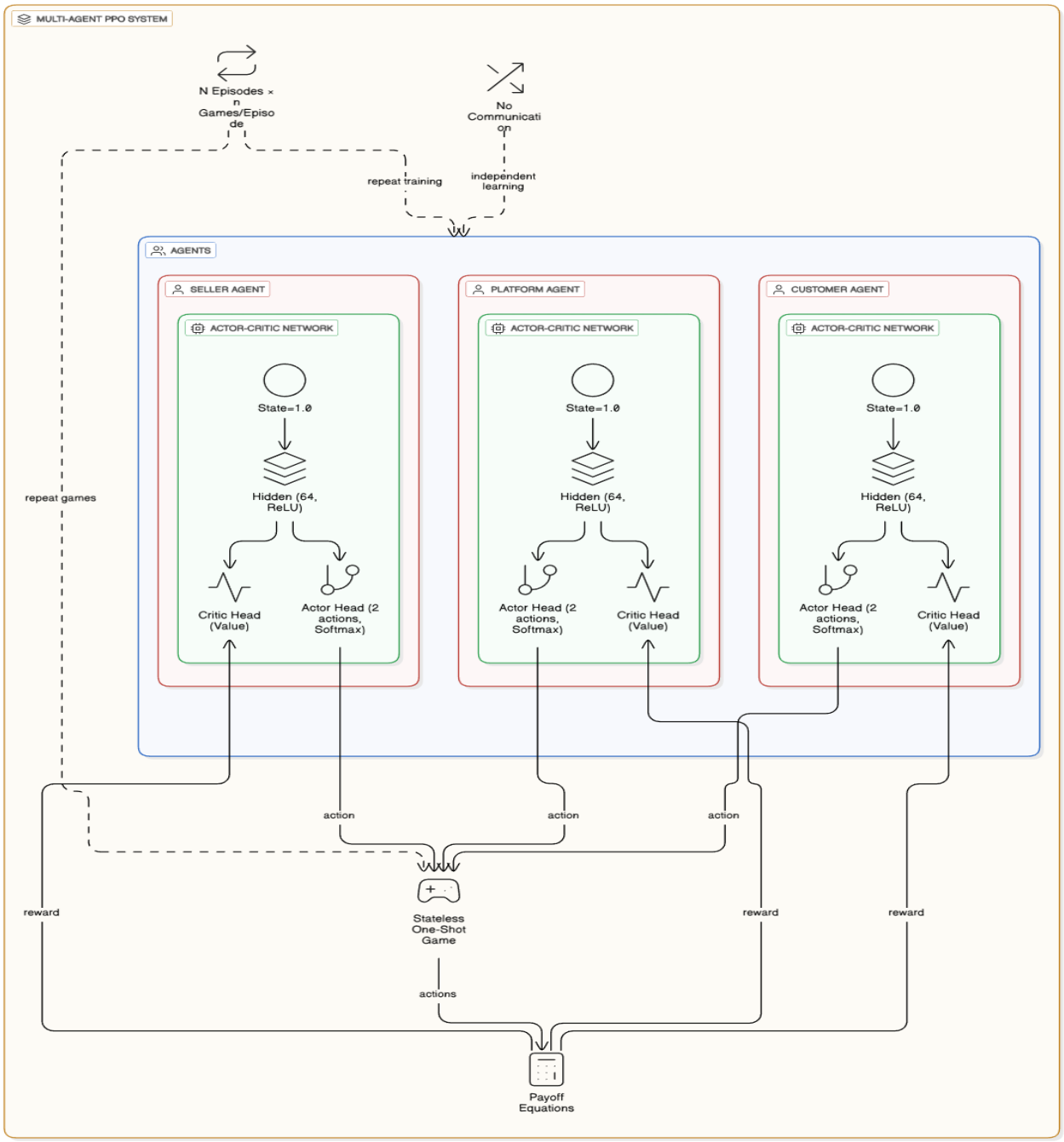

Figure 1. Model Architecture

4. Results and Discussion

The evolutionarily stable strategies (ESS) predicted by traditional evolutionary game theory for both parameter regimes examined were successfully replicated by the multi-agent PPO framework. These findings demonstrate that MARL can be used to effectively approximate theoretical equilibria without requiring explicit analytical derivation of replicator dynamics.

For the first parameter regime,

$$E_4(1,0,1), where - (P - C_m - \alpha P + \beta C_m + S + W_m) < 0,$$

The theoretical equilibrium predicts genuine product sales, consumer acceptance, and platform slack supervision. Convergence of the MARL approach to strategy probabilities closely aligned with the theoretical predictions was observed, indicating near-perfect agreement. For the second parameter regime,

$$E_3(0,0,1),\ where\ -(P - C_m - \alpha P + \beta C_m + S + W_m) < 0,$$

the theoretical equilibrium predicts fake product sales, consumer acceptance, and platform slack supervision. Strong convergence of the MARL results to this alternative equilibrium profile was achieved, confirming the robustness of the approach across different parameter settings. The convergence to $E_4(1,0,1)$, demonstrates that when economic conditions favor genuine products-specifically when price differentials, production costs, and reputation losses make fake products unprofitable-rational sellers naturally gravitate toward honest business practices, even with minimal platform supervision (Table 1).

Table 1. Comparison of Theoretical and MARL Results for Both Cases

Participant	Theoretical ESS	MARL Probability	MARL Strategy
E_4 (1,0,1) – Genuine Products with Slack Supervision			
Seller	1 (Genuine)	0.999987	Genuine (1.000)
E-commerce Platform	0 (Slack)	0.999583	Slack (1.000)
Customer	1 (Accept)	1.000000	Accept (1.000)
E_3 (0,0,1) – Fake Products with Slack Supervision			
Seller	0 (Fake)	0.999927	Fake (1.000)
E-commerce Platform	0 (Slack)	0.999459	Slack (1.000)
Customer	1 (Accept)	0.999999	Accept (1.000)

The alignment with $\boldsymbol{E_3(0,0,1)}$, shows that when the economic penalty for selling counterfeits (e.g. word of mouth loss, $\boldsymbol{W_m}$ = 30 instead of 50 and cost of manufacturing genuine product, $\boldsymbol{C_p}$ = 100) is insufficient to offset the profit advantage of fake products, sellers rationally choose dishonest practices. In both cases, the nearly perfect probabilities (approaching 1.0) indicate that the multi-agent PPO framework correctly identifies strong, stable equilibria where all participants have clear dominant strategies. Overall, the successful replication of both theoretical equilibria validates our MARL approach as a viable alternative to traditional evolutionary game theory methods, demonstrating its capability to handle multiple parameter regimes and converge to the appropriate stable strategies in each scenario.

4.1 Stability Test

To determine the stability and consistency of the learned equilibrium strategies, a stability test was conducted across a series of independent training sessions with different random seeds. Sensitivity of policy convergence to initial conditions is a critical issue in multi-agent reinforcement learning, since stochastic gradient updates and random action sampling during training can potentially lead to convergence toward different local optima or unstable equilibria. The stability analysis used five different random seeds. Each seed was trained independently on the three-player game for 3000 episodes with 100 samples per episode, following the same training protocol described. The strategy-selection probabilities of each agent were recorded at fixed intervals during training to obtain full convergence trajectories. Both parameter regimes analyzed in this study- E_4 (1,0,1) for genuine products and E_3(0,0,1) for fake products-were subjected to this stability test. In both cases, convergence trajectories were plotted in Figure 2- Figure 4 to show how agent strategies evolved from randomized initializations toward equilibrium. Mean and standard deviation statistics of the final converged probabilities across all seeds were used to assess convergence consistency. The primary objective of this analysis was to confirm that the multi-agent PPO framework consistently converges to the same

evolutionarily stable strategy regardless of random initialization, thereby demonstrating the robustness of the learned equilibrium. Low dispersion in the final strategy probabilities across different seeds would indicate stable convergence to a unique ESS, whereas high dispersion would suggest the presence of multiple equilibria or unstable training dynamics. This validation is essential to ensure that the observed equilibrium strategies are genuinely stable points rather than artifacts of specific initial conditions.

4.1.1 Case 1: $E_4(1, 0, 1)$ - Genuine Products with Slack Supervision

For the first case, $E_4(1, 0, 1)$, , the optimal strategies obtained by the multi-agent PPO framework were found to match the theoretical predictions. The seller's optimal strategy was identified as selling genuine products, the customer's optimal action was to accept the product, and the e-commerce platform's strategy was to adopt slack supervision.

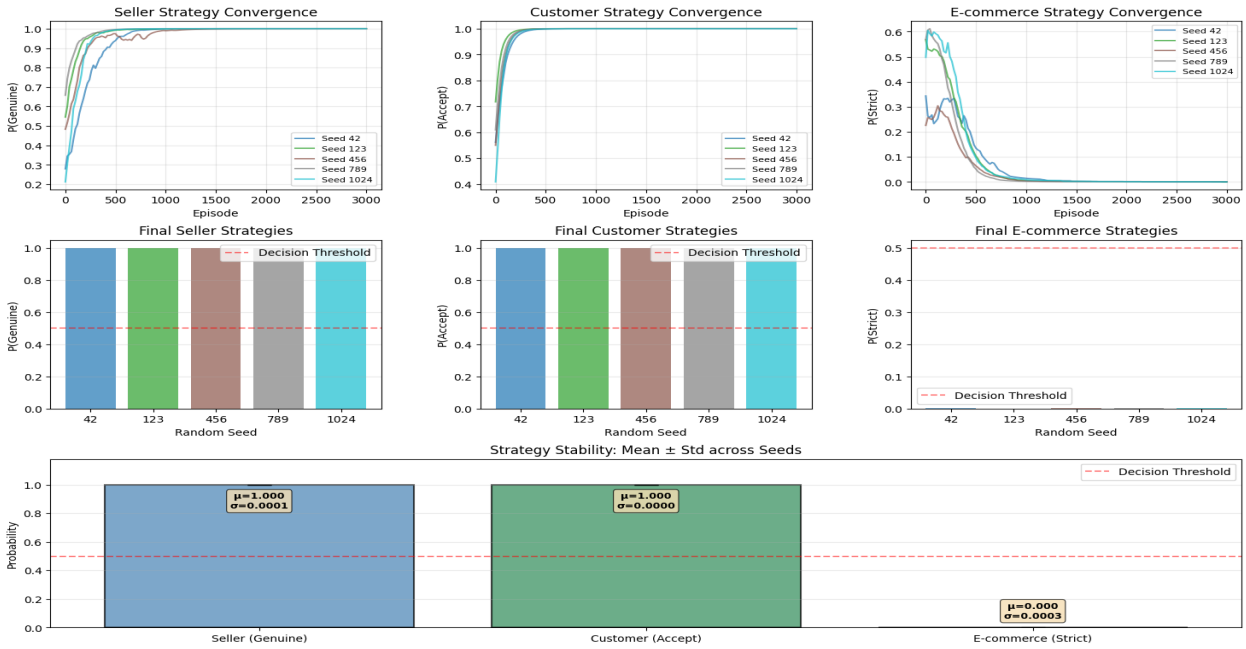

Figure 2. Stability Test Results of $E_4(1, 0, 1)$

Figures 3(a-d) illustrate how the expected payoff of each player was affected when one parameter was varied while all others were kept fixed.

In Figure 3(a), this subfigure demonstrates the impact of product price (P) in the genuine-product equilibrium (E_4). It shows that the seller's optimal strategy shifts at a certain price level (after the price reaches about 175) toward choosing the genuine-product strategy, which maximizes both seller and buyer payoffs. The platform's payoff and its slack-supervision strategy remain constant, indicating a market-based incentive for honest behavior.

In Figure 3(b), this subfigure shows the effect of manufacturing cost (C_m) on payoffs under the genuine-product regime. As increasing production costs reduce the seller's profit, the equilibrium strategies remain stable even at higher costs, demonstrating the robustness of the genuine-product equilibrium against moderate increases in manufacturing cost.

In Figure 3(c), this subfigure illustrates the impact of customer reward (Rc) for genuine items. An increase in Rc directly raises the payoff of customers without affecting the profits or strategies of the seller or the platform. This indicates that in a healthy market, improvements in product quality or reward mechanisms primarily benefit consumers.

In Figure 3(d), this subfigure plots payoffs against the platform's revenue (R_p). As platform revenue increases, only the platform's payoff rises, while seller and customer payoffs remain unchanged. This confirms that platform revenue does not influence the strategic decisions of the seller or customer under this equilibrium.

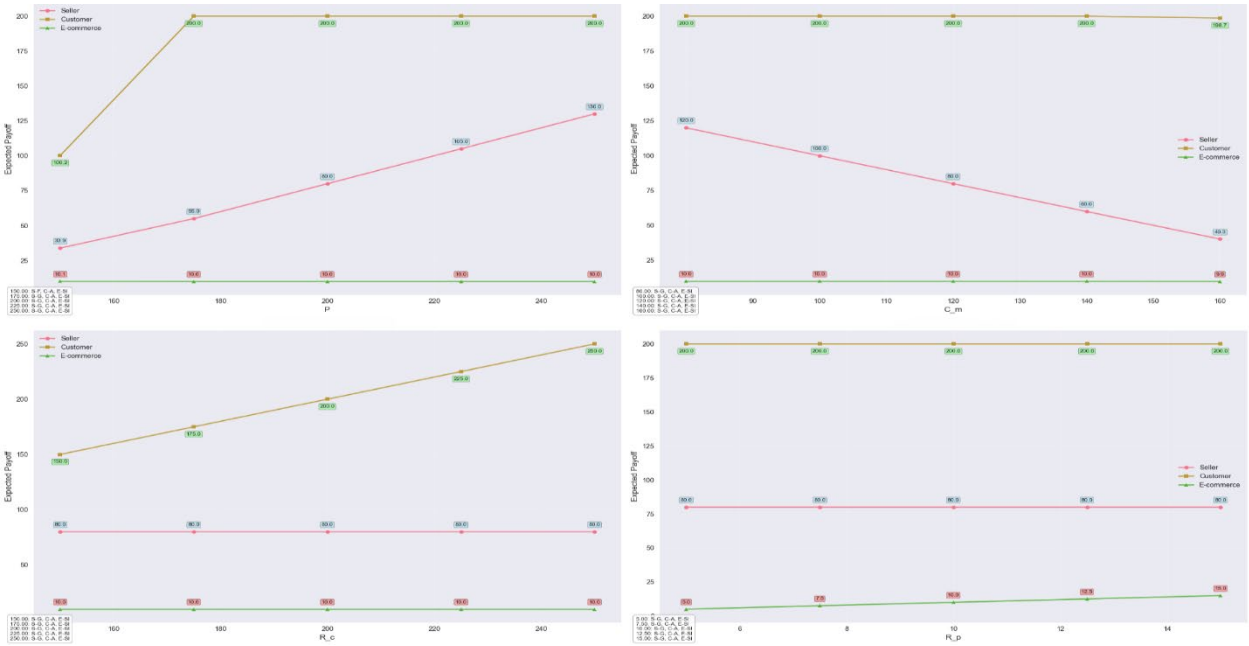

Figure 3. (a) Expected Payoffs vs P, (b) Expected Payoffs vs C_m, (c) Expected Payoffs vs R_c, (d) Expected Payoffs vs R_p

4.1.2 Case 2: $E_3(0, 0, 1)$ - Fake Products with Slack Supervision

For Case 2, $E_3(0, 0, 1)$, the optimal strategy was predicted by both the theoretical model and the MARL framework to involve the seller selling fake products, the customer accepting the products, and the e-commerce platform adopting slack supervision. The parameters P, C_m , $\alpha, \beta, \gamma, C_p, R_c, R_p, W_m$ were varied to examine their effects on payoffs and strategies (Figure 4).

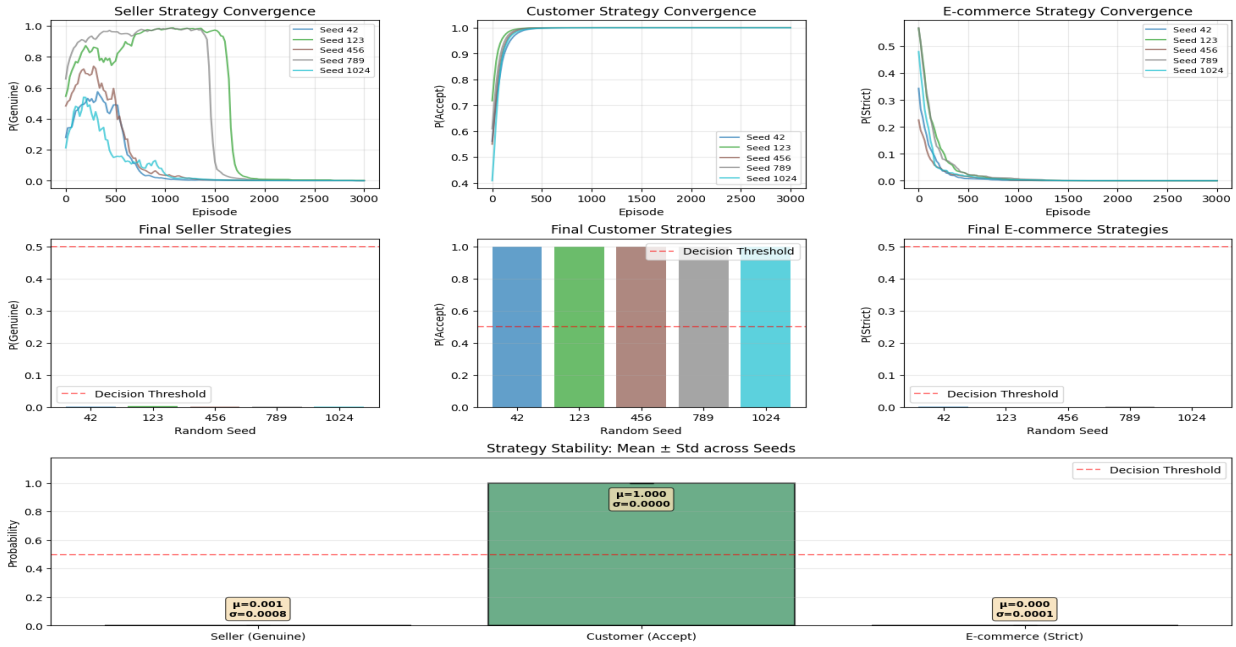

Figure 4. Stability Test Results of $E_3(0, 0, 1)$

In Figure 5(a), this subfigure shows payoffs in the fake-product equilibrium (E_3) as a function of price (P). The seller's profit increases with price, while the customer's and platform's payoffs remain low and fixed. Strategies do not change because supervision costs are high, illustrating a market failure where increasing prices benefit only counterfeit sellers. In Figure 5(b), this subfigure tests the effect of real manufacturing cost (C_m) in the fake-product regime. Sellers shift to genuine products at low manufacturing costs, improving customer welfare. As the cost rises, sellers revert to fake products. A specific cost threshold identifies where sellers switch back to honesty.

In Figure 5(c), this subfigure tests the effect of perceived fake-product value (α). A lower α encourages genuine sales and yields higher customer payoffs. As α increases, fakes become more attractive, prompting the seller to switch back to fake products and reducing customer payoff. This demonstrates how perception directly affects seller behavior.

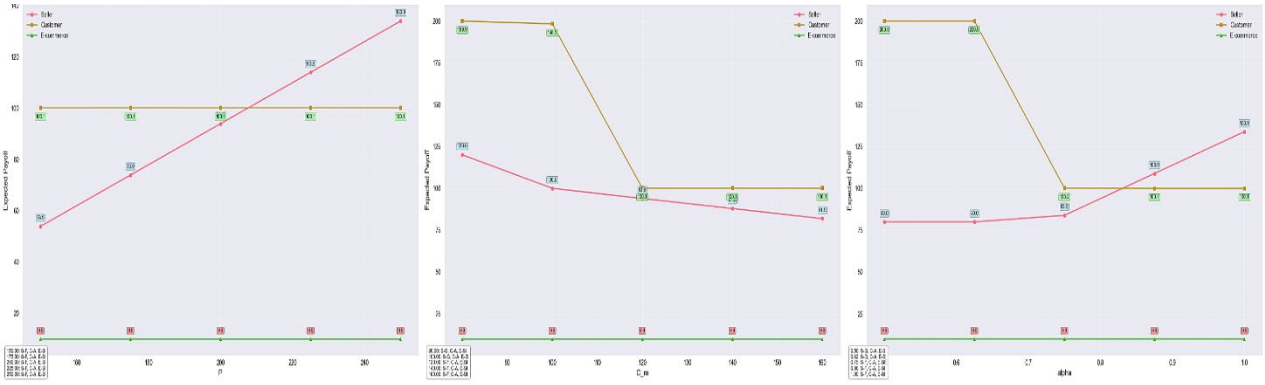

Figure 5. (a) Expected Payoffs vs P, (b) Expected Payoffs vs C_m, (c) Expected Payoffs vs alpha

In Figure 5(d), this subfigure displays the effect of the fake-production cost factor (β). Lower β makes fake production highly profitable, keeping customer payoff low. When β becomes sufficiently high, sellers shift to genuine products, significantly improving customer outcomes. This shows that production-cost policy levers can influence market honesty.

In Figure 5(e), this subfigure changes the relative quality of fakes (γ). As γ increases, customer payoff increases without altering equilibrium strategies. This indicates that improving counterfeit quality alone cannot shift the market away from fakes unless complemented by other economic or regulatory measures.

In Figure 5(f), this subfigure investigates the cost of strict supervision (C_p). Lower costs enable effective platform enforcement, discouraging counterfeits. Once C_p exceeds a certain threshold, the platform stops strict supervision, and sellers return to fake products. This identifies a critical supervision-cost level required for effective oversight.

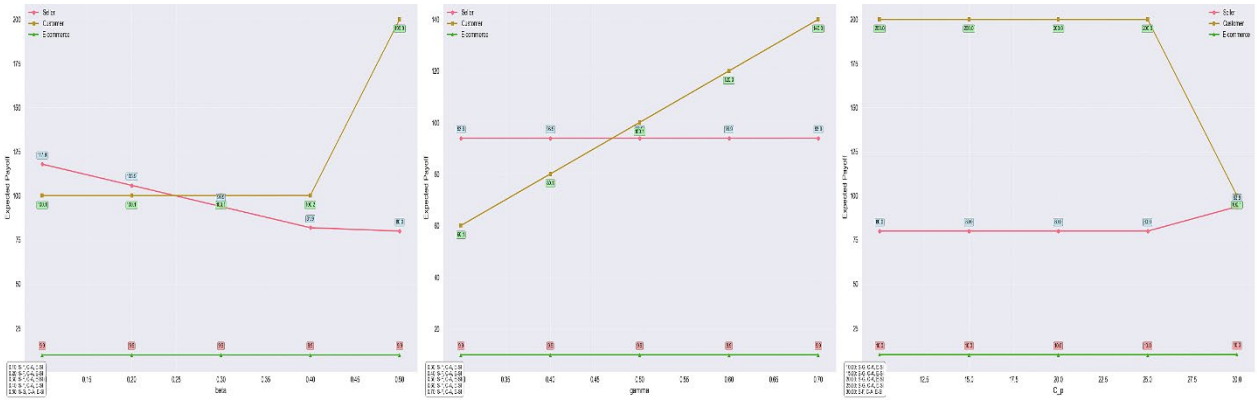

Figure 5. (d) Expected Payoffs vs beta, (e) Expected Payoffs vs gamma, (f) Expected Payoffs vs C_p

In Figure 5(g), this subfigure graphs payoffs against customer reward (R_c). Increasing R_c raises customer payoff without affecting the seller's strategy, which remains focused on fake products. This shows that improving rewards alone cannot break the counterfeit equilibrium.

In Figure 5(h), This subfigure presents the impact of platform revenue (R_p). Higher R_p increases only the platform's payoff and does not influence the strategies of sellers or customers, demonstrating that platform revenue does not provide a direct financial incentive to intervene in counterfeit markets.

Finally, in Figure 5(i), this subfigure shows losses from word-of-mouth (W_m) when selling fakes. Low reputational loss maintains the fake-product equilibrium. As W_m becomes significant, sellers shift to genuine products. This confirms that strong reputation systems can effectively deter counterfeit behavior even without direct monitoring.

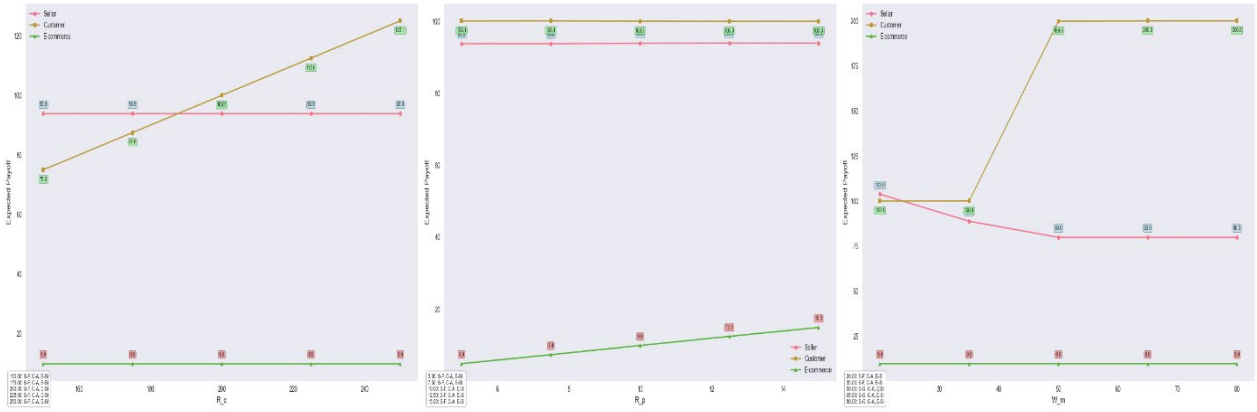

Figure 5. (g) Expected Payoffs vs R_c, (h) Expected Payoffs vs R_p, (i) Expected Payoffs vs W_m

4.2 Limitations and Future Work

Even though the study demonstrated that a multi-agent PPO approach can synthesize the equilibrium strategies, predicted by the evolutionary game theory in the two parameter regimes identified by the original paper, the current range of its application is limited. The rest of the parameter values were not provided, which constrained the generality of the findings; only these two equilibria were analyzed. In addition, the game was modeled as a fully known, stateless, one-step, interaction where the payoff functions are known and fixed, which does not mirror the complexity of sequential, state-dependent, or partially observable, e-commerce situations. Further studies will focus on its generalizability to other equilibria, sequential and dynamic equilibria, scaling to even larger strategy space, payoff uncertainty and different multi-agent reinforcement learning algorithms.

5. Conclusion

The stateless multi-agent Proximal Policy Optimization (PPO) architecture can estimate evolutionarily stable strategies (ESS) in general evolutionary games, thereby connecting the field of computational reinforcement learning with classical game theory. Two analytically derived equilibrium regimes-$E_4(1,0,1)$ and $E_3(0,0,1)$ were tested against the model, and the probabilities of the learned strategies were found to converge to values greater than 0.999, demonstrating that ESS can be recovered without relying on closed-form replicator dynamics. This approach serves as a scalable, simulation-based substitute for analyzing intricate multi-agent interactions. Sensitivity analysis revealed parametric thresholds that are critical in regulating changes in agent behavior, highlighting the usefulness of the framework in mapping strategic landscapes under different conditions. These results support the idea that multi-agent reinforcement learning is an effective method for studying strategic interaction in contexts where classical analytical approaches cannot be applied. However, the research was limited by its one-shot, stateless formulation and by the analysis of only two equilibrium regimes; it also assumed perfect information and fixed payoffs, which may not hold in more dynamic or partially observable environments.

References

Barfuss, W. and Meylahn, J. M., Intrinsic fluctuations of reinforcement learning promote cooperation, Scientific Reports, vol. 13, p. 1309, 2023.

Blanc, M. and Hansen, K. A., Computational complexity of multi-player evolutionarily stable strategies, in Computer Science – Theory and Applications, R. Santhanam and D. Musatov, eds., vol. 12730, Springer International Publishing, pp. 1–17, 2021.

Canese, S., Zecchin, M. and Neely, A., Multi-agent reinforcement learning: A review of challenges and applications, Applied Sciences, vol. 11, no. 11, p. 4948, 2021.

Gabler, V., Goswami, D. and Althoff, M., Decentralized multi-agent reinforcement learning based on hierarchical synergy of actor–critic and Q-learning, Frontiers in Robotics and AI, vol. 11, p. 1229026, 2024.

Guo, H., Zhao, X., Yu, H., Zhang, X. and Li, J., Game analysis of merchants and consumers confronting fakes on e-commerce platforms, Systems Science & Control Engineering, vol. 9, no. 1, pp. 198–208, 2021.

Huang, J. Y., Ding, R., Zhang, Z. and Zhang, S., A multi-level credit assignment method for multi-agent reinforcement learning, Applied Sciences, vol. 12, p. 6938, 2022.

Jia, L., Su, B., Xu, D., Wang, Y., Fang, J. and Wang, J., Policy optimization algorithm with activation likelihood-ratio for multi-agent reinforcement learning, Neural Processing Letters, vol. 56, p. 247, 2024.

Kargar, E., Iraji, S. and Meghdari, A., Multi-agent cooperative recurrent policy optimization, Frontiers in Robotics and AI, vol. 11, p. 1394209, 2024.

Le, L., Evolutionary game analysis for the regulation of merchants selling counterfeits on e-commerce platforms, SHS Web of Conferences, vol. 163, p. 02026, 2023.

Omidshafiei, S., Kramár, J., Hennes, D., et al., AlphaRank: Multi-agent evaluation by evolution, Scientific Reports, vol. 9, p. 9937, 2019.

Trivedi, S. and Hemachandra, N., Multi-agent natural actor-critic algorithms, Dynamic Games and Applications, vol. 13, pp. 1170–1205, 2023.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature, vol. 575, pp. 350–354, 2019.

Wong, A., Leibo, J. Z., Hughes, E., et al., Deep multiagent reinforcement learning: Challenges and directions, Artificial Intelligence Review, 2024.

Zhang, K., Yang, Z. and Başar, T., Decentralized multi-agent reinforcement learning with networked agents: Recent advances, Frontiers of Information Technology & Electronic Engineering, vol. 22, no. 6, pp. 802–814, 2021.

Zhou, Y., Jiang, J., Kilicarslan-Ayvaz, J., Tsourdos, A. and Savvaris, A., Multi-agent reinforcement learning for cooperative lane-change decision-making with delayed global reward and MCTS, Autonomous Intelligent Systems, vol. 2, p. 4, 2022.

Nomenclature

Symbol	Implication
P	the price of a genuine commodity
α	the ratio of the price of a fake commodity to the price of a genuine one (α ∈(0, 1])
C_m	the cost of purchasing a genuine commodity by the seller
β	the ratio of the cost of purchasing a fake commodity by the seller to the cost of purchasing a genuine one (β ∈(0, 1))
B	the loss of the seller due to returning
W_m	the potential loss of the seller brought by negative online word of mouth
C_p	the strict supervision cost of the e-commerce platform
R_p	the e-commerce platform's revenue after the seller successfully selling a commodity
F	economic penalty that the e-commerce platform imposes on the seller after finding a fake commodity
R_c	the utility of the customer brought by a genuine commodity
γ	the ratio of the utility of the customer brought by a fake commodity to the utility by a genuine one (γ ∈(0, 1))
I	the economic compensation paid by the seller to the customer
C_c	returning cost of the customer

Biographies

Mahamudul Hassan Siddique is an undergraduate final-year student in the Department of Industrial and Production Engineering at Bangladesh University of Engineering and Technology (BUET). He is studying the multi-tier supply chain optimization under the reinforcement learning in his thesis. His areas of research concern machine learning, deep learning, natural language processing, large language models, operations research, and optimization. He integrates the use of statistics and the AI techniques to come up with data-driven answers to complicated industrial challenges. His research is based on the interplay of conventional optimization with state-of-the-art computational models in aid of intelligent decision making.

Fahimul Haque is an undergraduate final year student of Industrial and Production Engineering with expertise in statistical learning, machine learning, optimization, and computational intelligence. His research includes nonlinear and explainable machine learning models for suicide mortality prediction using high-dimensional socio-economic data, Bayesian self-supervised MLPs for initialization-free constrained optimization, and multi-agent reinforcement learning (PPO) for evolutionary game equilibrium recovery. In manufacturing systems, he develops hybrid RSM–ML frameworks, symbolic regression, and ensemble models for surface roughness and flank wear prediction in advanced drilling processes. His work emphasizes model interpretability, generalization, and decision-oriented predictive analytics.