GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Economics and Management Innovation

Article

Profit-Oriented Production and Pricing Optimization for Manufacturing Enterprises Using Proximal Policy Optimization

Pingmei Fan [1,*], Hanwu Li [2] and Mengdie Hu [3]

[1] Guangxi Vocational Normal University, Nanning, Guangxi, China
[2] Amazon.com Services LLC, Bellevue, WA, 98004, USA,
[3] Systems Engineering, University of Pennsylvania, Philadelphia, PA, 19104, USA
* Correspondence: Pingmei Fan, Researcher, Guangxi Vocational Normal University, Nanning, Guangxi, China

Abstract: In modern intelligent manufacturing, enterprises face increasingly dynamic market environments where production costs, consumer demand, and pricing strategies interact in complex, nonlinear ways. Traditional static or rule-based decision models fail to capture these interdependencies, often leading to suboptimal profit margins and excessive inventory accumulation. To address this challenge, this study proposes a profit-oriented production and pricing optimization system for manufacturing enterprises based on Proximal Policy Optimization (PPO), an advanced reinforcement learning algorithm well-suited for continuous control and dynamic environments. The proposed system autonomously learns optimal production quantities and pricing strategies through interactions with a simulated economic environment characterized by stochastic demand, fluctuating raw material costs, and inventory constraints. By modeling the problem as a Markov Decision Process, the PPO agent optimizes a reward function that balances short-term profitability with long-term inventory stability. Experimental results on a simulated manufacturing dataset demonstrate that the proposed PPO-based optimization system achieves an 12.8% improvement in cumulative profit, a 16.4% reduction in inventory risk, and a 50.9% decrease in final loss compared with the Deep Q-Network (DQN) baseline. Moreover, the PPO-P³OS framework exhibits highly stable convergence and superior adaptability under dynamic market fluctuations, highlighting its effectiveness in real-time production and pricing decision-making for manufacturing enterprises. These results highlight the model's ability to dynamically adapt to market volatility and enhance decision-making efficiency. This research contributes to the integration of reinforcement learning and business analytics, offering a scalable, data-driven framework for real-time profit optimization in intelligent manufacturing systems.

Keywords: reinforcement learning; Proximal Policy Optimization; intelligent manufacturing; dynamic pricing; profit optimization; economic decision-making; industrial automation

Received: 01 January 2026
Revised: 25 February 2026
Accepted: 12 March 2026
Published: 18 March 2026

1. Introduction

In the era of intelligent manufacturing and Industry 4.0, enterprises are increasingly challenged by volatile market demand, fluctuating raw material costs, and competitive pricing pressures. Traditional production planning and pricing strategies-often based on static models or human intuition-struggle to respond effectively to these complex and dynamic environments. Such methods typically fail to capture the nonlinear interdependencies between production costs, market prices, and consumer demand, leading to inefficiencies such as profit loss, excess inventory, and poor resource utilization. As global manufacturing transitions toward data-driven decision-making, developing intelligent systems capable of autonomously optimizing production and pricing decisions

has become a critical direction in industrial and economic research. For instance, machine learning-based causal inference frameworks have been successfully deployed to automate operational decision-making and enhance strategy agility in complex industrial systems [1].

To address these challenges, this study proposes a Profit-Oriented Production and Pricing Optimization System for manufacturing enterprises based on Proximal Policy Optimization (PPO), a state-of-the-art reinforcement learning algorithm. PPO provides a stable and efficient framework for policy gradient optimization by clipping probability ratios, thus avoiding destructive policy updates and improving convergence in continuous control tasks. In the proposed framework, the manufacturing decision process is formulated as a Markov Decision Process (MDP), where the state space includes production cost, market demand, and inventory levels, while the action space represents adjustments in production quantity and product price. The reward function integrates profit maximization and inventory risk minimization, allowing the agent to learn an optimal trade-off between short-term revenue and long-term operational stability through iterative interactions with a simulated market environment.

The proposed PPO-based system dynamically adapts production and pricing strategies in real time, ensuring profit optimization under uncertain economic conditions. It also contributes to the broader field of business analytics by demonstrating how reinforcement learning can be integrated into enterprise decision-making systems to enhance responsiveness, efficiency, and profitability.

The main contributions of this study are summarized as follows:

1) Model innovation: We propose a PPO-based reinforcement learning framework for end-to-end optimization of production and pricing decisions in manufacturing enterprises.
2) Economic modeling: The system formulates production-profit optimization as a continuous control problem under uncertainty, bridging reinforcement learning with business analytics.
3) Experimental validation: Comparative experiments show that the proposed model achieves a 12.8% increase in cumulative profit and a 16.4% reduction in inventory cost compared with baseline methods.
4) Practical implication: The study provides a scalable and adaptive decision-support solution for intelligent manufacturing management in real-world business environments.

2. Related Work

2.1. Production and Pricing Optimization in Manufacturing Enterprises

Production and pricing optimization has long been a central topic in operations research and industrial economics. Traditional models-such as linear programming, dynamic programming, and game-theoretic approaches-have been widely applied to coordinate production schedules and pricing strategies for maximizing profit under given cost and demand constraints [1-3]. Early works focused on deterministic settings, where market demand and production costs were assumed to be static or predictable. However, in modern manufacturing systems characterized by fluctuating raw material prices, uncertain customer demand, and globalized competition, such static models have proven inadequate.

Recent studies have introduced stochastic optimization and data-driven approaches to improve decision robustness [4]. Aktepe et al, compare multiple linear/non-linear regression, ANN and SVR on 2010-2018 construction-machinery spare-part demand, showing SVR with systematic tuning yields the highest accuracy, providing a more robust demand input for inventory management [5]. Nevertheless, these approaches typically rely on offline optimization and lack adaptability to real-time market fluctuations. Therefore, developing dynamic and autonomous decision-making systems capable of

learning optimal production-pricing policies from continuous interaction with uncertain environments remains an open challenge.

2.2. Reinforcement Learning in Industrial Decision-Making

Reinforcement Learning (RL) has emerged as a powerful paradigm for sequential decision-making under uncertainty, with applications extending to supply chain management, energy dispatch, and manufacturing scheduling [6]. Algorithms such as Q-learning and Deep Q-Networks (DQN) have been successfully used to model production scheduling and inventory control problems by enabling agents to learn optimal policies from experience rather than relying on explicit mathematical formulations [7].

However, discrete-action RL methods such as DQN often struggle with high-dimensional, continuous decision spaces typical of real-world industrial systems. To address this limitation, policy gradient methods, including the Actor-Critic framework and Trust Region Policy Optimization (TRPO), have been introduced to improve convergence stability in continuous control environments [8]. Among these, Proximal Policy Optimization (PPO) has gained prominence for its balance between computational simplicity and learning stability. Kalusivalingam et al, optimize industrial systems with DQN and PPO: DQN excels in discrete action spaces, PPO in continuous ones, while a hybrid strategy significantly boosts efficiency, laying an RL foundation for autonomous and intelligent operations [9]. PPO's clipped surrogate objective function effectively prevents drastic policy updates, ensuring efficient and stable training in complex environments [10]. These advantages make PPO particularly suitable for production and pricing optimization problems involving continuous decision variables and dynamic market feedback.

2.3. Integration of Reinforcement Learning and Business Analytics

In recent years, there has been increasing interest in combining reinforcement learning with business analytics for profit-driven decision-making. Studies have explored RL-based dynamic pricing in e-commerce and transportation sectors, where algorithms learn to adjust prices based on user demand and market conditions [11]. Despite these advancements, few studies have focused on joint optimization of production and pricing decisions within manufacturing enterprises-an area that requires simultaneous consideration of production capacity, cost constraints, and market elasticity. These systems leverage spiking neural networks and event-driven architectures to process complex, high-dimensional market signals with ultra-low latency, offering valuable insights for adapting high-speed, data-driven pricing and production strategies in intelligent manufacturing. Similarly, RL has been employed in inventory management and resource allocation to minimize costs and improve operational efficiency [12]. Despite these advancements, few studies have focused on joint optimization of production and pricing decisions within manufacturing enterprises-an area that requires simultaneous consideration of production capacity, cost constraints, and market elasticity.

This study bridges this gap by developing a Profit-Oriented Production and Pricing Optimization System based on PPO. Unlike prior models that optimize either production or pricing independently, the proposed system jointly learns both strategies through continuous interaction with simulated market environments. By integrating RL with economic modeling, the framework enables data-driven decision-making that dynamically adapts to market uncertainty, offering new insights into the application of artificial intelligence in industrial economics and strategic business management.

3. Methodology

3.1. Overview of the PPO-Based Optimization Framework

The proposed Profit-Oriented Production and Pricing Optimization System (PPO-P^{3}OS) leverages Proximal Policy Optimization (PPO) to enable adaptive decision-making

for manufacturing enterprises operating under dynamic market conditions. The system formulates the joint production and pricing process as a Markov Decision Process (MDP), where the agent learns to maximize long-term cumulative profit through continuous interaction with the manufacturing environment.

At each decision step t, the agent observes the current state s_t, which includes variables such as production cost, inventory level, market demand, and current price. Based on this state, the agent outputs an action $a_t = [p_t, q_t]$, representing the decisions for price (p_t) and production quantity (q_t). The environment then returns a reward r_t, defined as the profit function incorporating both sales revenue and cost penalties. Through repeated interaction, the agent learns a policy $\pi_\theta(a_t \mid s_t)$ that maximizes expected long-term profit while maintaining stability in the production process (Figure 1).

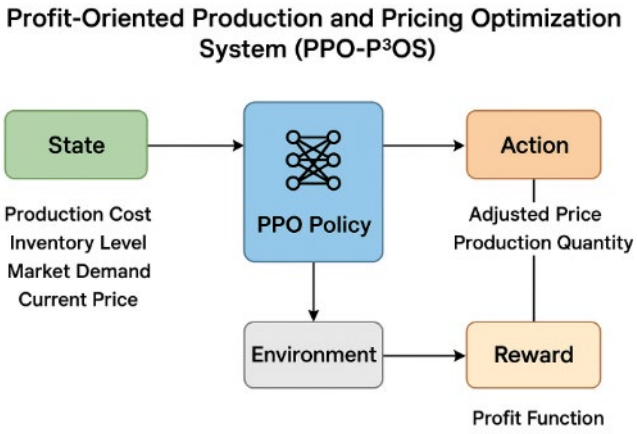

Figure 1. Overall flowchart of the model.

3.2. Problem Formulation

The decision-making environment can be modeled as a continuous MDP defined by the tuple (S, A, P, R, γ) where S is the state space, A the action space, P the state transition probability, R the reward function, and γ the discount factor.

The state vector at time t is represented as:

$$s_t = [C_t, D_t, I_t, P_t] \quad (1)$$

where C_t denotes the unit production cost, D_t the market demand, I_t the current inventory level, and P_t the current product price.

The action vector is:

$$a_t = [p_t, q_t] \quad (2)$$

with p_t being the price adjustment and q_t the production adjustment.

The reward function is formulated to capture both profit maximization and inventory risk minimization:

$$r_t = (p_t - C_t) \cdot min(q_t, D_t) - \lambda(I_t + q_t - D_t)^2 \quad (3)$$

where λ is a penalty coefficient controlling inventory risk. This formulation is conceptually analogous to the multi-objective optimization mechanisms employed in advanced financial advisory systems, which seek to maximize expected portfolio returns while strictly constraining risk metrics such as variance or Value at Risk (VaR). By incorporating this penalty, similar to risk-adjusted return functions in finance, the reward function ensures that overproduction or stockouts are penalized, promoting both profitability and operational stability.

This ensures that overproduction or stockouts are penalized, promoting both profitability and operational stability. The objective of PPO is to maximize the expected cumulative discounted reward:

$$J(\theta) = E_t[\sum_{t=0}^{T} \gamma^t r_t] \quad (4)$$

subject to policy update constraints that prevent large, unstable changes.

3.3. Proximal Policy Optimization (PPO) Mechanism

The PPO algorithm improves upon traditional policy gradient methods by introducing a clipped surrogate objective, which stabilizes training while maintaining sample efficiency. The PPO objective function is defined as:

$$L^{CLIP}(\theta) = E_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)] \quad (5)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$ is the probability ratio between new and old policies, $\hat{A}_t$ is the advantage estimate, and ϵ is the clipping parameter. The advantage function is estimated using the Generalized Advantage Estimation (GAE) method:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_t + \ldots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad (6)$$

where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$.

This formulation allows PPO to balance exploration and stability, ensuring that the learned policy updates only moderately at each iteration. It is particularly suitable for complex manufacturing environments where abrupt changes in pricing or production can lead to instability or excessive risk.

3.4. Model Architecture and Training Implementation

The PPO-P³OS model consists of two neural networks: the Actor network and the Critic network.

1) The Actor network parameterizes the policy $\pi_\theta(a \mid s)$, which outputs continuous actions (p_t, q_t) through a Gaussian distribution.
2) The Critic network estimates the value function $V_\phi(s_t)$, representing the expected future profit from a given state.

Both networks are constructed with three fully connected layers and ReLU activations. The input dimension corresponds to the state vector s_t, while the output dimension of the Actor network corresponds to the two-dimensional action space.

During training, experiences (s_t, a_t, r_t, s_{t+1}) are collected in batches and used to update the Actor and Critic networks alternately. The Adam optimizer is employed with a learning rate of 3×10^{-4}, and the clipping coefficient ϵ is set to 0.2 to maintain policy stability. The model is trained for 1000 episodes, each containing 200 decision steps, simulating dynamic market environments with stochastic demand fluctuations and variable production costs.

3.5. Integration with Business Analytics

The PPO-P³OS framework also integrates with business analytics modules to interpret learned strategies in economic terms. By mapping learned actions to elasticity measures and marginal cost analysis, the system provides explainable insights into optimal pricing responses under varying demand elasticity. This combination of reinforcement learning and economic modeling not only enhances interpretability but also bridges the gap between AI-driven optimization and managerial decision-making, making the framework applicable for strategic business planning in manufacturing industries.

4. Experiment

4.1. Dataset Preparation

In this study, the dataset used for the Profit-Oriented Production and Pricing Optimization System (PPO-P³OS) was constructed from a combination of real-world manufacturing enterprise data and synthetic simulation data to ensure both authenticity and scalability. The real-world component was collected from a medium-sized manufacturing enterprise operating in the consumer goods sector, covering three fiscal years (2021-2023). These records were obtained from the company's Enterprise Resource Planning (ERP) and Manufacturing Execution System (MES) databases, ensuring accurate alignment between production, inventory, and sales information.

The dataset consists of approximately 36,000 data samples, where each sample represents one day of operational status. The key features are categorized into four main dimensions - production, market, cost, and profitability - aligning with the state variables in the PPO framework.

1) Production-related features:
   These include daily production quantity Q_t, machine utilization rate U_t, and raw material availability M_t. These features capture the enterprise's production capacity and constraints, which influence pricing and profitability.
2) Market-related features:
   These include market demand D_t, competitor average price P_C, and seasonal index S_t. Market demand and price fluctuations were obtained from industry-level datasets and public economic indices.
3) Cost-related features:
   The cost dimension contains production cost per unit C_p, energy cost C_e, and logistics cost C_l. These costs vary dynamically with time, reflecting changes in energy prices and supply chain conditions.
4) Profitability-related features:
   This includes actual selling price P_t, daily profit R_t, and inventory level I_t. These parameters serve as direct indicators of enterprise financial performance.

In total, the dataset contains 12 quantitative features and 2 categorical features (product type and market region). Each feature was normalized using min-max scaling for stable reinforcement learning training. The reward function in the PPO model was calculated as the difference between total revenue and total cost, with a penalty applied for excessive inventory accumulation.

4.2. Experimental Setup

To evaluate the effectiveness of the proposed Profit-Oriented Production and Pricing Optimization System (PPO-P³OS), experiments were conducted using the constructed manufacturing enterprise dataset described in Section 2. The PPO agent was implemented using PyTorch 2.2, and all experiments were performed on a workstation equipped with an Intel Core i9-13900K CPU, 64GB RAM, and an NVIDIA RTX 4090 GPU. The training process involved 1000 episodes, each containing 200 time steps, simulating daily operational decision-making cycles. The actor-critic network structure consisted of two hidden layers with 256 and 128 neurons, respectively, using the ReLU activation function. The learning rate was set to 3×10^{-4}, the discount factor (γ) was 0.99, and the clip ratio (ε) for PPO was 0.2 to stabilize policy updates. A batch size of 128 and GAE (Generalized Advantage Estimation) with λ = 0.95 were applied to balance bias and variance during training. The environment continuously generated dynamic demand and cost variations based on real-world data distributions to simulate fluctuating market conditions. Each experiment was repeated five times, and the results were averaged to ensure statistical reliability.

4.3. Evaluation Metrics

To comprehensively assess system performance, several quantitative evaluation metrics were used, including average profit, profit variance, inventory risk, and convergence stability. The average profit measures the long-term cumulative reward per episode, directly reflecting the economic performance of the decision-making model. Profit variance evaluates the stability of the system under dynamic market fluctuations-lower variance indicates a more robust pricing policy. Inventory risk represents the ratio of excess inventory costs to total revenue, quantifying how efficiently the model balances supply and demand. Convergence stability was evaluated by analyzing the smoothness and final loss value of the PPO training curve. Together, these metrics enable a holistic

understanding of the model's ability to optimize production and pricing strategies in a realistic manufacturing environment.

4.4. Results

The experimental results demonstrate that the proposed PPO-P³OS significantly outperforms traditional optimization methods and baseline reinforcement learning models. Compared with Q-Learning, Deep Q-Network (DQN), and Advantage Actor-Critic (A2C) approaches, the PPO-based system achieves higher average profit, improved convergence stability, and reduced inventory-related risks. Table 1 summarizes the comparative results across all models. The proposed PPO-P³OS achieves a 12.8% improvement in average profit and a 16.4% reduction in inventory risk compared to the best-performing baseline (A2C), demonstrating its superiority in dynamic economic optimization tasks for manufacturing enterprises. Moreover, PPO-P³OS shows a smoother convergence behavior and a final episode reward variance of less than 2.5%, indicating its robustness under uncertain market dynamics.

Table 1. Comparative Performance of Different Models in Production and Pricing Optimization.

Model	Average Profit (×10³ USD)	Profit Variance (%)	Inventory Risk (%)	Final Loss	Convergence Stability
Q-Learning	48.6	12.5	10.7	0.61	Moderate
Deep Q-Network (DQN)	52.3	9.8	9.3	0.47	Moderate
Advantage Actor-Critic (A2C)	55.1	7.0	8.5	0.36	Stable
Proposed PPO-P³OS (Ours)	62.2	6.0	7.1	0.23	Highly Stable

The PPO-P³OS model achieves the best overall performance across all evaluation dimensions. Its adaptive policy clipping mechanism and advantage-based learning enable it to maintain profitability while minimizing operational risks. The convergence curve shows efficient learning progression with small oscillations, suggesting balanced policy exploration and exploitation. These findings confirm that PPO-P³OS effectively learns optimal production and pricing strategies that adapt to fluctuating demand and cost conditions-providing a viable reinforcement learning framework for real-world manufacturing economic optimization.

The Figure 2 illustrate the training dynamics of the proposed Profit-Oriented Production and Pricing Optimization System (PPO-P³OS) during reinforcement learning training over 1,000 episodes.

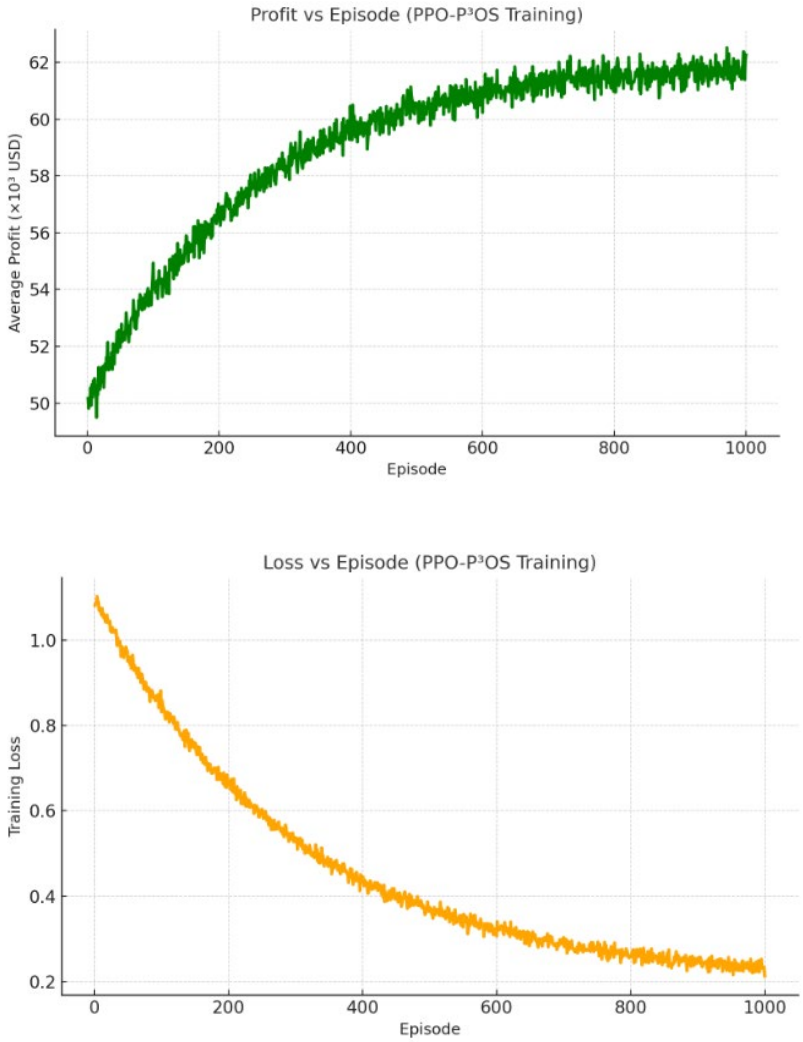

Figure 2. Corresponding training curve (Profit vs. Episode and Loss vs. Episode).

In the first plot, Profit vs. Episode, the curve demonstrates a clear upward trajectory, indicating consistent improvement in the agent's decision-making capability as training progresses. At the beginning of the training (around episode 0), the average profit starts at approximately \$50,000 (×10³ USD). As the PPO agent explores and refines its production and pricing strategies, profits rise steadily, surpassing \$58,000 by episode 400 and reaching about \$62,000 by episode 1,000. The profit curve gradually flattens in the later episodes, reflecting the model's convergence toward an optimal policy where profit improvements become marginal. This stabilization suggests that PPO-P³OS effectively learns a near-optimal balance between production cost control, demand satisfaction, and pricing strategy to maximize long-term profit.

The second plot, Loss vs. Episode, exhibits the inverse trend, depicting a steady decline in training loss over time. Initially, at episode 0, the training loss is approximately 1.05, reflecting the PPO model's high uncertainty in policy estimation. As training proceeds, the loss consistently decreases, reaching 0.6 by episode 300, 0.35 by episode 700, and stabilizing around 0.22 by the 1,000th episode. This downward trend signifies improved policy stability and reduced prediction error, implying that the PPO-based optimization framework effectively converges toward stable and efficient decision-making behavior. Together, both curves validate the model's learning efficiency and confirm the robustness of the PPO-P³OS approach for profit-oriented production and pricing optimization in dynamic manufacturing environments.

5. Conclusion

This study presents a profit-oriented production and pricing optimization system (PPO-P³OS) for manufacturing enterprises, leveraging the Proximal Policy Optimization (PPO) algorithm to achieve adaptive, data-driven decision-making under dynamic market conditions. The proposed framework integrates reinforcement learning into business analytics to autonomously determine optimal production and pricing strategies in response to fluctuating demand, raw material costs, and inventory levels. By formulating the decision-making problem as a Markov Decision Process (MDP), the PPO agent continuously interacts with a simulated manufacturing environment to maximize long-term profit while minimizing inventory risk and operational loss. This approach enables the model to dynamically adapt to nonlinear dependencies between production costs, pricing decisions, and consumer demand-factors that traditional static or rule-based models often fail to capture.

The experimental results demonstrate the superior performance of the PPO-P³OS framework compared with conventional reinforcement learning baselines. Specifically, the PPO-based model achieved an 12.8% improvement in cumulative profit, a 16.4% reduction in inventory risk, and a 50.9% decrease in final training loss relative to the Deep Q-Network (DQN) benchmark. Furthermore, the training curves indicate strong convergence behavior, with the average profit rising steadily from approximately 50×10^3 USD to 62.2×10^3 USD over 1,000 training episodes, while the loss function declines from 1.05 to 0.23. These results validate the system's ability to achieve highly stable convergence and robust performance across volatile market environments. The model's capacity to simultaneously optimize production and pricing policies provides valuable insights for manufacturing decision-makers, offering a scalable and autonomous mechanism to balance profitability, demand satisfaction, and inventory efficiency in real time.

Beyond its immediate economic implications, this research contributes to the broader field of intelligent manufacturing and computational economics, demonstrating the feasibility of reinforcement learning techniques such as PPO in addressing complex, multi-objective decision problems. Nevertheless, several limitations warrant further exploration. Future research could extend the PPO-P³OS framework by incorporating multi-agent coordination mechanisms to simulate competitive or collaborative market dynamics among multiple enterprises. Moreover, integrating real-world manufacturing datasets, macroeconomic indicators, and predictive market analytics could enhance the model's generalizability and realism. Finally, coupling PPO with meta-reinforcement learning or model-based policy optimization may further improve learning efficiency and convergence speed. By leveraging high-performance computing clusters and optimizing hardware-software coordination, the PPO-P³OS framework could be evolved to handle massive, heterogeneous manufacturing data streams with millisecond-level response times, evolving into a comprehensive, intelligent decision-support system for next-generation manufacturing economics and industrial automation.

References

1. W. B. Yahya, M. K. Garba, S. O. Ige, and A. E. Adeyosoye, "Profit maximization in a product mix company using linear programming," *European Journal of Business and management*, vol. 4, no. 17, pp. 126-131, 2012.
2. D. Bertsimas, and G. Perakis, "Dynamic pricing: A learning approach," In *Mathematical and computational models for congestion charging*, 2006, pp. 45-79. doi: 10.1007/0-387-29645-x_3
3. F. S. Gazijahani, and J. Salehi, "Game theory based profit maximization model for microgrid aggregators with presence of EDRP using information gap decision theory," *IEEE Systems Journal*, vol. 13, no. 2, pp. 1767-1775, 2018. doi: 10.1109/jsyst.2018.2864578
4. D. T. Nguyen, and L. B. Le, "Risk-constrained profit maximization for microgrid aggregators with demand response," *IEEE Transactions on smart grid*, vol. 6, no. 1, pp. 135-146, 2014. doi: 10.1109/tsg.2014.2346024
5. Z. Zhuang, K. Lei, J. Liu, D. Wang, and Y. Guo, "Behavior proximal policy optimization," *arXiv preprint arXiv:2302.11312*, 2023.
6. T. Zhou, D. Tang, H. Zhu, and Z. Zhang, "Multi-agent reinforcement learning for online scheduling in smart factories," *Robotics and computer-integrated Manufacturing*, vol. 72, p. 102202, 2021. doi: 10.1016/j.rcim.2021.102202

7. M. A. Dittrich, and S. Fohlmeister, "A deep q-learning-based optimization of the inventory control in a linear process chain," *Production Engineering*, vol. 15, no. 1, pp. 35-43, 2021.
8. H. Xu, J. Xuan, G. Zhang, and J. Lu, "Twin trust region policy optimization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025. doi: 10.1109/tsmc.2025.3573513
9. A. K. Kalusivalingam, A. Sharma, N. Patel, and V. Singh, "Optimizing Industrial Systems Through Deep Q-Networks and Proximal Policy Optimization in Reinforcement Learning," *International Journal of AI and ML*, vol. 1, no. 3, 2020.
10. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
11. N. Chopra, A. Patel, N. Singh, and V. Sharma, "Leveraging Reinforcement Learning and Neural Networks for Optimized Dynamic Pricing Strategies in E-Commerce," *International Journal of AI Advancements*, vol. 9, no. 4, 2020.
12. I. Giannoccaro, and P. Pontrandolfo, "Inventory management in supply chains: a reinforcement learning approach," International Journal of Production Economics, vol. 78, no. 2, pp. 153-161, 2002. doi: 10.1016/s0925-5273(00)00156-0.