

Article

Trustworthy Reinforcement Learning for AI-Driven Urban Decision-Making: Sustainable Dynamic Pricing and Resource Optimization for Smart City Operations

Žydrūnas Bautronis ^{1,2,*} and Robertas Alzbutas ^{1,2,*} ¹ Lithuanian Energy Institute, Breslaujos Str. 3, LT-44403 Kaunas, Lithuania; zydrunas.bautronis@lei.lt² Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Studentų Str. 50, LT-51368 Kaunas, Lithuania

* Correspondence: robertas.alzbutas@ktu.lt

Abstract

Rapid urbanization, increasing demand variability, and digitalisation of urban infrastructure require intelligent decision-support systems that can improve sustainability, resilience, and operational efficiency in smart city operations. Artificial intelligence (AI) and data-driven optimization offer strong potential for adaptive urban services, including dynamic pricing, demand response, resource allocation, and energy-aware management. However, many reinforcement learning applications still focus mainly on short-term performance while giving limited attention to transparency, fairness, stability, and accountability. This study proposes a trustworthy reinforcement learning framework for AI-driven urban decision-making, using sustainable dynamic pricing and resource optimization as mechanisms for adaptive and responsible decision-making. A custom reinforcement learning environment was developed using historical e-commerce transactional data as a methodological proxy to simulate interactions among demand, resource or inventory availability, service categories, price elasticity, and changing market conditions. Three reinforcement learning algorithms, namely Deep Q-Network, Proximal Policy Optimization, and Advantage Actor–Critic, were evaluated under comparable experimental conditions. Performance was assessed using profitability, decision stability, fairness-oriented pricing behavior, decision consistency, and interpretability. To improve transparency, trajectory-based policy audits and SHapley Additive exPlanations were applied to identify the main factors influencing pricing decisions. The results show that the Deep Q-Network agent achieved the most balanced performance, increasing total profit by 12.58% while recording no unethical price increases under low-demand conditions. Explainability analysis showed that stock or resource levels, demand shifts, and price elasticity were the strongest positive drivers of pricing actions, whereas inventory hoarding and unfavorable price increases reduced decision quality. The findings indicate that reinforcement learning can support sustainable and resilient urban decision-making when optimization objectives are combined with trustworthy AI principles. The proposed framework provides a practical basis for accountable AI-based decision-support systems in smart city operations, including demand-responsive services, resource optimization, sustainable dynamic pricing, and energy-aware management.



Academic Editor: Koichi Murata

Received: 28 June 2026

Revised: 12 August 2026

Accepted: 18 August 2026

Published: 2 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: trustworthy artificial intelligence; reinforcement learning; AI-driven urban decision-making; smart city operations; sustainable dynamic pricing; resource optimization; energy management; explainable AI; SHAP; demand response

1. Introduction

Rapid urbanization, increasing pressure on infrastructure, climate-related challenges, and continuously changing demand patterns are forcing cities to adopt more adaptive and intelligent operational systems. Modern cities increasingly rely on artificial intelligence, Internet of Things (IoT) infrastructure, and data-driven optimization. These technologies can improve sustainability, resilience, and efficiency in transportation, energy management, logistics, and resource allocation. As smart city ecosystems become increasingly interconnected and dynamic, traditional static optimization approaches are often unable to respond efficiently to evolving operational conditions.

Among various AI methods, reinforcement learning (RL) has attracted growing attention due to its ability to learn sequential decision-making policies through continuous interaction with the environment. Unlike rule-based or purely predictive systems, RL agents dynamically adapt their behavior according to environmental feedback and long-term reward optimization. This makes reinforcement learning particularly suitable for complex urban operations characterized by uncertainty, non-stationary demand, and resource constraints. Recent studies have demonstrated successful RL applications in areas such as dynamic pricing, smart mobility systems, energy demand-response optimization, ride-sharing coordination, and electric vehicle charging management.

Despite these advantages, the deployment of reinforcement learning systems in real-world operational environments remains associated with important trustworthiness challenges. Many RL-based optimization systems prioritize short-term reward maximization while paying limited attention to transparency, explainability, fairness, stability, and accountability. In urban systems, such limitations may lead to socially undesirable outcomes, including unfair pricing behavior, unstable resource allocation, and opaque automated decision-making processes that reduce public trust and complicate regulatory oversight.

Consequently, the concept of trustworthy AI has become increasingly important in AI-driven operational systems. Trustworthy AI extends beyond predictive performance and incorporates principles such as explainability, fairness, robustness, accountability, and transparency. These principles are particularly relevant in smart city environments because automated decisions may directly affect citizens, public services, infrastructure utilization, and operational sustainability. Ensuring that reinforcement learning systems remain interpretable, stable, and ethically aligned therefore represents a critical challenge for future AI-driven urban operations.

Dynamic pricing represents a suitable domain for evaluating trustworthy reinforcement learning behavior because it requires continuous adaptation to changing demand conditions while simultaneously balancing operational efficiency and socially acceptable decision-making. Although dynamic pricing has traditionally been associated with e-commerce and revenue optimization, similar sequential decision-making structures increasingly appear in smart city operations, including adaptive transportation pricing, congestion management, district energy balancing, electric vehicle charging coordination, and demand-responsive urban services.

The existing literature demonstrates strong optimization capabilities of RL-based pricing systems; however, most studies primarily focus on profitability and revenue maximization. Comparatively limited attention has been given to explainability mechanisms, fairness-oriented evaluation, policy stability, and ethical operational behavior. Furthermore, relatively few studies combine reinforcement learning optimization with explainability techniques and trustworthiness-oriented evaluation in dynamic decision-making environments.

Therefore, this study proposes a trustworthy reinforcement learning framework for AI-driven urban decision-making using sustainable dynamic pricing and resource optimization

as representative operational mechanisms. A custom RL environment was developed using historical operational and demand-related data to simulate interactions among demand fluctuations, resource availability, elasticity, and dynamic pricing decisions. Three reinforcement learning architectures, namely Deep Q-Network (DQN), Proximal Policy Optimization (PPO), and Advantage Actor–Critic (A2C), were evaluated under identical experimental conditions.

Unlike conventional RL studies, the proposed framework evaluates not only optimization performance but also fairness, explainability, policy stability, and decision trajectories. Additional transparency was achieved through SHapley Additive exPlanations (SHAP), policy interpretation analysis, and temporal cross-validation. The framework was designed not only to maximize operational efficiency but also to encourage stable, interpretable, and ethically aligned decision-making behavior.

The results demonstrate that reinforcement learning can improve operational performance while simultaneously maintaining important trustworthy AI properties. The proposed framework achieved a 12.58% increase in cumulative operational profit while avoiding unethical low-demand price increases and maintaining stable adaptive behavior across varying operational conditions. These findings suggest that reinforcement learning can provide a practical foundation for accountable AI-driven decision-support systems in sustainable smart city operations.

The main contributions of this study are threefold. First, the study proposes a reinforcement learning-based dynamic pricing framework that integrates operational efficiency with trustworthiness-oriented constraints. Second, it evaluates DQN, PPO, and A2C not only through profitability and reward-based metrics but also through behavioral stability, fairness-oriented pricing behavior, and explainability. Third, it demonstrates how trajectory auditing, SHAP analysis, policy interpretation techniques, and temporal cross-validation can be combined to assess whether RL-based decision systems are suitable for responsible deployment in adaptive urban operations.

2. Literature Review

2.1. *Why Trustworthy AI Matters?*

Artificial Intelligence (AI) refers to computational systems capable of performing tasks traditionally associated with human intelligence, including learning, decision-making, and problem-solving [1,2]. The rapid advancement of AI technologies has been driven by improvements in computational power, large-scale data availability, and increasingly sophisticated machine learning algorithms [3]. As AI systems become more deeply integrated into operational processes and automated decision-making, concerns regarding transparency, fairness, accountability, and reliability have intensified. Consequently, the concept of trustworthy AI has gained increasing importance in both academic research and practical applications [4].

Trustworthy AI extends beyond predictive performance and incorporates principles such as explainability, transparency, accountability, fairness, robustness, safety, and data governance [4]. These principles are particularly important in high-impact operational environments, where opaque or biased AI decisions may produce significant economic, social, or ethical consequences. Recent trustworthy AI frameworks emphasize that transparency and explainability should be accompanied by human oversight, technical robustness, fairness assessment, and appropriate governance mechanisms [4]. In AI-driven control systems, explainability is also increasingly considered an important design requirement that can improve transparency, accountability, and human understanding of automated decision-making [5]. At the same time, algorithmic bias remains a major concern, as AI models trained on incomplete, unbalanced, or historically biased data may reproduce

discriminatory or unfair outcomes [6]. Furthermore, trustworthy AI depends on reliable and appropriately governed data, as data quality and data governance directly affect the reliability and accountability of AI-supported decisions [4].

The importance of trustworthy AI becomes even more pronounced in reinforcement learning (RL) environments due to the adaptive and sequential nature of RL decision-making. Unlike traditional supervised learning approaches, RL agents continuously interact with dynamic environments and optimize actions through long-term reward maximization. Although this adaptive capability enables efficient optimization under changing conditions, it may also produce unstable, economically irrational, or ethically problematic behaviors if appropriate constraints and evaluation mechanisms are not incorporated into the learning process. Such risks are particularly relevant in RL-based recommendation systems, dynamic pricing applications, and autonomous operational systems, where automated decisions directly influence user behavior, resource allocation, and public trust [6].

Recent research increasingly treats explainability and fairness as integral components of trustworthy AI rather than as secondary evaluation criteria. Explainable-by-design approaches in deep reinforcement learning demonstrate that transparency mechanisms can be incorporated directly into AI-driven control architectures, while fairness-constrained RL pricing studies show that revenue optimization can be combined with explicit restrictions on discriminatory or temporally inconsistent pricing [5,7]. Consequently, trustworthy reinforcement learning should be viewed not only as a technical optimization problem but also as a governance and sustainability challenge requiring a balance between performance, fairness, transparency, and accountability.

Table 1 summarizes the main dimensions commonly associated with trustworthy AI frameworks and their relevance for adaptive AI-driven operational systems.

Table 1. Main dimensions of trustworthy AI frameworks and their relevance for adaptive AI-driven operational systems.

Dimension	Definition	Key Requirements
Data Quality	The foundational integrity of input information.	Reliability, completeness, and continuous monitoring.
Safety	The resilience of the AI agent in operational contexts.	Robustness in dynamic environments and fail-safe mechanisms.
Bias Mitigation	The pursuit of algorithmic fairness and equity.	Identifying, measuring, and reducing socio-technical biases.
Accountability	The governance structure of the AI system.	Assigning responsibility and ensuring auditability of decisions.
Transparency	The traceability of the system's internal logic.	Openness regarding system design, data provenance, and logic.
Explainability	The interpretability of model outputs for humans.	Human-understandable rationale for specific model behaviors.

2.2. Reinforcement Learning in Dynamic Pricing Systems

Dynamic pricing has become an increasingly important strategy in modern operational and commercial environments due to rapidly changing market conditions, fluctuating demand, and growing competition within digital ecosystems. Traditional pricing approaches are commonly based on predefined pricing rules, historical demand patterns, inventory-level adjustments, or competitor pricing analysis. Although such approaches remain widely used in practice, they often demonstrate limited adaptability and reduced responsiveness under highly dynamic and uncertain market conditions.

In contrast, reinforcement learning methods continuously interact with the environment and adapt pricing decisions according to changing operational conditions. By utilizing historical and dynamically updated environmental information, RL models can optimize long-term pricing and inventory decisions through iterative learning processes [8]. Such approaches improve pricing responsiveness while enabling more adaptive and context-aware pricing mechanisms based on consumer behavior, demand fluctuations, and operational constraints.

Recent studies demonstrate that reinforcement learning algorithms can significantly improve pricing optimization performance across multiple domains. RL-based pricing and resource scheduling have been investigated in energy systems [9], data-oriented markets [10], electric vehicle charging [11], airline revenue management [12], ride-sharing services [13], and perishable-product pricing and inventory management [8]. More recent studies have extended this direction toward fairness-constrained dynamic pricing, where deep reinforcement learning is used to optimize hotel pricing while explicitly imposing group and temporal fairness constraints [7]. Deep reinforcement learning (DRL)-based dynamic pricing has also been applied to multi-modal delivery systems, where pricing decisions influence customer choices and support the allocation of transportation resources across available delivery modes [14].

Recent MDPI studies further support the relevance of reinforcement learning and dynamic pricing in demand-responsive energy and smart-grid environments. Ismail and Baysal [15] developed an actor–critic reinforcement learning approach for dynamic pricing based on demand response in energy management, while Salazar et al. [16] proposed a reinforcement-learning-based pricing and incentive strategy for smart-grid demand response. Complementary work on incentive-based dynamic pricing from the customer perspective also demonstrates how real-time price signals and dynamic incentives can support load management in smart-grid systems [17].

However, despite the strong optimization capabilities demonstrated by RL systems, most existing studies primarily focus on profitability and revenue maximization. Comparatively limited attention has been given to explainability, fairness evaluation, transparency, policy stability, and ethical pricing behavior. As a result, important questions remain regarding whether RL-based pricing systems can be deployed responsibly in operational environments that directly affect users, infrastructure utilization, and public trust.

Therefore, integrating trustworthy AI principles into RL-based pricing systems represents an increasingly important research direction, particularly in adaptive urban and resource-sensitive operational contexts.

Table 2 summarizes the recent studies on reinforcement learning applications in dynamic pricing and related adaptive allocation systems.

Table 2. Summary of recent studies on reinforcement learning applications.

Authors	Domain	Method Used	Results
Watari et al. (2023) [9]	Power grid optimization	DRL	Grid load stability improved by 57%
Yavuz & Kaya (2024) [8]	Perishable good pricing	DQL, SAC	Pricing accuracy improved to 95–96%, decision time reduced by 70–80%.
Shen et al. (2024) [10]	Bond market liquidity	DQN	Enhanced municipal bond market liquidity
Bae et al. (2024) [11]	EV charging pricing	QL	Increased revenue through personalized pricing
Zhu et al. (2024) [12]	Airline pricing	PPO, TRPO	Achieved 99% proximity to theoretical maximum revenue

Table 2. *Cont.*

Authors	Domain	Method Used	Results
Bao et al. (2026) [13]	Ride-hailing and ride-pooling	DRL	Reduced pooling detour delay by 36.1% and total passenger waiting time by 20.1%
Wang et al. (2025) [7]	Hotel dynamic pricing	Two-stage DRL with fairness constraints	Optimized pricing while satisfying group and temporal fairness constraints
Zibaie et al. (2025) [14]	Multi-modal delivery pricing	DRL	Dynamic pricing used to influence customer choices and support transportation-resource allocation

2.3. Trustworthiness Challenges and Explainability in Reinforcement Learning

Despite the strong optimization capabilities of reinforcement learning (RL) algorithms such as DQN, PPO, and TRPO, their practical deployment in adaptive operational environments remains associated with several trustworthiness challenges related to convergence stability, scalability, explainability, and fairness. Unlike traditional predictive models, RL agents continuously interact with dynamic environments and optimize sequential decisions through long-term reward maximization. Although this adaptive behavior enables efficient optimization under changing conditions, it may also produce unstable or difficult-to-interpret decision patterns.

One major challenge involves convergence stability. Value-based algorithms such as Q-learning and DQN may experience unstable convergence behavior when combined with nonlinear function approximation methods, particularly deep neural networks [18]. In contrast, policy-gradient approaches such as PPO and TRPO generally demonstrate more stable learning dynamics, while model-based methods may provide clearer convergence guarantees. Furthermore, DQN-based systems are affected by correlated training sequences and unstable Q-value updates, which can reduce learning reliability and decision consistency. To mitigate these issues, DQN commonly incorporates experience replay mechanisms and separate target Q-networks that stabilize the learning process.

Additional challenges emerge in stochastic and high-dimensional operational environments. Under uncertain demand conditions, inaccurate Q-value estimation may reduce pricing quality and overall decision performance. Previous studies demonstrated that improved DQL variations, including pDQL1 and pDQL2, achieved greater robustness in stochastic environments through multi-demand modeling and average reward mechanisms [12]. Scalability also remains an important limitation, as deep RL models may become computationally expensive in large action and state spaces. Research in airline pricing optimization showed that DQN encountered memory utilization challenges, while PPO and TRPO demonstrated more computationally efficient performance [12]. Learning efficiency further affects practical applicability because convergence speed directly influences operational usability and adaptation capability in real-time environments [12,13].

Beyond stability-related limitations, reinforcement learning systems are frequently criticized for their “black-box” nature. In many decision-support settings, stakeholders cannot easily determine why specific actions are selected or which variables drive the learned policy. Recent research therefore increasingly considers explainability not only as a post hoc diagnostic tool but also as an architectural component of trustworthy AI-driven control systems. Explainable-by-design DRL approaches combining interpretable reasoning mechanisms with learned policies illustrate how transparency, accountability, and human understanding can be improved without abandoning adaptive control capabilities [5].

Complementary approaches include state–action explanations, policy summarization, trajectory auditing, and SHAP-based interpretation.

Fairness-related concerns are also becoming increasingly important in RL-based pricing systems. Existing work on personalized electric vehicle charging has highlighted the potential for pricing mechanisms to disadvantage particular users when individual information is incorporated into price formation [11]. More recent fairness-constrained DRL pricing research explicitly limits price differences across customer groups and across time, demonstrating that fairness can be incorporated directly into the optimization problem rather than assessed only after deployment [7]. These developments reinforce the need for RL pricing frameworks that evaluate profitability together with fairness, stability, transparency, and accountability.

Overall, the existing literature suggests that ensuring trustworthiness in reinforcement learning systems requires not only strong optimization performance but also methodological solutions that improve stability, explainability, fairness, robustness, and decision consistency under uncertain operational conditions. These challenges highlight the need for trustworthy RL frameworks capable of balancing adaptive optimization with transparent, accountable, and ethically aligned decision-making.

2.4. Research Gap in Relation to RL Applications

Although previous studies demonstrate the strong optimization capabilities of reinforcement learning in dynamic pricing systems, existing research primarily focuses on profitability and revenue maximization. Comparatively limited attention has been given to how RL-based pricing systems can remain interpretable, stable, fair, and trustworthy under dynamic operational conditions.

Furthermore, many existing studies evaluate RL performance mainly through financial metrics while overlooking broader trustworthiness-related considerations, including decision consistency, explainability, robustness, and transparency. Recent studies have begun to address individual trustworthiness dimensions, including fairness-constrained RL pricing [7] and explainability-oriented DRL control architectures [5]. However, integrated evaluation frameworks that simultaneously examine profitability, fairness-oriented pricing behaviour, explainability, policy stability, and trajectory-level decision auditing remain comparatively limited.

Therefore, this study addresses this gap by proposing a trustworthy reinforcement learning framework for dynamic pricing and adaptive operational decision-making that combines these dimensions within a unified evaluation framework. The proposed approach evaluates RL agents not only in terms of profitability but also through explainability, behavioral stability, fairness-oriented evaluation, and transparent decision-making behavior.

At the broader smart-city level, governance-aware multi-objective reinforcement learning combined with digital-twin policy validation has recently been proposed as a promising direction for aligning AI-based urban decision-making with sustainability, equity, safety, and public accountability objectives [19]. This further reinforces the need for RL frameworks that evaluate optimization performance together with explainability, fairness, stability, and auditability.

3. Methodology

3.1. Data Preparation and Environment Design

The effectiveness of reinforcement learning systems strongly depends on the quality, consistency, and realism of the underlying data. Historical e-commerce transactional data containing product prices, costs, quantities sold, product categories, and temporal purchase information were extracted from multiple related sources and preprocessed to

ensure consistency before model training. The dataset comprised 14,450 weekly aggregated transactional records spanning the period from 2016 to 2023 across multiple product categories. The data were sorted chronologically and aggregated into weekly records. Inconsistent records were removed, while negative or invalid product cost values were replaced with 90% of the corresponding product's average price to avoid unrealistic reward calculations and distorted optimization behaviour. Categorical variables were standardized before encoding, and weekly aggregation was applied to reduce short-term fluctuations while preserving temporal demand patterns for reinforcement learning.

Product categories were normalized into broader operational groups to reduce classification sparsity and improve training stability. To preserve temporal dynamics while reducing short-term noise, the data were aggregated into weekly intervals grouped by product category. Because historical observations were limited for long-horizon sequential learning, additional temporal sequences were generated using a seasonal naïve forecasting approach that preserved recurring seasonal patterns and operational variability. Although the dataset originated from e-commerce transactions, it was used as a methodological proxy for adaptive operational environments because it captures temporal demand fluctuations, resource constraints, and sequential pricing decisions that are also characteristic of many smart city applications. Consequently, the proposed framework should be interpreted as a transferable methodological approach rather than a domain-specific smart city implementation. Validation using real urban operational datasets remains an important direction for future research.

The base-demand variable was obtained from the preprocessed weekly category-level dataset and represents the historical reference demand for each product category and time period. It was derived from the observed transactional demand and sales patterns retained during weekly data aggregation.

Elasticity coefficients were assigned according to category-specific purchasing behavior and domain-informed operational assumptions. Essential product categories were associated with lower elasticity sensitivity, whereas non-essential or luxury-oriented categories were assigned higher elasticity values. Although simplified and not empirically calibrated using historical transaction data, these category-specific elasticity coefficients, derived from practical domain expertise, enabled the environment to simulate adaptive demand-response behaviour while providing a realistic approximation of customer price sensitivity.

The elasticity coefficients should therefore be interpreted as category-informed simulation parameters rather than econometrically estimated demand elasticities. This design choice allowed the environment to represent heterogeneous price sensitivity across product groups while keeping the simulation tractable. However, future work should estimate elasticity values directly from historical transaction data and examine the sensitivity of learned policies to alternative elasticity assumptions.

Consumer demand was modeled by adapting a normalized constant-elasticity price-response function, consistent with smart-pricing and demand-responsive decision frameworks in which pricing decisions are adapted to demand, infrastructure, and operational conditions [20]:

$$D_t = D_0 \cdot \left(\frac{P_t}{P_0} \right)^{-E_c \cdot \alpha_t}, \quad \alpha_t \sim U(0.95, 1.05), \quad (1)$$

where D_0 denotes the historical reference demand, P_t is the adjusted price, P_0 is the reference price, E_c is the category-specific price-elasticity coefficient for category c , and α_t is a stochastic elasticity-response strength multiplier. The stochastic multiplier α_t , sampled from a uniform distribution between 0.95 and 1.05 at each decision step, introduces small random variation in the elasticity response and the resulting demand.

The reinforcement learning environment was developed using the OpenAI Gymnasium framework and operated on an episodic basis, where one episode represented 52 weekly operational periods. At each time step, the agent observed operational state variables including current product price, estimated demand, inventory level, elasticity-related indicators, historical sales information, and category-level variables.

Categorical variables were encoded using One-Hot Encoding. The action space consisted of nine discrete pricing actions, ranging from -20% to $+20\%$ in 5% increments, allowing the agent to make gradual pricing adjustments while reducing excessive operational volatility and unstable pricing behavior.

3.2. Multi-Objective Reward Function

The proposed framework was formulated as a Markov Decision Process (MDP), represented by the tuple (S, A, P, R, γ) , where S denotes the state space, A represents the action space, P defines environment transitions, R is the reward function, and γ denotes the discount factor (for main notations used in RL environment, see Table 3 below).

Table 3. Main notation used in the reinforcement learning environment.

Symbol	Meaning
D_t	Adjusted demand at time step t
D_0	Historical reference demand
P_t	Adjusted price at time step t
P_0	Reference price
E_c	Category-specific price elasticity coefficient
α_t	Stochastic elasticity-response strength multiplier
S	State space
A	Action space
P	Transition function
R	Reward function
π	Pricing policy
γ	Discount factor
R_t	Total reward at time step t
Π_t	Operational profit at time step t
Q_t	Quantity sold at time step t
P_t	Price at time step t
C_t	Cost at time step t
λ_I	Inventory accumulation penalty
I_t	Inventory level at time step t
I_t^{left}	Remaining inventory at time step t
β_v	Sales-volume reward coefficient

Following recent RL formulations that evaluate policies through expected cumulative reward [20,21], the optimization objective is defined as:

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^T \gamma^t R_t \right], \quad (2)$$

where π denotes the pricing policy, T is the episode length (horizon), γ is the discount factor, and R_t is the reward obtained at time step t . For the simple reward case, the reinforcement learning problem could be considered with $\gamma = 1$.

To support trustworthy and operationally stable decision-making, the reward structure was designed as a multi-objective function that combines profitability, inventory efficiency, pricing stability, and fairness-oriented behavioral constraints:

$$R_t = \Pi_t - \lambda_I \cdot I_t^{\text{left}} + \beta_v \cdot Q_t + B_{\text{stock}} + B_{\text{conversion}} - P_{\text{lowdemand}} - P_{\text{highprice}} + B_{\text{discount}} + A_{\text{fatigue}} + B_{\text{clearance}}, \quad (3)$$

where $\Pi_t = (P_t - C_t)Q_t$ represents operational profit, P_t is the price, C_t is the cost, and Q_t is the quantity sold. $I_t^{\text{left}} = I_t - Q_t$ represents remaining inventory, and λ_I defines the inventory accumulation penalty. The term $\beta_v Q_t$ represents an additional sales-volume reward, where β_v is the sales-volume reward coefficient, which controls the strength of the incentive assigned to quantity sold. B_{stock} rewards inventory efficiency, and $B_{\text{conversion}}$ rewards stock balancing behavior or high-conversion behavior. $P_{\text{lowdemand}}$ penalizes low-demand or demand-drop conditions, whereas $P_{\text{highprice}}$ penalizes undesirable pricing behavior under unfavorable demand conditions. In addition, B_{discount} rewards proactive discounting strategies, while A_{fatigue} penalizes repetitive and unjustified price increases and $B_{\text{clearance}}$ rewards aggressive inventory clearance.

The reward function was designed to represent a constrained operational optimization problem rather than a pure revenue-maximization task. Profit was retained as the primary component, while additional reward and penalty terms were introduced to discourage inventory accumulation, excessive price increases under weak demand, repeated unjustified price increases, and economically inconsistent elasticity responses. This structure encouraged the agent to balance profitability with inventory turnover, demand responsiveness, pricing stability, and fairness-oriented behavior.

The proposed reward structure encouraged balanced adaptation while discouraging exploitative, unstable, or economically irrational pricing behavior. Unlike conventional profit-oriented reward formulations, the proposed approach incorporated trustworthiness-oriented reward components intended to improve policy stability, behavioral consistency, and transparency in dynamic decision-making settings.

The reward-function coefficients, penalties, bonuses, and behavioural thresholds were defined heuristically using domain knowledge and iterative experimentation during the development of the simulation environment. Relative thresholds were introduced to identify operationally undesirable conditions, including demand falling below 50% or 80% of baseline demand, remaining inventory exceeding 30% of available stock, high conversion rates above 70%, and repeated price increases across consecutive decision periods. These parameters were selected to encourage economically reasonable pricing behaviour while discouraging excessive inventory accumulation, low-demand price increases, and unstable pricing actions. The reward parameters were not empirically optimized through automated calibration and should therefore be interpreted as domain-informed design choices.

Because the reward terms jointly contribute to the cumulative reward, they operate as complementary incentives and penalties that encourage balanced pricing behaviour rather than optimizing a single objective. Their relative influence was determined heuristically during environment development to promote stable learning, economically reasonable pricing decisions, and fairness-oriented behaviour.

3.3. Reinforcement Learning Models and Evaluation Strategy

Three reinforcement learning architectures were evaluated in this study: Deep Q-Network (DQN), Proximal Policy Optimization (PPO), and Advantage Actor–Critic (A2C). All models were trained under identical experimental conditions using 350,000 training steps, a learning rate of 0.001, a discount factor of $\gamma = 0.8$, and an MlpPolicy (multilayer perceptron architecture). The multilayer perceptron policy was selected because the environment primarily consisted of structured numerical and categorical operational variables.

Additional hyperparameter optimization was performed for the DQN model using a structured manual parameter search consisting of 27 parameter combinations. The final

selected configuration utilized a learning rate of 0.001, a discount factor of $\gamma = 0.8$, an exploration fraction of 0.15, and a final exploration rate of 0.01.

PPO and A2C were trained using the commonly adopted default hyperparameter settings provided by the Stable-Baselines3 implementation, whereas a dedicated manual hyperparameter search was performed only for the DQN model (for a classic DQN see [22]) because it was selected as the primary candidate for the proposed trustworthy reinforcement learning framework. Consequently, the comparison among algorithms should be interpreted as indicative rather than as a comparison of individually optimized implementations. Future work will include systematic hyperparameter optimization for all evaluated algorithms to enable a more balanced comparative evaluation.

To improve transparency and evaluate decision-making behavior, the proposed framework incorporated trajectory auditing, SHAP-based explainability analysis, pricing action distribution analysis, inventory-demand balance evaluation, policy interpretation techniques, and policy stability assessment. Additional analysis using Principal Component Analysis (PCA) and K-means clustering was performed to explore behavioral similarities and policy adaptation patterns across trained agents.

Finally, model robustness and generalization performance were evaluated using a 10-fold cross-validation procedure with chronologically ordered data to assess stability under varying temporal operational conditions. The dataset was first sorted chronologically and partitioned into ten consecutive temporal folds of approximately equal size. For each fold, the selected fold was used for testing, while the remaining folds were used for training. The reinforcement learning model was retrained independently for each fold using identical hyperparameter settings, and the average test reward across all folds was used to assess model stability.

4. Results and Discussion

4.1. Reinforcement Learning Model Selection and Hyperparameter Optimization

The experimental evaluation focused on identifying a reinforcement learning (RL) model capable of balancing operational efficiency with trustworthy and explainable decision-making. Three RL architectures—Deep Q-Network (DQN), Proximal Policy Optimization (PPO), and Advantage Actor–Critic (A2C)—were evaluated within a discrete action space consisting of nine pricing actions ranging from -20% to $+20\%$. The action space was intentionally designed to avoid unrealistic price fluctuations while still providing sufficient flexibility for adaptive pricing decisions. The $\pm 20\%$ range represents moderate price adjustments that are more consistent with practical pricing scenarios, whereas larger changes could produce implausible pricing behaviour within the simulation. A 5% increment was selected as a compromise between decision granularity and computational efficiency, allowing the agent to explore meaningful pricing alternatives without unnecessarily increasing the size of the discrete action space. Such a structure reflects realistic operational pricing scenarios in which decisions are selected from a finite set of interventions rather than generated continuously.

The environment relied on structured operational variables, including demand, inventory levels, elasticity indicators, and historical pricing information, making multilayer perceptron architectures computationally efficient and sufficiently interpretable for the given task.

As illustrated in Figure 1, the comparison of rolling average rewards revealed substantial behavioral differences between the evaluated agents. PPO achieved the highest cumulative reward and demonstrated relatively stable convergence behavior. However, additional analysis indicated that the model occasionally increased prices during low-demand periods, suggesting stronger reward-seeking pricing tendencies. In practical urban

systems, such behavior may produce undesirable outcomes in citizen-facing services such as transportation, energy demand-response systems, or adaptive public utilities.

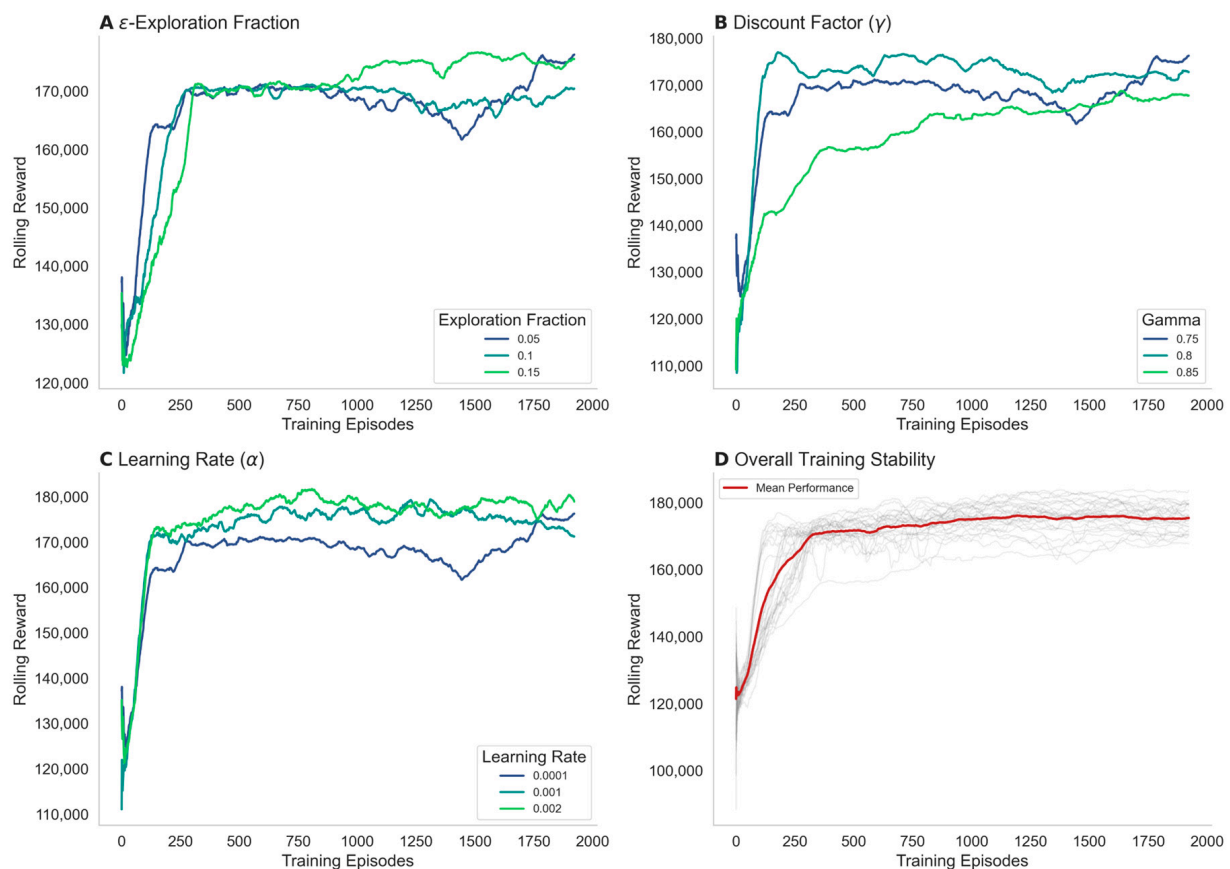


Figure 1. Hyperparameter optimization and training stability results for the DQN model: (A) influence of exploration fraction on rolling reward performance; (B) influence of discount factor γ ; (C) influence of learning rate; and (D) overall training stability across evaluated parameter combinations. In subfigure (D), the grey lines represent the rolling reward trajectories of individual training runs, while the highlighted line represents the mean performance across all runs.

In contrast, the A2C model demonstrated more conservative learning behavior throughout training. The agent frequently selected repetitive pricing actions and showed lower responsiveness to changing environmental conditions. Although such behavior improved predictability, it also reduced adaptability, which remains an important requirement in non-stationary operational environments.

Among the evaluated models, DQN demonstrated the most balanced exploration–exploitation behavior. Although the training dynamics exhibited greater variability compared to PPO, this reflected broader exploration of pricing strategies and active Q-value adaptation rather than premature convergence toward repetitive actions. Such adaptive learning behavior is particularly relevant in sustainable urban operations, where systems must continuously respond to fluctuations in demand, capacity, and resource availability.

To further improve model reliability, hyperparameter optimization was conducted using a structured grid search consisting of 27 parameter combinations. Different exploration fractions, discount factors, and learning rates were evaluated under identical experimental conditions. As presented in Figure 1A–D, the optimal configuration was achieved using a learning rate of 0.001, a discount factor of $\gamma = 0.8$, an exploration fraction of 0.15, and a final exploration rate of 0.01. This configuration provided the most balanced trade-off between short-term responsiveness and long-term operational stability.

The selected discount factor was particularly important because it enabled the agent to prioritize immediate demand-response signals while still accounting for future inventory and resource constraints. In adaptive urban operations, such balancing behavior is essential for reducing inefficient resource accumulation, unstable pricing policies, and excessive demand concentration.

The final model selection was therefore based not only on cumulative reward performance but also on broader trustworthiness-oriented criteria, including behavioral consistency, policy adaptability, explainability, and operational alignment. This multi-criteria evaluation approach is particularly important in sustainable AI deployment, where maximizing reward alone does not necessarily guarantee socially or operationally responsible behavior.

4.2. Performance Evaluation and Category-Specific Pricing Dynamics

The trained DQN agent demonstrated substantial operational improvements compared to the baseline rule-based pricing strategy. The baseline consisted of a manually designed rule-based pricing policy evaluated under the same experimental conditions as the reinforcement learning agent. Both approaches were assessed over 30 annual evaluation episodes, each comprising 52 weekly decision periods. Across these repeated evaluation runs, the RL-based approach increased cumulative profit by 12.58%, improving total profit from €4.18 million to €4.71 million. The cumulative total-profit comparison is reported here as the aggregate performance indicator, whereas the episode-level statistical comparison with the rule-based baseline is presented in Section 4.5.

As illustrated in Figure 2A, the category-level analysis revealed that the agent learned differentiated pricing strategies depending on demand elasticity, sales behavior, and inventory conditions. In highly elastic categories such as electronics, fashion, and leisure products, the agent actively applied dynamic pricing adjustments. Price reductions frequently coincide with increased sales volume, suggesting that the model successfully adapted to elasticity-related demand responses and optimized inventory turnover behavior.

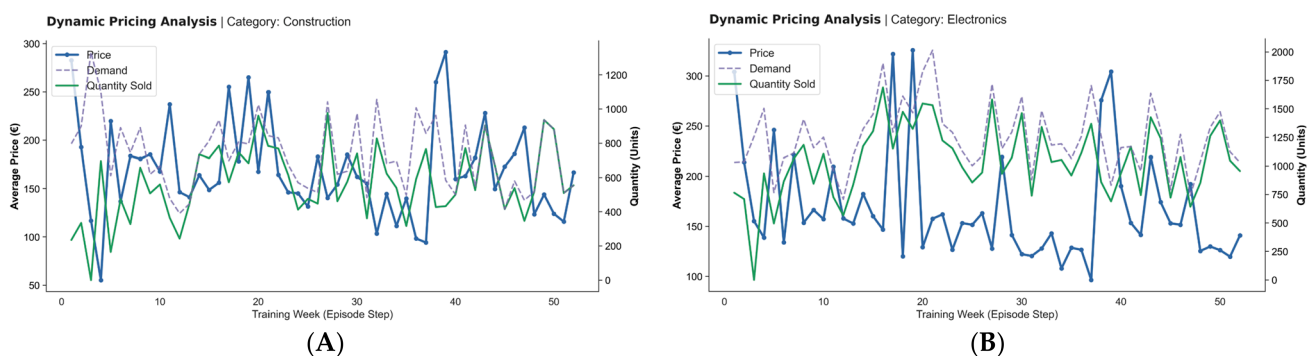


Figure 2. Dynamic pricing behavior learned by the DQN agent across different product categories: (A) construction category; (B) electronics category. The plots illustrate interactions between pricing actions, estimated demand, and quantity sold across weekly operational periods.

In categories associated with household and construction products (Figure 2B), pricing behavior was more volatile and less predictable. Although frequent pricing fluctuations were observed, these adjustments were not always strongly associated with improved sales performance. This suggests that demand dynamics in such categories may be influenced by additional factors not fully represented within the environment, including seasonal effects, delayed purchasing behavior, or supply-side variability. Nevertheless, the agent continued to demonstrate adaptive behavior rather than relying on static pricing policies.

The analysis additionally showed that the agent generally adjusted prices gradually rather than relying on abrupt pricing changes. Most actions remained within the

$\pm 5\text{--}10\%$ range, indicating that the learned policy prioritized stable long-term optimization instead of aggressive short-term profit extraction. Such gradual adaptation may be particularly beneficial in operational systems where excessive pricing volatility could reduce predictability and destabilize demand-response behavior.

The resource-oriented behavior of the agent was also evaluated by comparing demand levels against available inventory. Demand exceeded inventory capacity in only approximately 15% of episodes, suggesting that the agent maintained a relatively balanced relationship between demand stimulation and available resources. This finding highlights the potential applicability of RL-based adaptive pricing strategies in resource-constrained operational environments.

Overall, the category-level evaluation demonstrated that the proposed RL framework successfully adapted pricing behavior according to contextual operational conditions. Rather than applying a single optimization strategy across all categories, the agent developed differentiated pricing responses under varying demand and inventory scenarios. These findings support the broader applicability of reinforcement learning for adaptive operational decision-support systems operating under dynamic and resource-sensitive conditions.

4.3. Behavioral Integrity, Fairness, and Policy Stability

To move beyond purely reward-based evaluation, the trained DQN agent was additionally assessed using fairness-, explainability-, and stability-oriented indicators. Such evaluation is particularly important in AI-driven operational systems because maximizing cumulative reward alone does not necessarily guarantee trustworthy or socially acceptable behavior.

The behavioral analysis showed that the agent did not generate undesirable price increases during low-demand conditions, resulting in a 0% rate for this fairness-related violation indicator. This suggests that the reward structure and environmental constraints successfully discouraged exploitative pricing behavior under unfavorable demand conditions. Similar mechanisms may be particularly relevant in adaptive operational systems where excessive pricing responses could negatively affect fairness, predictability, or public trust.

Action distribution analysis further revealed that the agent remained adaptive while maintaining policy consistency. The frequency of pricing direction changes reached 51.84%, indicating active responsiveness to environmental signals, whereas repeated actions accounted for 23.64% of decisions, suggesting relatively stable behavior under recurring operational conditions. Larger pricing adjustments exceeding 10% represented 28.83% of all actions and typically occurred under stronger elasticity signals or changing inventory constraints. Nevertheless, most pricing interventions remained moderate, indicating that the learned policy generally avoided excessive operational volatility.

To further examine policy structure and behavioral differentiation, Principal Component Analysis (PCA) and K-means clustering were applied to the agent's Q-values. The resulting three-dimensional representation revealed three distinct behavioral clusters associated with different operational regimes. The first cluster appeared to correspond to discount-oriented behavior linked to inventory clearance and high-turnover products, whereas the second reflected more moderate pricing adaptation strategies. The third cluster suggested stronger pricing adjustments associated with higher elasticity tolerance and premium-oriented operational conditions.

The emergence of these distinct clusters suggests that the RL agent learned structured and context-dependent pricing behavior rather than relying on uniform optimization

patterns. Such differentiation may be particularly valuable in adaptive resource allocation environments where different demand profiles require distinct operational responses.

Model stability was additionally evaluated using 10-fold cross-validation with chronologically ordered folds. Across the validation folds, the average test reward reached approximately EUR 160,000 $\pm$ EUR 23,000 (see Table 4), indicating relatively stable performance under varying temporal conditions. Most validation folds achieved average rewards within the approximate range of EUR 150,000–180,000, while lower performance was observed mainly in the first and ninth folds. These lower values may reflect seasonal demand fluctuations or uneven inventory distribution during those periods. Despite these variations, no systematic performance degradation was observed across the remaining folds, suggesting that the proposed reinforcement learning framework generalized beyond a single favourable data partition and maintained stable decision-making under changing temporal conditions.

Table 4. Temporal cross-validation performance pattern across folds.

Fold Group	Observed Performance Pattern	Interpretation
Most folds	Average rewards approximately within EUR 150,000–180,000	Stable performance across most temporal partitions
First fold	Lower than the main stable range	Possible seasonal or inventory-distribution imbalance
Ninth fold	Lower than the main stable range	Possible seasonal or inventory-distribution imbalance
Overall	Approximately EUR 160,000 $\pm$ EUR 23,000	The model maintained broadly stable performance across temporal splits

These findings are particularly relevant because real-world operational environments are inherently non-stationary. Demand patterns, resource availability, and operational constraints continuously evolve. Consequently, AI-driven decision-support systems must demonstrate not only optimization capability but also adaptability, resilience, and behavioral stability under changing conditions.

Future work may extend the fairness evaluation by monitoring additional indicators, including pricing consistency across comparable operational scenarios, the frequency of price increases under vulnerable demand conditions, long-term pricing stability, and customer-level fairness metrics where such data are available.

4.4. Explainability and Transparency Through SHAP and Trajectory Auditing

In alignment with the Special Issue’s emphasis on trustworthy and accountable AI, additional explainability mechanisms were integrated into the evaluation framework. Reinforcement learning systems are frequently criticized for operating as black-box models whose internal decision logic may be difficult to interpret. To improve transparency and interpretability, trajectory auditing and SHAP analysis were applied to provide trajectory-level and global explanations of agent behavior.

As illustrated in Figure 3, trajectory auditing decomposed the reward function into its individual components, allowing the contribution of operational variables to overall reward formation to be evaluated directly. Figure 3 reports the reward-component audit, whereas the global SHAP analysis was used separately to summarize the state variables influencing pricing decisions. The analysis showed that the strongest positive contributors were profit generation, inventory clearance incentives, and sales-volume-related reward components. In contrast, negative reward contributions were primarily associated with

inventory accumulation, low sales efficiency, and undesirable pricing behavior under unfavorable demand conditions.

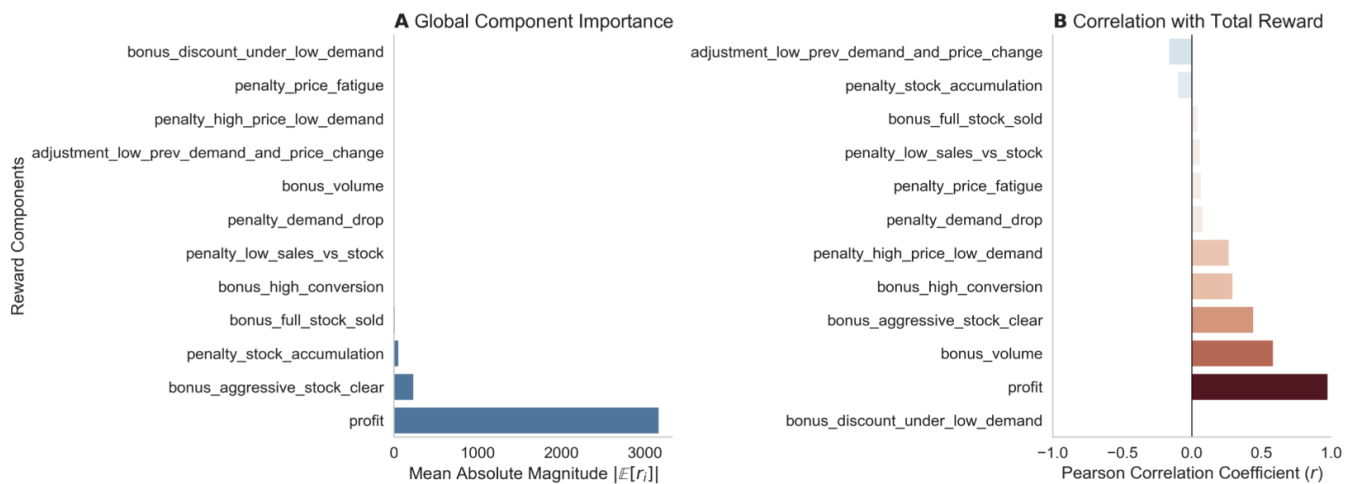


Figure 3. Explainability analysis of the trained DQN agent: (A) global importance of reward-function components; (B) correlation between reward components and total cumulative reward. In subfigure (B), the color gradient indicates the direction and magnitude of the Pearson correlation, ranging from negative (blue) to positive (red), consistent with the correlation coefficient shown on the horizontal axis.

The decomposition results suggest that the learned policy was not driven solely by short-term profit maximization. Instead, the reward structure encouraged more balanced operational behavior by simultaneously considering demand responsiveness, inventory efficiency, and pricing stability. Such multi-objective optimization may be particularly relevant in adaptive operational systems where efficiency objectives must coexist with broader sustainability and fairness considerations.

To further interpret the learned pricing policy, SHAP (SHapley Additive exPlanations) analysis was applied at the global level to identify the state variables that most strongly influenced pricing decisions across the evaluation dataset. The analysis identified stock levels, demand fluctuations, price elasticity, and previous pricing behavior as the most influential factors affecting the learned policy. Additional behavioral analysis demonstrated that price increases were predominantly selected under favourable demand conditions and constrained inventory, whereas lower demand or higher inventory levels generally resulted in more conservative pricing decisions. Most pricing adjustments remained within moderate ranges, indicating stable policy adaptation rather than aggressive short-term optimization. These behavioural patterns were consistent with the global SHAP findings and support the conclusion that the learned policy responded to economically meaningful operational signals rather than exhibiting unstable or random decision behaviour.

The combination of trajectory auditing and SHAP analysis substantially improved the transparency of the proposed framework. From the perspective of adaptive operational governance, such explainability mechanisms represent an increasingly important requirement because automated decisions may directly affect users, infrastructure utilization, and resource allocation. Consequently, explainability tools may support accountability, facilitate regulatory oversight, and improve trust in AI-assisted operational systems. While the present study focuses on global explainability, future work will extend the framework with case-specific local SHAP to further improve the auditability of individual pricing decisions.

4.5. Comparison with Baseline Pricing Strategies

To evaluate whether the proposed RL-based framework generated measurable operational value, the trained DQN agent was compared with four baseline pricing strategies: static pricing, random pricing, heuristic pricing, and a rule-based if–else strategy. Each strategy was evaluated across 30 annual episodes, with each episode consisting of 52 weekly decision periods.

The static baseline applied a constant price equal to the mean product price throughout each evaluation episode. The random baseline independently multiplied the observed average price at each step by a uniformly sampled factor between 0.95 and 1.05, using a fixed random seed for reproducibility. The heuristic strategy began from the initial average price and reduced the current price by 10% when inventory exceeded 110% of baseline demand. It increased the price by 5% when inventory fell below 70% of baseline demand and enforced a minimum selling price equal to 105% of product cost. The rule-based strategy applied a 5% price reduction when stock exceeded 120% of baseline demand, a 5% price increase when baseline demand exceeded 120% of available stock, and otherwise retained the observed average price. For all baseline strategies, sales were calculated as the minimum of baseline demand and available stock, and profit was calculated as the difference between selling price and product cost multiplied by the quantity sold.

The comparison demonstrated that the RL agent achieved the strongest cumulative performance and consistently outperformed the rule-based strategy, which represented the most competitive baseline. The episode-level statistical comparison showed that the RL-based approach generated an average profit of approximately EUR 174,000 per episode, whereas the rule-based strategy achieved approximately EUR 137,000 per episode. Overall, the RL agent exceeded the rule-based strategy by approximately EUR 37,000 per episode on average.

Bootstrap analysis indicated lower dispersion for the RL agent and substantially wider variability for the rule-based strategy, for which the interval was approximately EUR 111,000–162,000. A one-sided paired *t*-test confirmed that the RL agent significantly outperformed the rule-based strategy, with $t = 2.80$ and $p = 0.004$. The effect size, measured using Cohen's *d*, was approximately 0.5, indicating a medium practical effect. The Friedman test also confirmed statistically significant differences among pricing strategies, $\chi^2 \approx 46.7$, $p < 0.001$, with post hoc analysis identifying the RL agent as the main source of performance improvement.

These findings indicate that the RL framework did not merely outperform simplistic pricing baselines but also generated measurable improvement relative to a realistic manually designed rule-based strategy. From an operational perspective, this strengthens the argument that reinforcement learning can support adaptive decision-making under changing demand and resource conditions while maintaining stable and context-sensitive pricing behavior.

4.6. Implications for AI-Driven Urban Decision-Making

The experimental results suggest that the proposed reinforcement learning framework may provide a transferable methodological basis for AI-driven urban operations. Although the experimental environment relied on transactional pricing data, the underlying sequential decision-making structure shares similarities with broader smart city scenarios characterized by dynamic demand, resource limitations, and operational uncertainty.

The framework's ability to optimize operational performance while incorporating trustworthiness-oriented constraints is particularly relevant for adaptive and resource-sensitive urban systems. The observed 12.58% improvement in operational efficiency, combined with explainability mechanisms and fairness-oriented evaluation, suggests that

reinforcement learning may support adaptive decision-making while improving transparency and behavioral stability.

Related AI-based resource-allocation research in circular-economy digital marketplaces further demonstrates the potential of data-driven methods for demand–supply matching, dynamic pricing, and resource utilization in sustainability-oriented systems [23]. In energy demand-response systems, reinforcement learning could support adaptive pricing strategies that balance grid load while reducing excessive consumption peaks. In smart mobility systems, comparable methods may assist with congestion-sensitive pricing, ride-sharing coordination, or dynamic resource allocation. Similarly, in sustainable logistics and urban inventory management, RL-based systems may improve resource utilization while reducing waste and operational inefficiencies.

Despite these promising findings, several limitations should be acknowledged. The elasticity assumptions used within the environment were simplified and based on practical approximations rather than large-scale econometric estimation. A formal sensitivity analysis of the elasticity coefficients and demand-response parameters was not conducted; therefore, their influence on pricing actions, profitability, and fairness-related outcomes was not quantified systematically. Although temporal cross-validation and hyperparameter evaluation were used to assess model robustness, these analyses do not replace a dedicated parameter sensitivity study. In addition, the experiments relied on transactional e-commerce data rather than directly collected urban operational datasets. Consequently, the present study should be interpreted primarily as a transferable methodological framework rather than a fully deployed smart city implementation. Future research should systematically vary the elasticity and demand parameters to assess the robustness of the learned pricing policy under alternative demand assumptions.

Nevertheless, the results support the broader conclusion that reinforcement learning may contribute to more adaptive and resilient operational systems when optimization objectives are combined with trustworthy AI principles. By integrating adaptive optimization, explainability mechanisms, fairness-oriented evaluation, and stability analysis, the proposed framework provides a practical foundation for future research on accountable AI-driven decision-support systems in smart city environments.

5. Conclusions

This study proposed a trustworthy reinforcement learning framework for AI-driven operational decision-making using sustainable dynamic pricing and resource optimization as representative adaptive operational scenarios. The research addressed an important gap in the existing reinforcement learning literature by integrating operational optimization with trustworthiness-oriented principles, including explainability, fairness, stability, and transparency.

A custom reinforcement learning environment was developed to simulate interactions among demand dynamics, inventory availability, elasticity, and pricing behavior. Three reinforcement learning architectures—DQN, PPO, and A2C—were evaluated under identical experimental conditions. Among the evaluated approaches, DQN demonstrated the most balanced performance in terms of adaptability, operational consistency, and stable decision-making behavior.

The optimized DQN framework improved cumulative operational performance by 12.58% relative to the baseline rule-based strategy while maintaining context-dependent and behaviorally stable pricing adaptation across varying operational conditions. The results additionally demonstrated that trustworthiness-oriented evaluation mechanisms can be effectively integrated into reinforcement learning optimization frameworks through fairness-oriented analysis, stability assessment, trajectory auditing, and explainability methods.

To improve transparency and interpretability, SHAP-based analysis and reward decomposition techniques were incorporated into the evaluation framework. These methods indicated that the agent's decisions were primarily influenced by economically relevant operational variables, including demand fluctuations, inventory conditions, elasticity indicators, and sales-related signals. The integration of explainability mechanisms substantially improved the interpretability of the proposed framework compared to conventional black-box reinforcement learning approaches.

The findings suggest that reinforcement learning may support more adaptive and resilient operational systems when optimization objectives are combined with trustworthy AI principles. Importantly, the proposed framework directly addresses the research gaps identified in the literature by extending conventional reinforcement learning pricing approaches beyond short-term profit maximization. In particular, the integration of explainability, fairness, stability, and transparency demonstrates that trustworthy AI principles can be incorporated into reinforcement learning-based pricing without compromising operational performance. Although the current framework was evaluated using transactional pricing data, the underlying methodological structure may be transferable to broader smart city and resource-sensitive operational environments characterized by dynamic demand and operational uncertainty.

Several limitations should nevertheless be acknowledged. The elasticity assumptions used within the environment were simplified and based on practical approximations rather than large-scale econometric estimation. In addition, the experiments relied primarily on transactional e-commerce data rather than directly collected urban operational datasets. Furthermore, the proposed framework was evaluated in a single-agent reinforcement learning setting, where pricing decisions were made by a single autonomous agent. Future research may therefore focus on integrating real-world smart city datasets, multi-agent reinforcement learning environments, digital twin simulations, more advanced fairness-aware reward structures, and human-in-the-loop decision-making to investigate how interactions among multiple agents and human operators influence trustworthiness, coordination, and accountability in real smart city systems.

Overall, the proposed framework demonstrates that reinforcement learning can move beyond purely profit-oriented optimization toward more transparent, explainable, and responsible AI-driven decision-support systems. Compared with the conventional rule-based pricing strategy, the proposed framework improved cumulative operational performance by 12.58% while incorporating fairness, explainability, stability, and transparency mechanisms. Although the present study did not directly quantify long-term societal outcomes or public trust, these characteristics provide an important foundation for the responsible adoption of AI-driven operational systems in smart city environments. The findings suggest that operational performance and trustworthy AI objectives should be viewed as complementary rather than competing goals.

Author Contributions: Conceptualization, R.A.; methodology, Ž.B. and R.A.; software, Ž.B.; validation, Ž.B. and R.A.; formal analysis, Ž.B. and R.A.; investigation, Ž.B. and R.A.; resources, Ž.B. and R.A.; data curation, Ž.B.; writing—original draft preparation, Ž.B. and R.A.; writing—review and editing, Ž.B. and R.A.; visualization, Ž.B.; supervision, R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

A2C	Advantage Actor–Critic
AI	Artificial Intelligence
DQN	Deep Q-Network
IoT	Internet of Things
MDP	Markov Decision Process
PCA	Principal Component Analysis
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SHAP	SHapley Additive exPlanations

References

1. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson Education: London, UK, 2016.
2. Haenlein, M.; Kaplan, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *Calif. Manag. Rev.* **2019**, *61*, 5–14. [\[CrossRef\]](#)
3. Duan, Y.; Edwards, J.S.; Dwivedi, Y.K. Artificial intelligence for decision making in the era of Big Data—Evolution, challenges and research agenda. *Int. J. Inf. Manag.* **2019**, *48*, 63–71. [\[CrossRef\]](#)
4. Mora-Cantalops, M.; García-Barriocanal, E.; Sicilia, M.-Á. Trustworthy AI Guidelines in Biomedical Decision-Making Applications: A Scoping Review. *Big Data Cogn. Comput.* **2024**, *8*, 73. [\[CrossRef\]](#)
5. Jaziri, W.; Sassi, N. Explainable by Design: Enhancing Trustworthiness in AI-Driven Control Systems. *Mathematics* **2025**, *13*, 3805. [\[CrossRef\]](#)
6. Shankar, V. How AI is reshaping marketing. *J. Acad. Mark. Sci.* **2020**, *48*, 7–20.
7. Wang, X.; Xie, Y.; Jian, L.; Liu, W.; Lv, W. A Two-Stage Deep Reinforcement Learning-Driven Dynamic Discriminatory Pricing Model for Hotel Rooms with Fairness Constraints. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 337. [\[CrossRef\]](#)
8. Yavuz, T.; Kaya, O. Deep reinforcement learning algorithms for dynamic pricing and inventory management of perishable products. *Appl. Soft Comput.* **2024**, *163*, 111864. [\[CrossRef\]](#)
9. Watari, D.; Taniguchi, I.; Onoye, T. Duck curve aware dynamic pricing and battery scheduling strategy using reinforcement learning. *IEEE Trans. Smart Grid* **2023**, *15*, 457–471. [\[CrossRef\]](#)
10. Yu, K.; Shen, Q.; Lou, Q.; Zhang, Y.; Ni, X. A deep reinforcement learning approach to enhancing liquidity in the U.S. municipal bond market: An intelligent agent-based trading system. *Int. J. Eng. Manag. Res.* **2024**, *14*, 113–126. [\[CrossRef\]](#)
11. Bae, S.; Kulcsar, B.; Gros, S. Personalized dynamic pricing policy for electric vehicles: Reinforcement learning approach. *Transp. Res. Part C Emerg. Technol.* **2024**, *161*, 104540. [\[CrossRef\]](#)
12. Zhu, X.; Jian, L.; Chen, X.; Zhao, Q. Deep Q-learning for multi-flight dynamic pricing: Maximizing revenue with a novel utility function in airline revenue management. *Comput. Ind. Eng.* **2024**, *193*, 110302. [\[CrossRef\]](#)
13. Bao, Y.; Gao, J.; He, J.; Oliehoek, F.A.; Cats, O. A timely match for ride-hailing and ride-pooling services using a deep reinforcement learning approach. *Transp. Res. Part C Emerg. Technol.* **2026**, *187*, 105644. [\[CrossRef\]](#)
14. Zibaie, A.; Beliaev, M.; Alizadeh, M.; Pedarsani, R. Dynamic Pricing for Multi-Modal Meal Delivery Using Deep Reinforcement Learning. *Future Transp.* **2025**, *5*, 112. [\[CrossRef\]](#)
15. Ismail, A.; Baysal, M. Dynamic Pricing Based on Demand Response Using Actor–Critic Agent Reinforcement Learning. *Energies* **2023**, *16*, 5469. [\[CrossRef\]](#)
16. Salazar, E.J.; Jurado, M.; Samper, M.E. Reinforcement Learning-Based Pricing and Incentive Strategy for Demand Response in Smart Grids. *Energies* **2023**, *16*, 1466. [\[CrossRef\]](#)
17. Alquthami, T.; Milyani, A.H.; Awais, M.; Rasheed, M.B. An Incentive Based Dynamic Pricing in Smart Grid: A Customer’s Perspective. *Sustainability* **2021**, *13*, 6066. [\[CrossRef\]](#)
18. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
19. Alharbi, S. MORL-SGF: A Governance-Aware Multi-Objective Reinforcement Learning Framework with Digital Twin Policy Validation for Sustainable Smart Cities. *Systems* **2026**, *14*, 294. [\[CrossRef\]](#)

20. Pergamalis, C.; Tsampasis, E.; Gkonis, P.K.; Elias, C.N. Smart Pricing for Smart Charging: A Deep Reinforcement Learning Framework for Residential EV Infrastructure. *Future Internet* **2026**, *18*, 241. [[CrossRef](#)]
21. Wang, J.; Gao, R.; Zha, H. Reliable Off-Policy Evaluation for Reinforcement Learning. *Oper. Res.* **2024**, *72*, 699–716. [[CrossRef](#)]
22. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)] [[PubMed](#)]
23. Sheikh, A.A.; Simske, S.J.; Chong, E.K.P. Evaluating Artificial Intelligence Models for Resource Allocation in Circular Economy Digital Marketplace. *Sustainability* **2024**, *16*, 10601. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.