DOI: https://doi.org/10.15276/ict.03.2026.24
UDC 004.8:658.7:519.816

Event-driven supply chain architecture based on Q-Learning under uncertainty

Vitaliy M. Kobets[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@ksu.ks.ua. Scopus Author ID: 55965109600
Dmytro M. Klionov[1)]
ORCID: https://orcid.org/0009-0004-8537-3806; dklonov@ksu.ks.ua. Scopus Author ID: 55965109600
[1)] Kherson State University, 27, Universytetska St. Kherson, 73000, Ukraine

ABSTRACT

The paper considers building an Event-Driven Architecture (EDA) for Supply Chain Management (SCM) systems using Q-learning under uncertainty. It is substantiated that in the presence of high demand volatility, delivery delays, and input data distortions, deterministic and classical statistical forecasting models commit significant errors, leading to increased operational costs due to stockouts or excess inventory. The authors emphasize the expediency of transitioning from scheduled cyclical forecast calculations to real-time response to events in the logistics network. The objective of the study is to develop an event-driven decision-making architecture that combines an event bus, a Reinforcement Learning-based controller, and an asymmetric reward function. The main focus is on abandoning the direct recovery of the demand function in favor of assessing the optimality of actions in specific system states and balancing holding costs and shortage risks. The paper systemizes key solutions: the formalization of a Markov Decision Process (MDP) without constructing a mathematical model of the environment (a model-free approach); asynchronous event processing in network nodes; and the use of Q-learning as an interpretable and computationally lightweight method, serving as an alternative to Deep Reinforcement Learning (Deep RL). Practical analysis proves that applying the proposed approach ensures system resilience to noisy data and minimizes time lags during managerial decision-making. The conclusions outline perspectives for further research: extending the architecture to multi-agent systems (Multi-Agent RL) and developing node-interaction protocols for distributed SCM networks.

Keywords: Supply chains; event-driven architecture; Q-learning; decision-making algorithms; demand uncertainty; artificial intelligence in logistics

For citation: **Kobets V. M., Klionov D. M. "Event-driven supply chain architecture based on Q-Learning under uncertainty".** ***Informatics. Culture. Technology*****. 2026; Vol. 3 No. 1 (3): 288–297. DOI: https://doi.org/10.15276/ict.03.2026.24**

Relevance. Managing modern Supply Chains (Supply Chain Management, SCM) in a dynamic market environment expands beyond the framework of classical deterministic planning and poses a complex challenge for adaptive engineering. As modern logistics and production networks integrate hundreds of distributed nodes, even a minor delivery delay or a small deviation in order volume at a single stage can scale into significant financial losses for the entire system.

In practice, this drives two key trends: on the one hand, the focus shifts from cumbersome classical reviews and deterministic models toward the direct integration of artificial intelligence tools into the decision-making loop [1]; on the other hand, there is a transition from hard-to-interpret neural networks and Deep Reinforcement Learning (DeepRL) toward transparent RL algorithms that enable explicit analysis and control of selected strategies [6], [7]. Such a transformation is inevitable, as most traditional inventory management methods rely on the assumption of demand stationarity. In contrast, in real-world market conditions with high noise levels (unpredictable impact factors), the accuracy of purely theoretical forecasting declines sharply. Consequently, rigid supply schedules result in either excess inventory accumulation or stockouts. At the same time, attempts to complicate models using multi-echelon DeepRL methods encounter the "black box" problem and high computational resource requirements [3], [4], [7].

The research goal is to develop and substantiate an Event-Driven Architecture (EDA) for supply chain management based on Q-learning, which maximizes rewards of supply chain participants under high environmental uncertainty.

Modern scientific literature highlights 5 main research directions for improving SCM efficiency.

1. *Clarity of System Requirements***:** System stability and Service Level are defined as early as the non-functional requirements stage [6].

© Kobets V., Klionov D.. 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

2. *Event-Driven Architecture*: Transitioning from heavy batch processing to a software Event Bus – a specialized infrastructure layer that provides asynchronous event transmission and processing (e.g., StockLow or ShipmentArrived) among autonomous system nodes without tight direct coupling [3], [4].

3. *Mathematical Formalization*: Representing the network as a Markov Decision Process (MDP) with discrete state (S) and action (A) spaces, alongside an asymmetric reward function (R) [7].

4. *Transparent Artificial Intelligence*: Utilizing tabular Q-learning models with a clear economic interpretation of Q-values [2], [8].

5. *Monitoring*: Continuous monitoring of Q-value convergence and the capability for real-time controller re-training upon changes in external conditions [1], [10].

Practical analysis shows that basic architectural decisions made during the system design phase exert the greatest influence on reducing operational costs. They define the boundaries and operational rules for distributed nodes, without which further operational tuning would be ineffective.

Traditional deterministic mathematical models of multi-echelon supply chains formalize the network as a directed graph $G = (V, E)$, where $V = M \cup K \cup P$ (manufacturers $m \in M$, transporters $k \in K$, and retail sales points $p \in P$).

Under centralized planning, the objective of the manufacturer $m \in M$ is to maximize the net present value (NPV) over the planning horizon H:

$$\max_{q_{mt}} \sum_{t=1}^{H} \frac{1}{(1+r)^t} \left[P_{mt}^{M} \sum_{p \in P} x_{mpt} - C_{mt}^{P}(q_{mt}) - C_{mt}^{H}(I_{mt}) \right] \tag{1}$$

where:

- $\max\limits_{q_{mt}}$ – the maximization operator over output volumes q_{mt} (the manufacturer strives to achieve maximum profit).
- $\sum_{t=1}^{H}$ – summation over all time periods t from 1 to H, where H is the total planning time horizon.
- $\frac{1}{(1+r)^t}$ – the discount factor bringing future cash flows to their present value.
- r – the discount rate (cost of capital or rate of return) per period.
- t – the ordinal number of the time period ($t = 1, \dots, H$).
- $P_{mt}^{M} \sum_{p \in P} x_{mpt}$ – total revenue (income) of manufacturer m in period t.
- P_{mt}^{M} – wholesale selling price per unit of product offered by manufacturer m in period t.
- $\sum_{p \in P} x_{mpt}$ – total delivery volume from manufacturer m to all retail sales points $p \in P$ in period t.
- x_{mpt} – volume of goods shipped from manufacturer m to retail sales point p in period t.
- P – the set of all retail sales points ($p \in P$).
- $C_{mt}^{P}(q_{mt})$ – production cost function of manufacturer m in period t as a function of total output q_{mt}.
- q_{mt} – total volume of output (produced goods) by manufacturer m in period t.
- $C_{mt}^{H}(I_{mt})$ – holding and storage costs for finished goods inventory in manufacturer m's warehouse in period t.
- I_{mt} – residual finished goods warehouse inventory of manufacturer m at the end of period t.

For transporters $k \in K$, the objective is to minimize total fuel costs and environmental taxes along all routes (edges of the graph) $(i, j) \in E$:

$$\min_{x} \sum_{t=1}^{H} \sum_{(i,j) \in E} \left(C_{ijkt}^{\text{fuel}} \cdot d_{ij} \cdot \omega(x_{ijkt}) + E_t^{\text{tax}} \cdot \phi_{ij}(x_{ijkt}) \right), \tag{2}$$

where:

- $k \in K$ – a specific transporter from the set of all transport operators K.
- min – minimization of the transporter's k total costs.
- $\sum_{t=1}^{H}$ – summation over all time periods t from 1 to H.
- $\sum_{(i,j)\in E}$ – summation over all routes (edges E of the transportation network).
- $C_{ijkt}^{fuel} \cdot d_{ij} \cdot \omega(x_{ijkt})$ – fuel costs of transporter k.
- C_{ijkt}^{fuel} – fuel cost per distance unit for transporter k on route (i,j) in period t.
- d_{ij} – distance between nodes i and j.
- $\omega(x_{ijkt})$ – fuel consumption coefficient depending on the cargo volume x_{ijkt} transported by transporter k.
- $E_t^{tax} \cdot \phi_{ij}(x_{ijkt})$ – environmental tax in period t.
- E_t^{tax} – emission tax rate in period t.
- $\phi_{ij}(x_{ijkt})$ – volume of harmful emissions generated during the transportation of cargo x_{ijkt}.

Objective function of the retail sales point ($p \in P$):

$$\max_{x_{mpt}} \sum_{t=1}^{H} \left[P_{pt}^{final} \sum_{m\in M} x_{mpt} - C_{pt}^{S} \cdot \max\left(0, D_{pt} - \sum_{m\in M} x_{mpt}\right) - C_{pt}^{H}(I_{pt}) \right], \tag{3}$$

where:

- $\max\limits_{x_{mpt}}$ – the maximization operator for the total profit of retail sales point $p \in P$ over incoming order volumes x_{mpt}.
- $\sum_{t=1}^{H}$ – summation over all time periods t from 1 to H.
- $P_{pt}^{final} \sum_{m\in M} x_{mpt}$ – total revenue from retail sales at point p in period t.
- P_{pt}^{final} – final retail price per product unit at sales point p in period t.
- $\sum_{m\in M} x_{mpt}$ – total product delivery volume received by sales point p from all manufacturers $m \in M$ in period t.
- x_{mpt} – volume of goods shipped from manufacturer m to retail sales point p in period t.
- $C_{pt}^{S} \cdot \max(0, D_{pt} - \sum_{m\in M} x_{mpt})$ — penalty for stockout (unfulfilled consumer demand) at sales point p in period t.
- C_{pt}^{S} – penalty rate (financial loss) per unit of unfulfilled order at point p in period t.
- D_{pt} – actual consumer demand at retail sales point p in period t.
- $\max(0, D_{pt} - \sum_{m\in M} x_{mpt})$ — stockout volume (the non-negative difference between actual demand and received product volume; equals 0 if delivery covers or exceeds demand).
- $C_{pt}^{H}(I_{pt})$ – holding costs for unused inventory stored at retail sales point p in period t.
- I_{pt} – residual warehouse inventory at retail sales point p at the end of period t.

To overcome the limitations of deterministic planning, a Q-learning model (model-free approach) is applied based on Temporal Difference (TD) updating of Q-values via the Bellman equation:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \cdot \left[R_{t+1} + \gamma \max_{a'\in A} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \right], \tag{4}$$

where:

- $Q_{t+1}(s_t, a_t)$ – the updated action-value estimate for executing action a_t in state s_t, stored in the Q-table for step $t+1$.
- $Q_t(s_t, a_t)$ – the current (prior) expected action-value for action a_t in state s_t.
- α – the learning rate ($\alpha \in (0,1]$), which determines the degree to which the old estimate is adjusted by newly acquired information.
- R_{t+1} – the scalar immediate reward received by the agent from the environment immediately after executing action a_t at step t.
- γ – the discount factor ($\gamma \in [0,1)$), defining the relative weight of future rewards compared to immediate outcomes.
- $\max_{a' \in A} Q_t(s_{t+1}, a')$ – the maximum expected utility estimate across all possible candidate actions $a' \in A$ available in the subsequent state s_{t+1}.
- $\left[R_{t+1} + \gamma \max_{a' \in A} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)\right]$ – the Temporal Difference Error (TD-Error), representing the discrepancy between:
  - $R_{t+1} + \gamma \max_{a' \in A} Q_t(s_{t+1}, a')$ – the TD-Target (the observed immediate reward plus the discounted maximum future value estimate);
  - $Q_t(s_t, a_t)$ – the previous forecast (prior expected Q-value).

The immediate reward function of Transporter $k \in K$ in the simulation environment accounts for the monetary equivalent of penalties for delivery deadline violations:

$$R_{kt} = \text{Revenue}_{kt} - \sum_{(i,j) \in E} E_t^{tax} \cdot e_{ij}\left(x_{ijk,t}\right) - \lambda \cdot C_{pt}^{S} \cdot \max\left(0, \tau_{ij} - dd_{pt}\right), \tag{5}$$

where:

- R_{kt} – the net scalar reward returned to the Q-algorithm for Transporter $k \in K$ at step t.
- Revenue_{kt} – total contractual revenue of Transporter k for provided transportation services in period t.
- $\sum_{(i,j) \in E} E_t^{tax} \cdot e_{ij}\left(x_{ijk,t}\right)$ – total deducted environmental taxes across active routes $(i,j) \in E$ when transporting cargo volume $x_{ijk,t}$.
- $\lambda \cdot C_{pt}^{S} \cdot \max\left(0, \tau_{ij} - dd_{pt}\right)$ – monetary penalty incurred by the transporter for delivery deadline breach on routes servicing sales point p:
  - λ – scaling and normalization factor for the financial penalty ($\lambda \geq 0$);
  - C_{pt}^{S} – penalty rate of sales point p for stockout or delayed delivery in period t;
  - τ_{ij} – actual transit time spent by transport on route (i,j);
  - dd_{pt} – contracted delivery deadline for order fulfillment to retail sales point p in period t;
  - $\max\left(0, \tau_{ij} - dd_{pt}\right)$ – delivery time delay (if $\tau_{ij} \leq dd_{pt}$, the value strictly equals 0).

Engineering Practices and Architecture**.** For direct decision-making within the control loop, a Q-learning controller (Q-controller) is utilized – an algorithmic module based on automatic control theory and reinforcement learning. Its function is to estimate the $Q(s,a)$ function (*Quality* – the expected utility of executing action a in state s), enabling the agent to select an optimal logistics strategy without constructing an explicit mathematical model of the environment.

To formalize the operational execution, event processing, and adaptive behavior within the proposed Event-Driven Architecture (EDA), the interaction loop across the Event Bus, Q-controllers, and the core Supply Chain Model (SCM) environment is structured as a sequential workflow (Fig. 1).

The operational sequence depicted in Fig. 1 function through three continuous stages.

1. *Event Propagation & Strategy Formulation***:** At each execution step, Manufacturers (M_1, M_2) and Transporters (T_1, T_2) receive state update notifications via the Event Bus. Participants adapt to organizational conditions by choosing whether to integrate vertically or operate independently. Operational control variables serve as adaptation mechanisms: Manufacturers

determine output volumes (q_i), while Transporters establish freight tariffs (t_i). Strategies are selected by querying their respective Q-Controllers.

2. *Market Execution & Payoff Computation*: The SCM Model receives aggregated decision events from the Event Bus, evaluates vertical integration criteria, and computes dynamic market-clearing prices in response to stochastic demand fluctuations.

3. *Q-Controller Adaptation & Convergence:* Realized financial returns are routed back to Q-Controllers as reward signals. The Q-learning modules compute temporal-difference errors to update action-value matrices.

The learning process terminates when the variation in participants' objective reward functions stabilizes within a predefined threshold:

$$|R_{t+1} - R_t| < \varepsilon. \qquad (6)$$

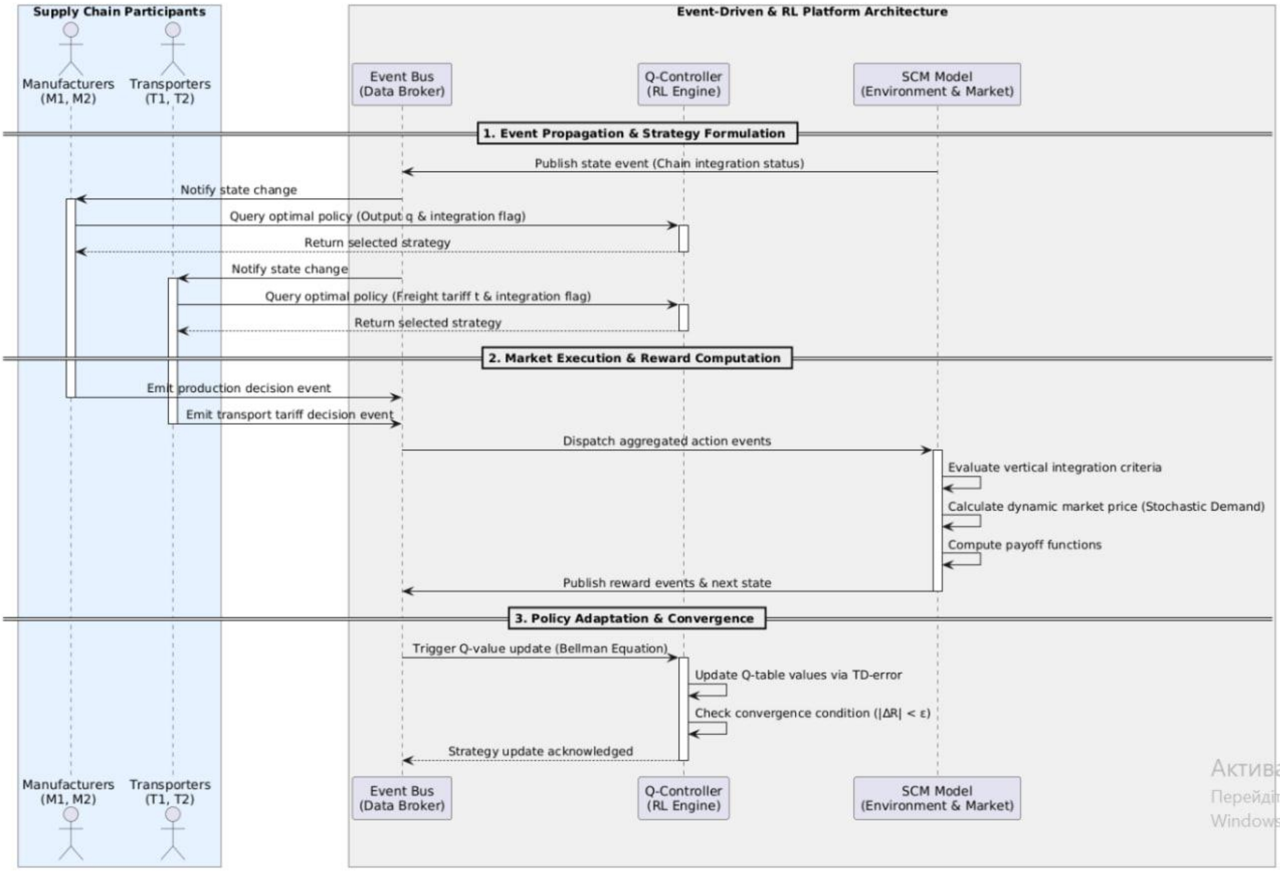

Fig. 1. **Sequence diagram of event flows, decision-making, and adaptive policy updates across SCM participants, Event Bus, and Q-controllers**

***Source*: compiled by the authors**

Objective Functions and Mathematical Formulation. To model participant decision-making under varying structural conditions, payoff functions are defined for each organizational state.

Decentralized Framework

Under decentralized operation (absence of integration), the objective functions for Manufacturer i and Transporter i in chain $i \in \{1,2\}$ are formulated as:

$$\Pi_{M_i}^{dec} = (P - v_i - t_i) \cdot q_i, \quad \Pi_{T_i}^{dec} = (t_i - w_i) \cdot q_i \qquad (7)$$

where v_i represents production unit costs; w_i denotes transportation costs per unit; t_i is the unit tariff set by transporter i; q_i is the output volume of manufacturer i and P is the market price governed by stochastic inverse demand:

$$P = \max(0, A - B(q_1 + q_2) + \xi), \quad \text{where } \xi \sim \mathcal{N}(0, \sigma^2) \qquad (8)$$

Vertically Integrated Framework. When vertical integration is unlocked and mutually selected, the joint supply chain generates an integrated profit based on the combined operational cost C_{int}:

$$\Pi_i^{int} = \left(P - c_{int,i}\right) \cdot q_i \tag{9}$$

Total chain earnings are split equally between the integrated entities due to equal power on profit distribution:

$$\Pi_{M_i}^{int} = 0.5 \cdot \max\left(0, \Pi_i^{int}\right), \quad \Pi_{T_i}^{int} = 0.5 \cdot \max\left(0, \Pi_i^{int}\right) \tag{10}$$

Q-Learning Control Model. Agents optimize their long-run cumulative rewards without prior knowledge of global environment transitions. The action-value function $Q(s,a)$ for each agent's Q-Controller is iteratively updated according to the Bellman optimality equation:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \cdot \left[R_{t+1} + \gamma \max_{a' \in A} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)\right] \tag{11}$$

where:

- $s_t \in S$ – represents the current state space vector at time t (incorporating the vertical integration status of both competing chains).
- $a_t \in A$ – is the executed action tuple at time t, defined as:
  - $a_{M_i} = (a_{int}, q_i)$ – for Manufacturers, combining integration decision $a_{int} \in \{0,1\}$ and production output volume q_i;
  - $a_{T_i} = (a_{int}, t_i)$ – for Transporters, combining integration decision $a_{int} \in \{0,1\}$ and freight tariff t_i.
- R_{t+1} – is the scalar financial payoff (reward) received by the agent after transitioning to state s_{t+1}.
- α – is the learning rate parameter ($\alpha \in (0,1]$).
- γ — is the discount factor for future rewards ($\gamma \in [0,1)$).
- $\max_{a' \in A} Q_t(s_{t+1}, a')$ – explicitly denotes the maximum expected Q-value across all available candidate actions $a' \in A$ in the next state s_{t+1}.

To operationalize the proposed Event-Driven Architecture (EDA) shown in Fig. 1, each functional stage of the system lifecycle is mapped onto the underlying architecture components (Event Bus, Q-Controllers, and SCM Environment Model). Table 1 formalizes how high-level architectural requirements are implemented as specific operational mechanisms across the distributed supply chain network.

Table 1. **Distribution of practices across the development stages of an Event-Driven SCM system**

No.	Development / Operational Stage	Targeted EDA Component(s)	System Architectural Mechanism & Implementation
1	Requirements Formalization	SCM Model, Manufacturers (M_i), Transporters (T_i)	Establishes baseline operational constraints (v_i, w_i) and service level metrics. Defines payoff structure (Π_i^{dec}, Π_i^{int}) balancing inventory holding costs against stockout penalties
2	EDA Infrastructure & Event Routing	Event Bus	Implements asynchronous message queues for event dispatches (StockLow, OrderPlaced, ShipmentArrived). Decouples state distribution from decision computation to eliminate latency
3	MDP State Space Formalization	Q-Controller, SCM Model	Discretizes state space vector $s_t \in S$ (representing integration status across chains) and action spaces $a_t = (a_{int}, q_i)$ or (a_{int}, t_i) to prevent the curse of dimensionality in model-free environments
4	Q-Controller Execution & Policy Update	Q-Controller (RL Engine)	Evaluates action utility via Bellman updates (eq. 11). Handles cold-start initialization using prior simulation history and enables real-time Q-table updates upon receiving reward events (R_{t+1}) via the Event Bus
5	Online Monitoring & System Convergence	Event Bus, Q-Controller	Monitors temporal difference errors in real time. Validates system convergence using the reward stabilization condition $\|R_{t+1} - R_t\| < \varepsilon$ (Eq.1) and triggers adaptive controller re-training upon structural market shifts

***Source*: compiled by the authors**

As detailed in Table 1, the architecture ensures that high-level economic objectives are translated into low-level software events processed by the Q-controllers. By anchoring each engineering practice to structural nodes, the system maintains real-time responsiveness without relying on centralized deterministic solvers.

Simulation Results

To clarify the experimental scope within the broader system framework, the simulation experiment in this study specifically evaluates the Supply Chain Interaction Layer (which governs dynamic vertical integration and pricing decisions between competing Manufacturers M_i and Transporters T_i). While the overall architecture spans multi-tier event handling, this experiment isolates the strategic equilibrium and Q-learning convergence mechanisms within a duopoly supply chain under stochastic market demand, serving as a functional validation module prior to full-scale network deployment. To validate the efficiency of the proposed approach, a simulation model of a duopoly market structure under demand uncertainty was developed using the Mesa 3 framework (Python). The model encompasses two parallel supply chains $i \in \{1,2\}$, each consisting of a Manufacturer (M_i) and a Transporter (T_i).

The inverse demand function incorporates random stochastic noise distributed according to a normal distribution $\varepsilon_t \sim N(0,\sigma^2)$:

$$P = \max(0, A - B(q_1 + q_2) + \xi), \quad \text{where } \xi \sim \mathcal{N}(0,\sigma^2) \tag{12}$$

where A represents the baseline market capacity parameter ($A = 280$), B is the demand slope parameter ($B = 1$), $\xi \sim \mathcal{N}(0,2)$ denotes normal stochastic noise with mean 0 and variance $\sigma^2 = 2$, and q_1, q_2 are the output volumes of manufacturers 1 and 2, respectively. Marginal production costs are defined as v_1, v_2, unit transportation costs as w_1, w_2, and total combined vertical integration costs as $c_{int,1}, c_{int,2}$.

The simulation was executed over 50 independent Monte Carlo runs, each lasting 1,200-time steps. The empirical results obtained were compared with the theoretical robust Cournot model [11] under decentralized planning with tariffs t_1 and t_2, respectively (Table 2).

Table 2. **Dynamics of market indicators across experiment stages under uncertainty (Mean ± SD over 50 Monte Carlo runs)**

Metric	Stage 1: Decentralization (steps 50-199)	Stage 2: Adaptation (steps 300-599)	Stage 3: Final state (steps 900-1149)	Theoretical Cournot Model
Output volume q_1	13.05 ± 6.15	13.00 ± 6.18	13.21 ± 6.34	30.28
Output volume q_2	13.55 ± 6.50	13.50 ± 6.41	13.51 ± 6.36	30.28
Delivery tariff t_1	18.21 ± 7.05	17.26 ± 7.25	16.89 ± 7.30	37.33
Delivery tariff t_2	18.46 ± 7.05	17.01 ± 8.24	16.84 ± 8.26	37.33
Market price (P)	88.16 ± 12.91	88.19 ± 13.37	87.90 ± 13.45	47.33
Supply Chain 1 integration share (%)	0.0 % (SD: 0.0%)	1.6 % (SD:12.6 %)	2.5 % (SD: 15.7 %)	-
Supply Chain 2 integration share (%)	0.0% (SD: 0.0%)	6.8% (SD: 25.2%)	7.1 % (SD: 25.7 %)	-
Reward M_1	728.86 ± 233.23	722.38 ± 202.86	728.42±205.58	549.90
Reward T_1	222.71 ± 141.67	218.53 ± 151.36	223.44 ± 162.33	824.85
Reward M_2	748.32 ± 238.75	721.00 ± 198.05	719.04 ± 194.87	549.90
Reward T_2	236.39 ± 153.37	255.03 ± 183.05	254.22 ± 181.95	824.85

Source: **calculated by the authors based on the results of the simulation experiment**

The adapted estimation of final Q-values for dominant states is presented in Table 3. The difference between the Cournot theoretical model and the experimental results is that, during the experiment, participants in the supply chains find a local optimum, whereas the Cournot model calculates a global optimum. The initial parameters of the participants and the discrete step at which changes occur in the supply chain system influence both the rate of adaptation and the isolation of the local optimum. When participants choose short-run guaranteed low rewards rather than short-run losses, a significant increase in profit is achieved with a certain probability. The current choice made by supply chain participants between guaranteed low profits and risky, high future profits, taking into account the discounting of future profits, accounts for the difference between the experimental Q-table approach and the theoretical predictions of the Cournot model.

Table 3. **Final Q-value estimates for Supply Chain 1 and Supply Chain 2 agents**

State $s = (S_1, S_2)$	Supply Chain	Integration Strategy	$Q(s,a)$ Manufacturer	$Q(s,a)$ Transporter	Dominant Choice (Q)
(0, 0)	SC1	Integration = 0	728.42	223.44	Dominant decentralization
(0, 0)	SC2	Integration = 0	719.04	254.22	Strategic decentralization
(0, 1)	SC2	Integration = 1	741.20	278.50	Local shift toward integration ($\Delta Q_{T2} = +24.28$)
(1, 1)	Both (SC1, SC2)	Integration = 1	635.10	198.20	Economically inefficient symmetric integration

Source: **compiled by the authors based on Q-table convergence from simulation results**

Analysis of Results and Discussion

An analysis of the obtained empirical data allows for the following scientific and practical conclusions:

1. **The Effect of "Conservative Output" Under Noisy Demand.** In the deterministic theoretical Cournot model, the output volume of each manufacturer reaches $q^* = 30.28$, forming a total supply of $Q = 60.56$ at a market price of $P = 47.33$. In contrast, under stochastic noise ε_t, Q-agents formed a restrained strategy with output volumes $q_1 = 13.21$ and $q_2 = 13.51$ (totaling $Q \approx 26.72$). This confirms that, under uncertainty, agents prefer a higher price ($P \approx 87.9$) to volume, thereby minimizing the risk of losses from stockouts or price collapses.

2. **Asymmetric Formation of Bargaining Power.** Agent learning led to a gradual reduction in transportation tariffs (from $t_1 = 18.21, t_2 = 18.46$ in Stage 1 to $t_1 = 16.89, t_2 = 16.84$ in Stage 3). Consequently, Manufacturers achieved higher average rewards ($R_{M1} = 728.42$ and $R_{M2} = 719.04$) than predicted by the theoretical decentralized model ($R_M^* = 549.90$), effectively capturing a portion of the transporters' profits.

3. **Limitations of Vertical Integration in a Noisy Environment.** Unlike deterministic models, where vertical integration is economically advantageous due to the elimination of double marginalization, agents in a noisy environment maintain a predominantly decentralized structure (integration shares are 2.5 % for SC 1 and 7.1 % for SC 2). This occurs because manufacturers, by restricting their own output volumes, increase product prices and thus secure higher profits without requiring long-run vertical integration. The feasibility of vertical integration emerges only when demand does not decline over a specific time interval. The integration of both supply chains leads to a significant increase in sales volume by eliminating double marginalization, which lowers prices and profits for both supply chains under the given logistics system parameters, ultimately rendering vertical integration economically unviable.

Conclusions. As a result of this research, an event-driven supply chain management architecture based on Q-learning was developed and scientifically substantiated. It has been experimentally proven that responding to events as they occur (via trigger handling through an Event Bus), combined with evaluating action values in Q-tables, and eliminates reliance on distorted static demand forecasts. The study confirmed that foundational architectural decisions regarding constraint formalization and the discretization of state and action spaces provide the most significant effect. Experimental validation in the Mesa 3 environment confirmed that, under uncertainty, Q-agents self-organize into a stable asymmetric Nash equilibrium, yielding a more risk-averse pricing strategy than deterministic calculations in the Cournot quantity-competition model.

As for prospects for future research then to fully generalize the experimental conditions across the entire Event-Driven Architecture (EDA), future research directions will focus on: (i) scaling the multi-agent Q-learning environment from a duopoly setting to a heterogeneous multi-echelon network incorporating real-time inventory management at retail nodes ($p \in P$), (ii) stress-testing the event-bus throughput and latency under high-frequency real-time event streaming, and (iii) integrating dynamic emission taxes (E_t^{tax}) alongside adaptive penalty structures (λ) to evaluate cross-layer system resilience across complex, non-stationary logistics ecosystems.

Conflict of Interest: The authors declare that they have no conflict of interest, including financial, personal, authorship, or any other nature that could have influenced the research or the results published in this article**.**

Funding: This research received no external funding.

Data Availability: All data are available in digital or graphical form within the main text of the manuscript.

Use of Artificial Intelligence Tools: During the preparation of this manuscript, Artificial Intelligence (AI) tools-specifically ChatGPT, Google Gemini, and Claude-were utilized for grammatical and stylistic editing. All text sections processed with the assistance of these tools were thoroughly reviewed and edited by the authors. AI was not used to generate text, formulas, or tables, nor to formulate scientific claims, derive results, or draw conclusions. All scientific statements, results, and conclusions represent the authors' own intellectual contribution**.**

REFERENCES

1. Samuels, A. “Examining the integration of artificial intelligence in supply chain management from Industry 4.0 to 6.0: a systematic literature review”. *Frontiers in Artificial Intelligence*. 2025; 7: 1477044. DOI: https://doi.org/10.3389/frai.2024.1477044.

2. Jiang, Z., Wu, D. & Cao, Y. “New-generation AI-driven intelligent decision-making and inventory optimization in the full lifecycle of complex product manufacturing integrating LSTM and Q-learning”. *Scientific Reports*. 2026; 16 (1): 11077. DOI: https://doi.org/10.1038/s41598-026-41629-6.

3. Yang, Y., Wang, M., Wang, J., Li, P. & Zhou, M. “Multi-Agent deep reinforcement learning for integrated demand forecasting and inventory optimization in sensor-enabled retail supply chains”. *Sensors*. 2025; 25 (8): 2428. DOI: https://doi.org/10.3390/s25082428.

4. Lu, Y. “A multimodal deep reinforcement learning approach for IoT-driven adaptive scheduling and robustness optimization in global logistics networks”. *Scientific Reports*. 2025; 15 (1): 25195. DOI: https://doi.org/10.1038/s41598-025-10512-1.

5. Wang, Y. & Liang, X. “Application of reinforcement learning methods combining graph neural networks and self-attention mechanisms in supply chain route optimization”. *Sensors*. 2025; 25 (3): 955. DOI: https://doi.org/10.3390/s2503955.

6. Huang, Y. & Li, W. “Interpretable reinforcement learning with structured policies for prescriptive supply chain analytics”. *Scientific Reports*. 2026; 16: 61734. DOI: https://doi.org/10.1038/s41598-026-61734-w.

7. Liu, X., Hu, M., Peng, Y. & Yang, Y. “Multi-agent deep reinforcement learning for multi-echelon inventory management”. *Production and Operations Management*. 2024; 34 (7): 1836–1856. DOI: https://doi.org/10.1177/10591478241305863.

8. Eze, J., Duan, Y., Eze, E., Ramanathan, R. & Ajmal, T. “Machine learning-based optimal temperature management model for safety and quality control of perishable food supply chain”. *Scientific Reports*. 2024; 14 (1): 27228. DOI: https://doi.org/10.1038/s41598-024-70638-6.

9. Fan, X., Zou, J., Liu, Y., Li, D., Zhou, D., Chi, H. & Li, D. “Transformative artificial intelligence integration in aquatic supply chains: synergizing precision aquaculture with intelligent logistics and data-driven consumption”. *Food Chemistry X*. 2026; 35: 103833. DOI: https://doi.org/10.1016/j.fochx.2026.103833.

10. Saeed, A., Jian, M., Imran, M., Dong, Z. S. & Freen, G. "Sustainable–resilient and climate-responsive e-waste supply chain network design under uncertainty using four-valued refined neutrosophic optimization". *Scientific Reports*. 2026; 16: 56894. DOI: https://doi.org/10.1038/s41598-026-56894-8.

11. Kobets, V., & Poltoratskiy, M. "Forming an evolutionarily stable firm strategy under Cournot competition using social preferences". *Communications in Computer and Information Science*. 2014; 469: 343-361. DOI: https://doi.org/10.1007/978-3-319-13206-8_17.

Received 26.07.2026
Received after revision 24.08.2026
Accepted 02.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.24
УДК 004.8:658.7:519.816

Керована подіями архітектура ланцюжків постачання на основі Q-навчання в умовах невизначеності

Кобець Віталій Миколайович[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@ksu.ks.ua. Scopus Author ID: 56006224700
Кльонов Дмитро Михайлович[1)]
ORCID: https://orcid.org/0009-0004-8537-3806; dklonov@ksu.ks.ua. Scopus Author ID: 55965109600
[1)] Херсонський державний університет, вул. Університетська, 27. Херсон, 73000, Україна

АНОТАЦІЯ

У роботі розглянуто концепцію побудови керованої подіями архітектури (Event-Driven Architecture, EDA) для систем управління ланцюгами постачання (Supply Chain Management, SCM) із використанням Q-learning в умовах невизначеності. Обґрунтовано, що за наявності високої волатильності попиту, затримок у постачаннях та викривлень у вихідних даних детерміновані й класичні статистичні моделі припускаються значних помилок, що призводить до зростання операційних витрат через дефіцит або надлишок товарного запасу. Автори підкреслюють доцільність переходу від циклічного розрахунку прогнозів за розкладом до реагування на події в логістичній мережі у реальному часі за допомогою програмної шини подій (Event Bus). Метою дослідження є розробка керованої подіями архітектури прийняття рішень, яка поєднує шину подій, контролер на основі навчання з підкріпленням (Reinforcement Learning, RL) та асиметричну функцію винагороди. Основна увага приділяється відмові від прямого відновлення функції попиту на користь оцінки оптимальності дій у конкретних станах системи та пошуку балансу між витратами на утримання запасів і ризиками дефіциту. У роботі систематизовано ключові рішення: формалізація марковського процесу прийняття рішень (Markov Decigion Process, MDP) без побудови математичної моделі середовища (model-free підхід); асинхронна обробка подій у вузлах мережі; використання Q-learning як інтерпретованого та обчислювально легкого методу. Практичний аналіз доводить, що застосування запропонованого підходу забезпечує стійкість системи до зашумлених даних та мінімізує часові затримки під час прийняття управлінських рішень.

Ключові слова: ланцюжки постачання; керована подіями архітектура; Q-навчання; алгоритми прийняття рішень; невизначеність попиту; штучний інтелект у логістиці

ІНФОРМАЦІЯ ПРО АВТОРІВ

Vitaliy M. Kobets - Doctor of Economic Science, Professor, Department of Computer Science and Software Engineering. Kherson State University. 27, Universitetska Str. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@ksu.ks.ua. Scopus Author ID: 56006224700
***Research field*:** data science in economics; evolutionary microeconomics; robo-advisors

Кобець Віталій Миколайович – доктор економічних наук, професор, професор кафедри Комп'ютерних наук та програмної інженерії. Херсонський державний університет, вул. Університетська, 27. Херсон, 73000, Україна
***Галузь досліджень*:** інтелектуальний аналіз даних в економіці; еволюційна мікроекономіка; автоматизовані фінансові консультанти.

Dmytro M. Klonov - PhD Student, Lecturer, Department of Computer Science and Software Engineering. Kherson State University, 27 Universytetska St. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0009-0004-8537-3806; dklonov@ksu.ks.u. Scopus Author ID: 55965109600
***Research field*:** software engineering

Кльонов Дмитро Михайлович – аспірант, викладач кафедри Комп'ютерних наук та програмної інженерії. Херсонський державний університет, вул. Університетська, 27. Херсон, 73000, Україна
***Галузь досліджень*:** програмна інженерія