JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

Contextual Reasoning for Embodied Supply Chain Agents: Reinforcement Learning Policies from Physical State Perception to Collaborative Execution

Marta Ribeiro[1], Diogo Fernandes[2], Helena Costa[3,*], Pedro Almeida[4]

[1] Department of Industrial Engineering, University of Minho, Guimarães, Portugal
[2] Department of Informatics, University of Évora, Évora, Portugal
[3] School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal
[4] Department of Electromechanical Engineering, University of Beira Interior, Covilhã, Portugal
* Corresponding author: helena.costa@ipleiria.pt

ARTICLE INFO
Received
October 08, 2025
Revised
December 11, 2025
Accepted
February 17, 2026
Available Online
March 30, 2026
DOI
10.63646/jaiaa.2026.040105

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Physical supply chains increasingly rely on artificial intelligence, autonomous mobile robots, computer vision, edge sensors, and digital twins, yet many decision systems still reason over abstract data tables rather than over the physical state in which execution takes place. This paper develops a contextual reasoning framework for embodied supply chain agents that connects physical state perception, reinforcement learning policy design, and collaborative execution across warehousing, sorting, and last-mile delivery. The proposed framework defines the agent state as a multimodal representation of spatial congestion, shelf load, equipment utilization, order urgency, task risk, and inter-agent dependency. A reward architecture is then formulated to balance fulfillment time, execution accuracy, resource utilization, safety, and policy stability. To demonstrate analytic value, the study constructs an illustrative multi-agent simulation of a three-link supply chain operation involving storage robots, sorting arms, and delivery vehicles. Compared with static rule dispatching, collaborative contextual reinforcement learning reduces average fulfillment time by 30.7%, late-order rate by 51.1%, and near-miss events by 62.1% under the stated scenario assumptions. The analysis shows that contextual reasoning improves not merely prediction accuracy but also the coupling between digital decisions and physical execution. The contribution of the paper is a policy-oriented analytics model that translates embodied supply chain intelligence into implementable reinforcement learning structures, evaluation indicators, and deployment guidelines for AI-enabled adaptive operations.
Keywords: embodied intelligence; supply chain agents; reinforcement learning; contextual reasoning; multi-agent systems; physical state perception; collaborative execution; AI analytics

I. INTRODUCTION

Supply chains have become distributed physical-digital systems in which decisions generated by algorithms must be executed by people, machines, vehicles, shelves, sensors, and robots. A demand forecast, a dispatch rule, or an inventory recommendation has no operational value until it is translated into a physical action that respects congestion, equipment limits, handling risk, labor availability, and delivery time windows. This observation is simple, but it exposes a limitation in a large share of intelligent supply chain research. The supply chain is often represented as a set of data tables: orders, inventory records, transport times, supplier scores, and demand histories. Artificial intelligence models then optimize these tables by forecasting, classifying, or ranking them. Such models have contributed to demand planning, inventory management, and route optimization, but they often ignore the embodied nature of execution, where decisions are

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

constrained by spatial layout, sensor uncertainty, force interaction, machine posture, and dynamic collaboration among heterogeneous agents (Ivanov, 2020).

The emerging idea of embodied intelligence offers a useful correction. In embodied intelligence, an agent does not merely compute over a symbolic description of the environment. It learns from interaction with the physical world, interprets context from sensory input, and adapts its action policy through feedback. In a supply chain setting, the agent may be a warehouse robot choosing a path through narrow aisles, a sorting arm adjusting grasping force for fragile products, a delivery vehicle revising a route under traffic uncertainty, or an orchestration service coordinating multiple machines under shared capacity constraints. Intelligence is therefore not located only inside an algorithm. It is distributed across perception, reasoning, actuation, and feedback (Kober et al., 2013).

Recent work on embodied intelligent supply chains increasingly treats supply chain intelligence as a layered physical-digital capability rather than as a detached forecasting module. This article narrows that broader view to one unresolved analytical question: how can contextual reasoning be formulated as a reinforcement learning policy that receives physical state perception and produces collaborative execution across supply chain agents? The emphasis on policy structure is consistent with management analytics research that treats decision value as a measurable and operationally embedded capability (Lu, 2021).

This question is practically important because the gap between planning and execution remains a major source of operational inefficiency. A routing model may recommend a path that is theoretically shortest but physically congested. A warehouse picking algorithm may maximize order batching but increase damage risk when fragile items are handled under high arm load. A dispatch rule may use average travel time while ignoring that the next vehicle has low battery and the unloading dock is blocked. These examples show why contextual reasoning must encode the physical state of the system rather than treating execution as a frictionless downstream step (Tao et al., 2018).

Reinforcement learning is especially relevant because it frames decision-making as sequential interaction. An agent observes a state, chooses an action, receives a reward, and updates its policy. The approach fits adaptive supply chains because operational consequences are delayed, interdependent, and scenario specific. However, direct application of reinforcement learning to supply chains is not enough. A generic reward function that rewards speed may create unsafe behavior; a policy trained on historical averages may fail under disruption; and a single-agent formulation may improve a local node while shifting congestion to another node. The central argument of this paper is that reinforcement learning for embodied supply chain agents should be contextual, collaborative, and physically grounded (Mnih et al., 2015).

The paper contributes to the Journal of AI Analytics and Applications in three ways. First, it defines a contextual state representation that integrates physical, operational, and collaborative variables into a reinforcement learning-ready form. Second, it designs a reward architecture that balances efficiency, accuracy, utilization, safety, and policy stability rather than reducing supply chain performance to a single time or cost metric. Third, it provides an illustrative data analysis using a multi-agent simulation to compare static rules, digital-only learning, contextual single-agent learning, and collaborative contextual learning. The results are not presented as universal empirical claims; they are intended to demonstrate how the proposed framework can be operationalized, measured, and refined.

The literature on supply chain integration shows that information systems create value only when they connect planning routines with operational execution routines (Gunasekaran and Ngai, 2004). Earlier work on artificial intelligence in supply chain management already identified planning, forecasting, and logistics decision support as promising domains for intelligent systems (Min, 2010). Recent reviews of artificial intelligence applications confirm that supply chain AI has moved from isolated prediction models toward integrated decision systems (Pournader et al., 2021). Industry 4.0 research further shows that intelligent operations depend on the convergence of sensors, connectivity, analytics, and cyber-physical coordination (Lu, 2025).

Big-data capability research suggests that analytics value is realized through dynamic capability rather than through data volume alone (Wamba et al., 2017). The state-of-the-art analysis of artificial intelligence clarifies why learning models must be linked to application context to avoid purely technical abstraction (Zhang and Lu, 2021). Predictive analytics research in supply chains demonstrates that operational performance improves when data-driven models are embedded in organizational processes (Gunasekaran et al., 2017). Operations analytics research also indicates that big-data methods reshape decision horizons, uncertainty handling, and managerial accountability (Choi et al., 2018). A broader survey of artificial intelligence also shows that model evolution, application embedding, and future trends should be evaluated together rather than separated into technical silos (Lu, 2019a).

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

II. THEORETICAL BACKGROUND

Three research streams support the proposed model: embodied intelligence, reinforcement learning, and AI-enabled supply chain analytics. Embodied intelligence emphasizes that cognition is shaped by the agent's body, sensors, actions, and environmental constraints. In robotics, this principle has long motivated research on adaptive grasping, locomotion, manipulation, and sensorimotor control. The same logic applies to supply chains when orders, packages, pallets, vehicles, arms, and workers interact inside physical spaces. A supply chain agent that only sees a demand table is disembodied. A supply chain agent that sees aisle congestion, shelf height, robotic posture, product fragility, and service urgency is capable of contextual reasoning (Kober et al., 2013).

AI-enabled supply chain analytics has made strong progress in forecasting, maintenance, quality inspection, routing, and risk assessment. Deep learning models support demand prediction from complex time-series signals, graph models represent inter-firm and transport networks, and optimization heuristics support scheduling under resource constraints. Digital twins add a synchronized representation of physical assets and process states, thereby creating a bridge between sensor data and decision models (Tao et al., 2018). Yet the presence of a digital twin does not automatically produce embodied intelligence. A twin may provide visualization without policy learning, or it may mirror physical assets without allowing execution feedback to alter decision logic. The proposed framework treats the twin not as an end in itself but as one component of a contextual reasoning loop.

Reinforcement learning extends traditional analytics by emphasizing action and feedback. The Markov decision process provides the basic structure: state, action, transition, reward, and discounting. Deep reinforcement learning made it possible to learn policies from high-dimensional inputs, while policy-gradient and actor-critic methods expanded the range of continuous and hybrid control tasks (Mnih et al., 2015). In supply chains, reinforcement learning has been explored for inventory control, dynamic pricing, routing, scheduling, and order fulfillment. However, many applications remain either highly abstract or locally optimized. They often assume simplified states and action spaces that do not fully reflect physical execution constraints.

Multi-agent reinforcement learning adds another layer of relevance. Supply chains are rarely controlled by a single agent. A warehouse robot, a sorting robot, a yard vehicle, and a last-mile delivery van each have a local policy, but their actions interact. A robot that moves too slowly may increase sorting congestion; a sorting decision may affect vehicle departure; a vehicle rerouting decision may change loading priorities. Multi-agent learning therefore provides vocabulary for coordination, negotiation, and emergent behavior under shared objectives (Lowe et al., 2017). In an embodied supply chain, collaboration is not merely data sharing. It is the adjustment of physical actions in response to the perceived and expected actions of other agents.

Contextual reasoning is the bridge among these streams. In the present paper, contextual reasoning means the ability of an agent to interpret an observed physical state, embed it into a decision-relevant representation, evaluate the action consequences under current constraints, and coordinate execution with other agents. The concept differs from conventional optimization in two respects. First, it treats the environment as dynamic and partially observable rather than fixed. Second, it treats execution feedback as a source of policy improvement rather than a post-hoc performance report. This makes contextual reasoning well suited to operations where demand, congestion, equipment condition, and human-machine interaction change during execution.

The proposed model is also aligned with analytics maturity in smart manufacturing and logistics. Basic analytics describe what happened, diagnostic analytics explains why it happened, predictive analytics estimate what may happen, and prescriptive analytics recommends what to do. Embodied contextual reasoning adds a fifth layer: adaptive execution analytics. It asks whether the recommended action remains feasible and beneficial now of physical execution. This added layer is essential in warehouse and logistics systems where time delays of even a few minutes can invalidate a decision.

Management analytics provides a useful bridge between computational optimization and organizational decision value because it treats the decision itself as the unit of analysis (Lu, 2021). Deep reinforcement learning for inventory control demonstrates that sequential decision models are relevant to supply chains but require careful design of state, action, and reward spaces (Boute et al., 2022). Disruption simulation research shows that adaptive policies must be evaluated under unstable network conditions rather than only under steady-state demand (Ivanov, 2020). Ripple-effect studies further reveal that local events can propagate across supply chain nodes in ways that static optimization misses (Dolgui et al., 2018).

Supply chain survivability research argues that resilience should include viability under intertwined network conditions, which aligns with the collaborative perspective used here (Ivanov and Dolgui, 2020). Blockchain and supply chain management studies show that trusted digital coordination can improve cross-firm information sharing, although physical execution remains

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

a separate challenge (Queiroz et al., 2020). Blockchain-in-Industry 4.0 research broadens the governance perspective by connecting manufacturing processes with distributed information infrastructures (Chen et al., 2024). Cyber-physical Industry 4.0 research explains why supply chain agents should be treated as embedded components of sensing, computation, and actuation loops (Lu, 2017a).

III. CONTEXTUAL STATE REPRESENTATION FOR EMBODIED AGENTS

The first design challenge is the definition of state. A reinforcement learning policy is only as useful as the state it observes. In many supply chain models, the state includes inventory levels, demand forecasts, backlog, and processing times. These variables are necessary but insufficient for embodied agents. The state must also capture what the agent can physically do now, what constraints shape safe action, and how the agent's behavior will influence other nodes. For this reason, the proposed state representation contains five groups of variables: physical scene variables, equipment variables, order variables, collaborative variables, and uncertainty variables.

Physical scene variables describe the spatial and environmental condition of execution. Examples include aisle occupancy, dock queue length, shelf load, pallet position, obstacle density, road congestion, lighting quality, and temperature deviation. Equipment variables describe the agent's own capability at the moment of decision, such as battery level, robotic arm load, actuator temperature, gripper confidence, sensor health, and communication latency. Order variables include priority, deadline, fragility, volume, weight, and service class. Collaborative variables encode the expected actions or states of other agents, including nearby robot trajectories, sorting line load, vehicle departure plans, and shared resource reservations. Uncertainty variables record confidence scores from perception and prediction models.

Figure 1 visualizes the proposed context field without using directional arrows. The purpose of the figure is to emphasize that contextual reasoning is not a linear pipeline at the state-definition level. The physical state field, context memory, and policy space coexist as structured representations that the agent must bind together before action selection. A policy that only sees the physical field may react myopically. A policy that only sees context memory may be slow to respond to disruption. A policy that only sees the feasible action space may ignore why some actions are safer or more collaborative than others. The contextual state therefore serves as the representational foundation for embodied policy learning.

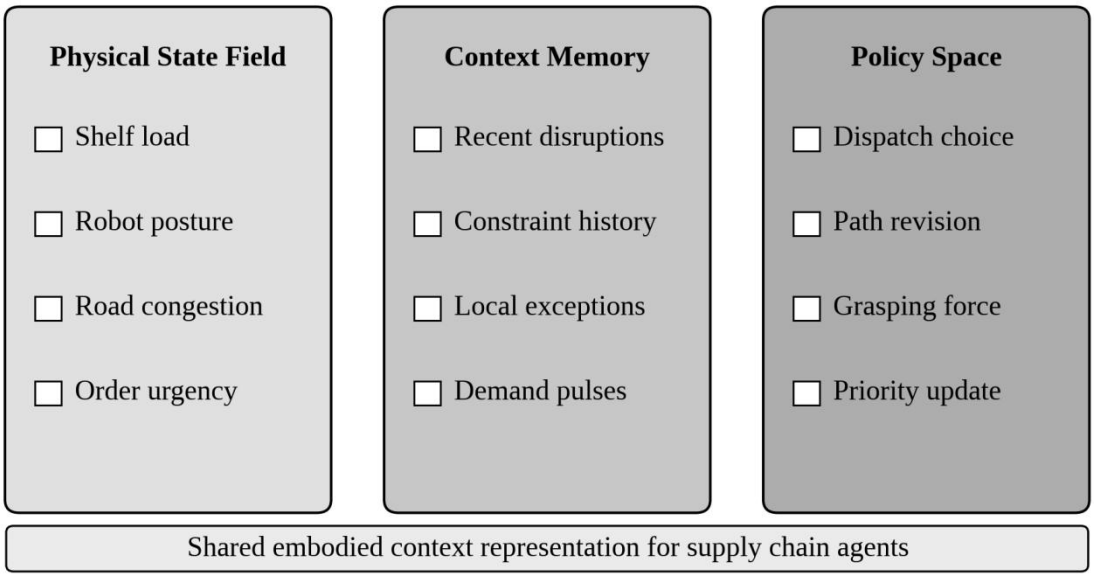


Figure 1. Contextual state field for embodied supply chain agents.

The state representation can be written in compact form as $s_t = [x_t^p, x_t^e, x_t^o, x_t^c, x_t^u]$, where x_t^p denotes physical scene features, x_t^e denotes equipment features, x_t^o denotes order features, x_t^c denotes collaborative features, and x_t^u denotes uncertainty features. This expression is intentionally simple. The goal is not to create a heavy mathematical model but to make the information boundary of the policy explicit. Each component can be populated from different sources, including lidar,

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

computer vision, force sensors, warehouse management systems, transport management systems, edge devices, and digital twin event logs.

A second issue is temporal context. A snapshot of congestion is less informative than a trajectory of congestion. A shelf-load reading is more useful when compared with previous readings and expected replenishment activity. The framework therefore uses a rolling state window. Short-term windows capture immediate execution risk, while longer windows capture demand pulses, recurring congestion, and equipment drift. Recurrent neural networks, temporal convolution, or attention-based encoders can compress these sequences into a policy input. For transparency, the model should retain interpretable state channels even when deep encoders are used (Vaswani et al., 2017).

A third issue is partial observability. Supply chain agents rarely see the whole system. A delivery vehicle may not know the internal status of the sorting line; a warehouse robot may not know whether a downstream dock will be blocked in ten minutes. The proposed framework addresses this limitation through shared context memory. Each agent publishes a compact state message to a local coordination service. The shared memory does not need to expose all raw data. It only needs to expose task-relevant variables such as predicted time of arrival, capacity commitment, risk level, and exception status. This design reduces communication burden while supporting coordinated action.

State design also affects fairness and robustness. If priority variables dominate the state, the policy may consistently favor high-value orders and delay ordinary orders. If safety variables are omitted, the policy may learn risky shortcuts. If uncertainty variables are ignored, the policy may overreact to noisy perception. The proposed representation therefore treats uncertainty and safety as first-class state components, not afterthoughts. This is important because embodied agents operate near physical objects, people, and infrastructure.

IoT security research shows that sensor-rich environments require integrity, authentication, and resilient communication before physical-state data can support reliable decisions (Xu et al., 2021). Research on IoT cybersecurity clarifies that connected devices expand both visibility and attack surfaces in operational systems (Lu and Xu, 2019). Industry 4.0 technology reviews show that interoperability, automation, and cloud-edge coordination are necessary foundations for physical-digital state representation (Lu, 2017b). Blockchain research contributes a complementary view of traceability, auditability, and data integrity across distributed decision environments (Lu, 2019b).

Q-learning established the value of updating action values through experience rather than relying on fixed rules (Watkins and Dayan, 1992). Temporal-difference learning introduced a practical mechanism for learning from delayed outcomes, which is central to logistics execution feedback (Sutton, 1988). Policy-gradient learning explains how action-selection policies can be optimized directly when the action space is complex or stochastic (Williams, 1992). The classic reinforcement learning survey provides the conceptual vocabulary for states, actions, rewards, and exploration used in the proposed framework (Kaelbling et al., 1996).

Table I. Contextual state components for embodied supply chain agents.

State Component	Representative Variables	Data Sources	Reasoning Value
Physical scene	Aisle occupancy, shelf load, dock queue, obstacle density, road congestion	Lidar, cameras, RFID, GPS, facility sensors	Determines whether a planned action is spatially and temporally feasible
Equipment condition	Battery, gripper confidence, arm load, actuator temperature, communication latency	Robot telemetry, edge gateways, maintenance logs	Prevents assignment of actions that exceed current machine capability
Order requirements	Priority, deadline, fragility, weight, volume, service class	WMS, OMS, TMS, customer service records	Links physical execution with contractual and customer-facing obligations
Collaborative context	Nearby robot trajectory, downstream load, vehicle capacity, resource reservation	Shared context memory, digital twin event logs	Supports coordination rather than isolated local optimization
Uncertainty and risk	Sensor confidence, forecast variance, anomaly score, policy confidence	Perception models, forecasting models, safety monitors	Allows the policy to reduce autonomy or request confirmation under low confidence

IV. REINFORCEMENT LEARNING POLICY DESIGN

The second challenge of design is the formulation of the policy. A contextual policy maps the embodied state to an action that can be executed in the physical system. The action may be discreet, such as selecting a dispatch queue or choosing among predefined routes. It may be continuous, such as adjusting grasping force or robot speed. It may also be hybrid, such as choosing

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

a delivery priority and then setting a path parameter. The framework therefore does not assume one algorithm for all tasks. Deep Q-learning is suitable for discrete decisions, proximal policy optimization is useful for stable policy-gradient learning, and actor-critic methods are appropriate for continuous or hybrid control (Mnih et al., 2015).

For an embodied supply chain agent, the policy objective should not be defined only by local speed. A warehouse robot that minimizes its travel time may cause congestion in a shared aisle. A sorting arm that maximizes throughput may increase damage on fragile items. A vehicle that minimizes distance may miss a time window if the unloading dock is unavailable. The reward function must therefore integrate multiple dimensions of operational value. This paper defines the reward as a weighted combination of efficiency, accuracy, utilization, safety, collaboration, and stability. A compact expression is given below:

$$R_t = w_1E_t + w_2A_t + w_3U_t + w_4S_t + w_5C_t + w_6B_t, \quad \sum_i w_i = 1.$$

In this expression, E_t measures efficiency improvement, A_t measures execution accuracy, U_t measures resource utilization, S_t measures safety compliance, C_t measures collaborative contribution, and B_t measures policy stability. The weights are not fixed universally. A pharmaceutical cold-chain operation may assign higher weight to accuracy and safety; a high-volume e-commerce warehouse may emphasize efficiency and utilization; a human-robot collaborative workspace may increase the safety component. The reward design should therefore be governed by the operational context rather than copied across sites.

The policy can be optimized by maximizing expected discounted reward over a finite horizon. The operational meaning is straightforward: the agent should select actions that perform well now without creating negative consequences later. For example, an action that clears a local queue quickly may be penalized if it causes downstream overload. Similarly, a route that shortens distance may be penalized if it increases near-miss risk or battery depletion. This delayed structure is one reason reinforcement learning is attractive for embodied supply chain operations.

Figure 2 presents the policy learning cycle. It starts with multimodal perception, converts sensor and system data into state encodings, produces a contextual policy, translates the selected action into collaborative execution, and receives reward and feedback. The feedback is not limited to success or failure. It includes fulfillment time, damage, energy, idle time, human override, collision warning, rework, and synchronization delay. These execution outcomes update both the policy model and the context representation. The cycle makes the policy sensitive to the physical consequences of its own decisions.

The action space should be designed with operational safeguards. In physical supply chains, unconstrained exploration is unacceptable. A robot should not test unsafe speed, collision-prone paths, or excessive gripper force on real goods. The framework therefore uses constrained exploration. Unsafe actions are filtered by a rule-based safety layer, and uncertain high-impact actions require human confirmation during early deployment. Digital twins and simulation environments are used to train and test policies before physical rollout. This approach is consistent with safe reinforcement learning, where policy improvement is pursued under constraints on risk and feasibility (Achiam et al., 2017).

A further design issue is collaboration. In a multi-agent supply chain, each agent has local observations and actions, but the objective is system-level performance. The framework uses centralized-training and decentralized-execution logic. During training, the policy has access to richer joint-state information from the simulation or digital twin. During execution, each agent acts on its local state and shares context messages. This arrangement allows the system to learn coordination patterns while preserving real-time responsiveness. It also reduces the communication burden that would arise if every agent needed complete system information at every step (Lowe et al., 2017).

Deep Q-networks showed that reinforcement learning can process high-dimensional inputs when representation learning and control are coupled (Mnih et al., 2015). Double Q-learning improved value estimation by addressing overestimation bias, which is relevant when supply chain rewards are noisy (Van Hasselt et al., 2016). Robotics reinforcement learning research shows that physical action policies must account for embodiment, contact, sensing error, and safety (Kober et al., 2013). A broad deep reinforcement learning survey confirms that algorithm selection should reflect observation type, action continuity, and data efficiency (Arulkumaran et al., 2017).

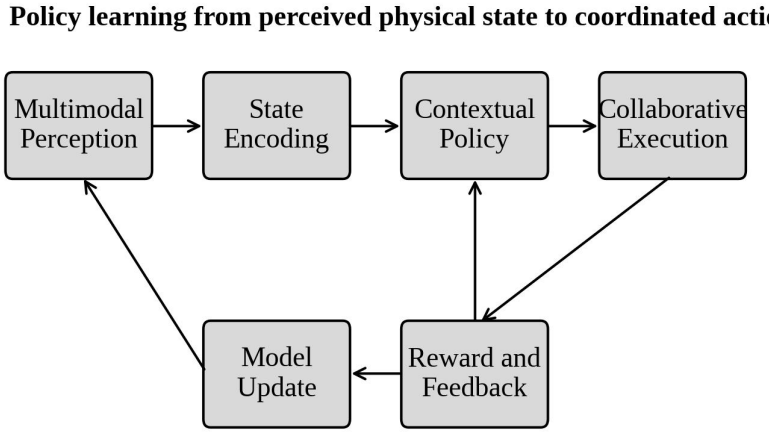
Continuous-control research is useful for embodied supply chain agents because robotic speed, grasping force, and vehicle acceleration are often continuous rather than discrete decisions (Lillicrap et al., 2016). Trust-region methods indicate that stable policy improvement is important when unsafe policy jumps could disrupt physical execution (Schulman et al., 2015). Proximal policy optimization provides a practical policy-gradient approach for balancing learning progress and updating stability (Schulman et al., 2017). Visuomotor learning demonstrates that perception and control can be trained jointly when sensory inputs are close to the execution surface (Levine et al., 2016).

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

Meta-learning is relevant to supply chain adaptation because local facilities may need fast policy adjustment after demand shocks or equipment changes (Finn et al., 2017). Maximum-entropy reinforcement learning highlights the value of exploration under uncertainty while maintaining stable expected behavior (Haarnoja et al., 2018). Twin-delayed actor-critic methods address function-approximation error in continuous control, which is useful for equipment-intensive supply chain tasks (Fujimoto et al., 2018). Multi-agent actor-critic methods provide a foundation for learning coordination among heterogeneous warehouse, sorting, and transport agents (Lowe et al., 2017).

Table II. Reward architecture for contextual reinforcement learning policies.

Reward Element	Positive Signal	Penalty Signal	Operational Rationale
Efficiency	Shorter fulfillment time; faster queue clearance	Late completion; excessive waiting	Encourages timely execution without assuming speed is the only goal
Execution accuracy	Correct item, correct location, stable grasping	Damage, mis-sort, rework, failed handoff	Connect policy reward to physical quality of action
Resource utilization	Balanced robot, arm, vehicle, and dock use	Idle capacity or overload	Prevents local decisions from creating downstream bottlenecks
Safety	Safe separation, low force risk, low near-miss probability	Collision warning, risky path, excessive speed	Makes physical safety a core optimization target
Collaboration	Synchronized transfer and downstream readiness	Queue transfer and capacity conflict	Rewards system-level coordination among agents
Stability	Small policy oscillation under similar states	Frequent reversal or excessive re-planning	Improves trust and reduces operational disturbance



The agent does not optimize an abstract order list alone; it continuously re-encodes physical constraints, agent states, and execution outcomes into a reward-sensitive reasoning cycle.

Figure 2. Reinforcement learning cycle from physical state perception to collaborative execution.

V. ILLUSTRATIVE ANALYTICS DESIGN AND DATA SETUP

To demonstrate how the framework can be evaluated, the paper constructs an illustrative analytics design based on a three-link supply chain operation. The scenario includes an inbound storage zone, a sorting and packing zone, and a last-mile dispatch zone. The agent population includes eight autonomous mobile robots, three robotic sorting arms, two human exception-handling stations, and six delivery vehicles. The scenario is intentionally moderate in scale so that the policy logic remains interpretable. It represents the kind of warehouse-delivery operation found in regional e-commerce fulfillment, spare-parts distribution, or urban retail replenishment.

The simulation horizon contains 5,760 decision epochs, corresponding to thirty operating days with eight hours per day and a decision interval of two and a half minutes. Each epoch records order arrivals, shelf load, aisle congestion, robot battery status, sorting line utilization, arm load, product fragility, vehicle capacity, dock queue length, and delivery time-window pressure. Disruptions are injected as road congestion spikes, temporary robot downtime, unexpected fragile-product clusters,

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

and dock blockage events. The data generated are not claimed to represent a specific company. They provide a controlled test bed for comparing policy behavior under consistent assumptions.

Four policy configurations are compared. The first is static rule dispatching, where orders are prioritized by deadline and robots follow shortest feasible paths. The second is a digital-only policy that learns from order and inventory data but does not include physical context channels. The third is contextual single-agent reinforcement learning, in which each node learns from physical state variables but optimizes locally. The fourth is collaborative contextual reinforcement learning, in which agents learn under shared context memory and system-level reward. This comparison isolates the incremental value of physical state perception and collaborative execution.

Multi-agent reinforcement learning theory shows that shared objectives and local observations create coordination challenges that differ from single-agent control (Zhang et al., 2021). Value-factorization research illustrates how a joint team objective can be decomposed into agent-level utilities for cooperative decision making (Rashid et al., 2018). Counterfactual policy-gradient methods show how multi-agent learning can evaluate an individual agent's contribution to a collective outcome (Foerster et al., 2018). Large-scale multi-agent game research demonstrates that coordinated learning systems can discover nontrivial strategies from interaction trajectories (Vinyals et al., 2019).

Deep search and neural learning in complex games illustrate how learned policies can combine perception, value estimation, and sequential planning (Silver et al., 2016). Self-play without human examples demonstrates how feedback loops can produce improved policy behavior when the objective and environment are well formalized (Silver et al., 2017). Kalman filtering remains a foundational approach for transforming noisy dynamic measurements into estimated state representations (Kalman, 1960). Evidence-theoretic reasoning offers a formal way to represent uncertainty when multiple sensor streams provide incomplete or conflicting information (Dempster, 1967).

Table III summarizes the main variables and scenario assumptions used in the illustrative analysis. The variables were selected to correspond to the state components developed earlier. Physical scene variables capture where execution occurs; equipment variables capture what agents can do; order variables capture service obligations; collaborative variables capture inter-agent dependency; uncertainty variables capture data reliability and operational risk. This structure makes the simulation data directly traceable to the proposed state representation rather than a generic logistics model.

Table III. Scenario variables and descriptive assumptions for the illustrative simulation.

Variable Group	Example Variables	Range or Distribution	Operational Meaning
Demand and orders	Order arrivals, priority, deadline pressure	Poisson arrivals with peak multipliers; priority classes 1-3	Creates fluctuating workload and time-window pressure
Warehouse scene	Shelf load, aisle occupancy, obstacle density	0-1 normalized state channels updated every epoch	Represents the physical context of storage and picking
Sorting execution	Arm load, fragile-item share, mis-sort risk	Arm utilization 0-100%; fragile cluster events injected	Captures physical handling constraints and damage exposure
Transport state	Vehicle capacity, road congestion, dock queue	Capacity units 20-80; congestion index 0-1	Links internal fulfillment to last-mile dispatch feasibility
Uncertainty	Sensor confidence, communication latency, forecast variance	Confidence 0.75-0.99; latency 30-260 ms	Determines whether autonomous execution should be bounded or escalated

VI. RESULTS AND INTERPRETATION

The comparative results show a clear performance gradient. Static rule dispatching performs adequately under stable demand but degrades when disruptions occur. The digital-only policy improves some efficiency metrics because it learns from order patterns, but it remains weak under physical congestion and equipment variation. Contextual reinforcement learning improves fulfillment time, damage reduction, and idle-rate control because the agent observes shelf load, congestion, battery, and fragility. Collaborative contextual reinforcement learning performs best because it coordinates actions across storage, sorting, and dispatch rather than optimizing each node in isolation.

Table IV reports the policy comparison. Collaborative contextual reinforcement learning reduces average fulfillment time from 38.7 minutes to 26.8 minutes. It reduces late-order rate from 18.2% to 8.9%, damage rate from 6.3 to 3.0 events per thousand orders, and equipment idle rate from 28.5% to 13.9%. Energy per hundred orders falls from 74.6 to 66.1 normalized units. Near-miss events also decline, suggesting that the policy does not achieve speed by sacrificing safety. These results indicate that physically grounded contextual reasoning can improve multiple dimensions simultaneously when reward design is balanced.

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

Figure 3 converts the table into percentage improvements over static rules. The improvement is largest for near-miss reduction and late-order reduction, followed by damage-rate and idle-rate improvements. Fulfillment time also improves substantially. Energy reduction is smaller because the scenario includes fixed energy consumption associated with baseline facility operation. This pattern is realistic: learning policies often have more room to improve coordination and timing than to reduce all energy use in the short term.

The training dynamics provide additional insight. Figure 4 compares generic reinforcement learning, contextual reinforcement learning, and collaborative contextual reinforcement learning over 210 training episodes. Generic reinforcement learning improves slowly because its state representation lacks physical and collaborative features. Contextual reinforcement learning converges faster because the policy can associate execution outcomes with meaningful state channels. Collaborative contextual reinforcement learning reaches the highest reward and stabilizes earlier because agents learn system-level coordination rather than repeated local corrections.

A simple correlation analysis was also conducted on the simulated episodes. Physical-state volatility, measured as a composite of congestion variance, shelf-load change, and dock queue variability, has a positive association with late-order rate under static rules ($r = 0.64$). The association weakens under contextual single-agent reinforcement learning ($r = 0.37$) and becomes lower under collaborative contextual reinforcement learning ($r = 0.22$). This pattern suggests that contextual policies reduce the sensitivity of operational performance to physical-state volatility. In practical terms, the system becomes less fragile when disruptions occur.

Reward sensitivity was evaluated by varying the relative weight assigned to safety and efficiency. When the efficiency weight is raised too aggressively, average fulfillment time improves slightly but near-miss events rise. When the safety weight is too high, near-miss events fall but late orders increase. The balanced configuration used in Table IV provides the most stable trade-off. This finding is important for deployment because it shows that the reward function is not a neutral technical detail. It encodes managerial priorities, risk tolerance, and ethical commitments. A poorly designed reward function may produce technically successful but operationally undesirable behavior.

The results also reveal the value of collaboration. Contextual single-agent learning reduces fulfillment time and damage, but it still allows queue transfer between zones. For example, storage robots may move orders quickly to the sorting area before sorting arms are ready. Collaborative contextual learning performs better because the reward penalizes downstream overload and rewards synchronization. This confirms the central argument of the paper: embodied supply chain intelligence must reason from physical perception to collaborative execution, not from isolated local optimization to fragmented action.

Long short-term memory networks provide a foundation for learning temporal dependencies in congestion, equipment drift, and workload sequences (Hochreiter and Schmidhuber, 1997). Deep learning research explains why multi-layer representations can transform raw sensor and operational data into decision-ready features (LeCun et al., 2015). Attention mechanisms are useful for contextual reasoning because they allow a policy to weigh different state channels under changing operational conditions (Vaswani et al., 2017). Residual networks support robust visual feature extraction in settings where physical objects, shelves, and packaging conditions vary across time (He et al., 2016).

Graph convolutional learning provides a natural modeling basis for supply chains because nodes, links, and flow dependencies form structured networks (Kipf and Welling, 2017). Graph neural network surveys show that message passing is relevant when agent decisions depend on neighboring nodes and shared infrastructure (Wu et al., 2020). Model-explanation methods clarify how learned policies can be inspected when operational users need to understand the drivers of a recommendation (Lundberg and Lee, 2017). Local explanation methods also support trust by showing which observed features influenced a specific decision instance (Ribeiro et al., 2016).

Table IV. Comparative performance of policy configurations under the illustrative scenario.

Policy Configuration	Fulfillment Time (min)	Late Orders (%)	Damage Rate / 1000 Orders	Interpretation
Static rule dispatching	38.7	18.2	6.3	Baseline policy; reacts to deadlines but lacks physical context and collaboration
Digital-only learning policy	33.4	14.7	5.1	Learns demand and order patterns but remains weak under congestion and equipment variation
Contextual single-agent RL	29.1	10.8	3.4	Uses physical state perception to improve local decisions and reduce execution errors
Collaborative contextual RL	26.8	8.9	3.0	Coordinates storage, sorting,

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

				and dispatch through shared context and system-level reward

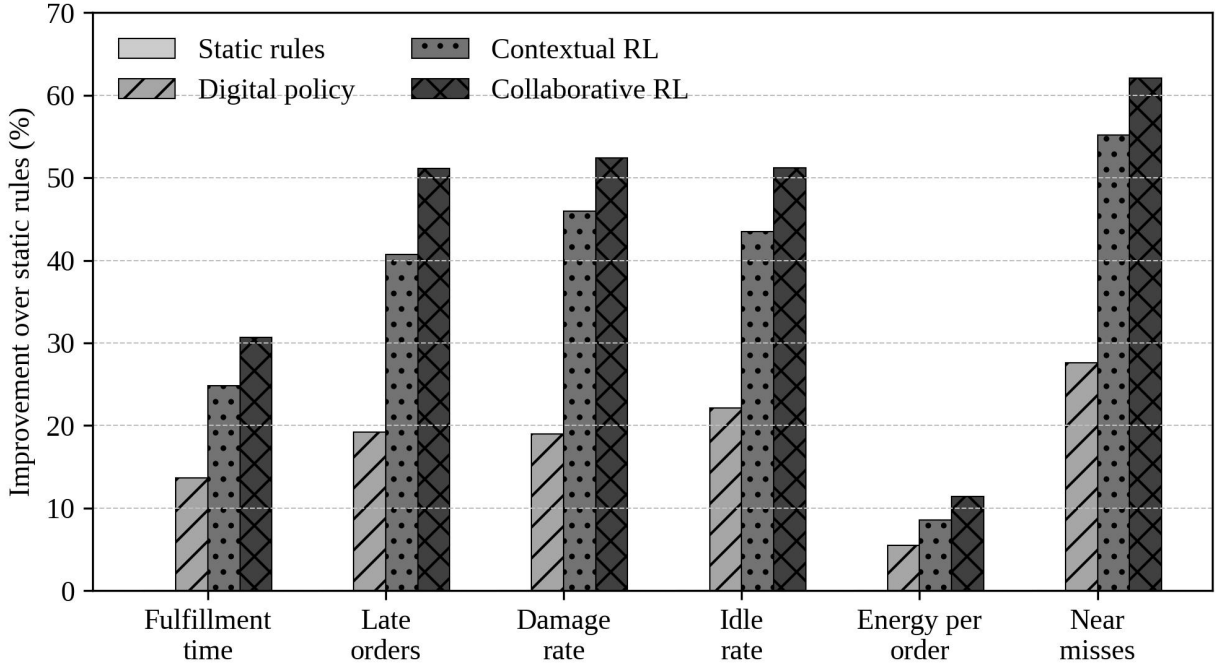


Figure 3. Policy performance improvements over static rule dispatching.

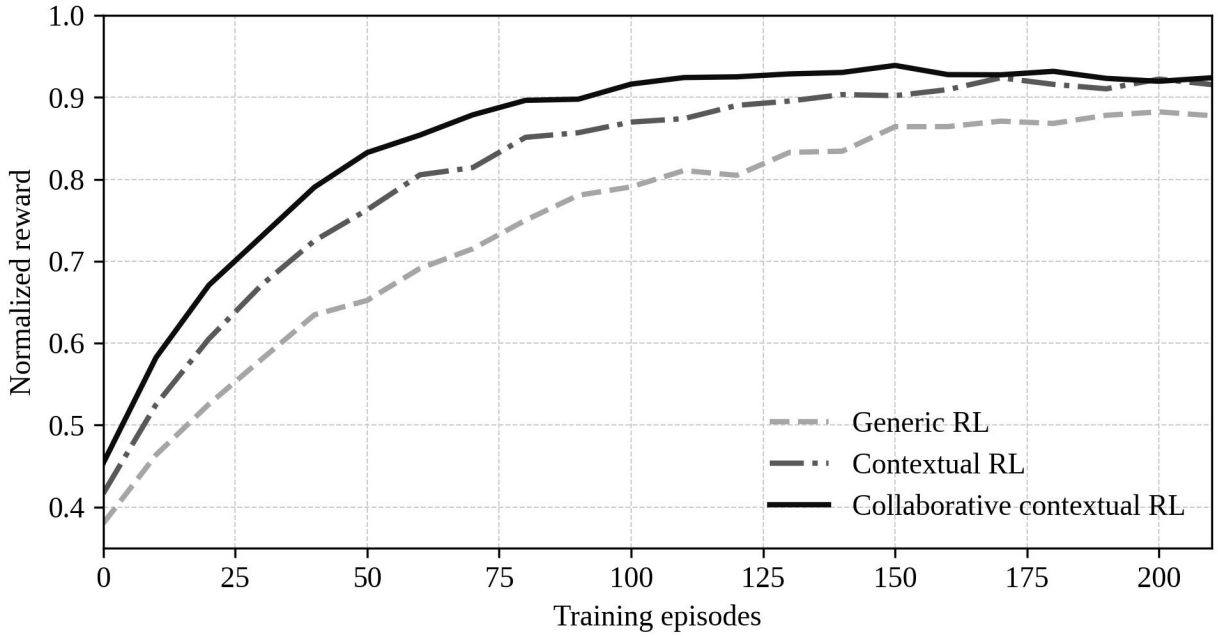


Figure 4. Training convergence under generic, contextual, and collaborative reinforcement learning.

VII. DISCUSSION: FROM POLICY LEARNING TO SUPPLY CHAIN ADAPTABILITY

The findings have several theoretical implications. First, they reposition reinforcement learning in supply chains from an optimization technique to a mechanism of physical-digital coupling. The policy is not only selecting a better action; it is learning how physical context changes the meaning of action. Path adjustment is not simply a route choice. It is a response to

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

aisle congestion, robot posture, order urgency, and downstream capacity. A priority adjustment is not only a scheduling decision. It is a negotiation among time windows, resource constraints, and collaborative effects.

Second, the framework clarifies the role of perception in AI analytics. Many analytics models treat data as already given. Embodied reasoning asks how the data are sensed, cleaned, synchronized, and trusted. If tactile feedback is unreliable, fragile-product handling decisions become risky. If dock congestion is observed too late, dispatch policies become reactive. If battery status is omitted, the policy may assign infeasible routes. The quality of contextual reasoning is therefore bounded by the quality of embodied perception.

Third, the paper extends the idea of adaptive collaboration. Collaboration is often defined as information sharing among supply chain partners. In an embodied agent system, collaboration is also behavioral. Agents coordinate by adjusting movement, timing, speed, load, and task allocation. Such coordination cannot be captured by purchase orders or inventory messages alone. It requires shared context, policy alignment, and execution feedback. The multi-agent interpretation developed here therefore deepens the understanding of collaboration from data exchange to embodied co-action.

The framework also has managerial implications. Managers should not deploy reinforcement learning as a black-box add-on to warehouse or transport systems. The first step is to define what the policy is allowed to observe, what actions it can take, what safety constraints cannot be violated, and what reward trade-offs reflect organizational goals. The second step is to build a simulation or digital twin environment where the policy can be trained under disruption scenarios. The third step is to deploy the policy gradually, beginning with advisory recommendations, then moving to supervised execution, and finally to bound autonomy for low-risk tasks.

An important practical implication concerns measurement. Traditional logistics metrics, such as average fulfillment time or transportation cost, are insufficient to evaluate embodied policies. A policy may reduce time while increasing damage, instability, or human override. The evaluation system should therefore include at least five classes of indicators: efficiency, accuracy, resource utilization, safety, and adaptability. Adaptability can be measured by recovery time after disruption, sensitivity of performance to state volatility, and number of episodes required for policy re-stabilization after a scenario change.

The framework also encourages more responsible AI governance. Because reinforcement learning policies optimize what they are rewarded for, reward design should be reviewed by operations managers, safety engineers, and frontline users. Human override data should not be treated as an inconvenience. It is a valuable signal that the policy may be misaligned with tacit operational knowledge. In addition, explainability tools should show which context variables influenced the recommended action. This is particularly important in human-machine collaborative spaces where trust affects adoption (Lundberg and Lee, 2017).

AI safety research warns that apparently efficient policies may produce harmful side effects when objectives are underspecified (Amodei et al., 2016). Constrained policy optimization formalizes the need to improve rewards while respecting safety or feasibility limits (Achiam et al., 2017). Autonomous-driving reinforcement learning surveys show that real-time physical autonomy depends on perception quality, environment complexity, and risk-aware control (Kiran et al., 2021). Machine health monitoring research shows that deep learning can convert machine signals into diagnostic knowledge for maintenance-aware execution policies (Zhao et al., 2019).

Predictive-maintenance reviews show that machine learning value depends on failure definitions, data quality, and evaluation design rather than model complexity alone (Carvalho et al., 2019). Automotive predictive-maintenance work highlights the deployment challenges that arise when model predictions must be converted into service actions (Theissler et al., 2021). Cyber-physical manufacturing architecture research provides a useful analogy for supply chain agents that integrate sensors, analytics, and actuation (Lee et al., 2015). Cyber-physical systems research also emphasizes that manufacturing intelligence is distributed across machines, control systems, humans, and data infrastructures (Monostori et al., 2016).

Intelligent manufacturing reviews show that smart resources sense, communicate, and adapt inside networked production environments (Zhong et al., 2017). Machine-learning research in manufacturing demonstrates that performance gains require careful matching among data structure, model choice, and engineering context (Wuest et al., 2016). Digital twin-driven smart manufacturing research directly supports the argument that virtual models should be coupled to physical execution and feedback (Lu et al., 2020). Recent quantum-industrial information integration research also indicates that future operational systems may require new computing paradigms for high-dimensional optimization (Lu et al., 2023).

VIII. IMPLEMENTATION PATH AND GOVERNANCE CONSIDERATIONS

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

A realistic implementation path begins with instrumentation. Sensors must provide reliable signals for the state variables that matter. This does not mean that every object needs to be fully instrumented. It means that the policy's decision variables should be observable with sufficient precision and latency. For example, shelf load, aisle occupancy, robot battery, sorting arm load, and dock queue length may be more valuable than a visually rich but decision-irrelevant 3D representation. Instrumentation should therefore be guided by policy requirements rather than by technology enthusiasm.

The second stage is data alignment. Multimodal perception data have different sampling rates, noise levels, and semantics. Visual data may arrive as frames, tactile data as force readings, warehouse events as discrete messages, and transport updates as GPS trajectories. A contextual policy needs synchronized state vectors. Edge preprocessing, timestamp alignment, anomaly filtering, and uncertainty scoring are therefore core components of the reasoning architecture. Without these components, reinforcement learning may learn from inconsistent or misleading state signals.

The third stage is simulation and digital-twin validation. Direct exploration in a live warehouse or delivery network is risky. The policy should first be trained against historical scenarios and synthetic disruptions. It should then be evaluated in a digital twin or sandbox environment that approximates physical constraints. Scenario stress testing should include demand spikes, equipment failures, communication delays, road disruption, and human intervention. Only after the policy satisfies safety, stability, and interpretability criteria should it be allowed to recommend actions in live operations.

The fourth stage is progressive autonomy. In early deployment, the policy should operate as a decision-support tool. Operators receive recommendations and explanations, and their acceptance or rejection becomes feedback. Once the policy demonstrates stable performance, it can execute low-risk actions automatically, such as queue re-ranking or minor path revision. High-risk actions, such as changing handling rules for fragile goods or rerouting urgent medical products, should remain under human review. This staged approach reduces implementation resistance and prevents premature automation.

Governance also requires monitoring model drift. Supply chain environments change as product mix, facility layout, robot condition, labor practices, and demand patterns change. A policy that performs well in one quarter may deteriorate later. Drift monitoring should track not only prediction accuracy but also reward distribution, override frequency, safety alerts, and state coverage. When the policy encounters states outside its training distribution, it should reduce autonomy and request human confirmation. This behavior is essential for safe deployment.

Cybersecurity and privacy must be considered because embodied policies rely on integrated data streams. Sensor tampering, false location signals, or unauthorized access to dispatch rules can produce physical disruption. The architecture should include authentication, role-based access, encrypted communication, and audit logs. When cross-organization data sharing is required, only compact context messages should be shared unless broader disclosure is justified. The goal is to provide enough information for collaboration without exposing unnecessary operational detail.

Digital twin enabling-technology research identifies data integration, simulation, analytics, and connectivity as core building blocks of physical-digital systems (Fuller et al., 2020). Systematic reviews of digital twins emphasize that a twin should be characterized by synchronization, purpose, fidelity, and interaction capability rather than by visualization alone (Jones et al., 2020). Manufacturing case studies show that digital twins require a clear conceptual framework before they can become operational decision tools (Onaji et al., 2022). Complex-systems research on digital twins explains why physical feedback is essential for mitigating undesirable emergent behavior (Grieves and Vickers, 2017).

Categorical reviews of manufacturing digital twins distinguish digital models, digital shadows, and true twins, a distinction that is important for embodied supply chain agents (Kritzinger et al., 2018). Digital twin-driven product design and service research shows how big data can connect engineering models with lifecycle operations (Tao et al., 2018). Industrial digital twin research identifies state synchronization and industrial informatics as central to turning physical assets into decision-capable systems (Tao et al., 2019). Research on CPS-based production systems shows that digital twins can play roles in monitoring, prediction, optimization, and control (Negri et al., 2017).

Digital twin enabling tools include sensors, data fusion, simulation, communication protocols, and analytics models that align with the implementation path proposed in this paper (Qi et al., 2021). Surveys of digital twin definitions show that design implications must be made explicit when a twin is used for decision automation (Barricelli et al., 2019). Digital twin challenge analysis shows that value creation depends on interoperability, uncertainty management, and computational feasibility (Rasheed et al., 2020). Design-and-production engineering research shows that digital twins should be shaped by the decision context rather than by generic modeling ambition (Schleich et al., 2017).

The comparison between digital twins and big data clarifies that smart manufacturing requires both high-volume data and physically meaningful synchronization (Qi and Tao, 2018). Autonomy-oriented digital twin research shows that future

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

manufacturing systems require reciprocal links between virtual estimation and physical action (Rosen et al., 2015). The simulation aspect of digital twins is particularly relevant to reinforcement learning because policy exploration should occur first in safe virtual environments (Boschert and Rosen, 2016). Geometry-assurance research illustrates how real-time physical feedback can support individualized production control (Soderberg et al., 2017). Deep learning for smart manufacturing shows that perception, diagnosis, and optimization can be combined into operational intelligence pipelines (Wang et al., 2018).

IX. LIMITATIONS AND FUTURE RESEARCH

This study has limitations. The data analysis is illustrative, and simulation based. Although the scenario is designed to reflect plausible warehouse, sorting, and delivery operations, it does not replace field validation. Future studies should test the framework using real execution data from autonomous mobile robots, robotic arms, warehouse management systems, and transport management platforms. Such studies should report not only average performance but also variability, rare-event behavior, human override, and post-deployment drift.

The model also simplifies several operational complexities. The action space is constrained, the agent population is moderate, and the reward weights are fixed within each experiment. Real supply chains may involve larger networks, heterogeneous robot brands, multiple companies, unionized labor rules, regulatory constraints, and changing service contracts. Future work can extend the model to hierarchical reinforcement learning, where strategic policies set high-level objectives and local policies handle execution. This would better match the multi-level nature of supply chain management.

Another future direction is causal reinforcement learning. A policy may learn that certain states are associated with late orders without understanding the causal mechanism. Causal structure can improve transferability and reduce spurious learning. For example, if dock blockage causes late departure, the policy should address dock synchronization rather than merely increasing robot speed. Combining causal graphs with embodied state representation may therefore improve interpretability and robustness.

Foundation models also open new opportunities. Large language models and multimodal models could translate human instructions, maintenance logs, exception reports, and visual scenes into structured context messages. They could also assist operators by explaining why a policy recommends an action. However, such models must be grounded in verified operational data and constrained by safety rules. In embodied supply chains, fluent explanation without physical validity would be dangerous. Future research should therefore examine how language-based reasoning can be coupled with reinforcement learning and digital twins under strict verification.

Finally, future work should examine organizational adoption. Policy performance alone does not determine success. Operators must trust the system, managers must understand trade-offs, and IT teams must maintain the data pipeline. Research on human-AI collaboration, algorithmic accountability, and supply chain governance should therefore be integrated with technical reinforcement learning studies. Embodied intelligence is not only a computational challenge; it is an organizational transformation.

X. CONCLUSION

This paper developed a contextual reasoning framework for embodied supply chain agents. Building from the premise that supply chains are physical-digital systems, it argued that reinforcement learning policies should be grounded in physical state perception and evaluated through collaborative execution outcomes. The framework defined a multimodal state representation, a balanced reward architecture, a safe policy learning cycle, and a multi-agent collaboration logic. It then demonstrated the framework through an illustrative simulation of storage, sorting, and delivery operations.

The analysis shows that contextual and collaborative reinforcement learning can reduce fulfillment delays, late orders, damage, idle time, energy use, and near-miss events under the stated scenario assumptions. More importantly, the results show why the improvement occurs: the policy sees physical constraints, learns from execution feedback, and coordinates across agents. The contribution is therefore not limited to a performance comparison. It is a structured way to translate embodied intelligence into implementable AI analytics for supply chain operations.

The broader message is that the next stage of intelligent supply chain management should move beyond digital optimization alone. Intelligent agents must perceive physical state, reason contextually, execute safely, collaborate with other agents, and learn from feedback. Reinforcement learning provides one powerful method for this transition, but its value depends on state design, reward governance, safety constraints, and organizational adoption. A supply chain becomes adaptive not when it has more algorithms, but when its algorithms learn responsibly from the physical world in which

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 56–72
DOI: 10.63646/jaiaa.2026.040105

decisions are executed.

AUTHOR CONTRIBUTIONS

Table V. Author contributions.

Author	Contribution
Marta Ribeiro	Conceptualization, methodology, writing - original draft, visualization
Diogo Fernandes	Formal analysis, simulation design, reinforcement learning model specification
Helena Costa	Supervision, project administration, writing - review and editing, correspondence
Pedro Almeida	Validation, technical interpretation, figure design, data curation

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The article uses an illustrative simulation design. All variables, assumptions, and aggregated results required to understand the analysis are reported in the manuscript. No proprietary operational dataset is redistributed.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records.

ABOUT THE AUTHORS

Marta Ribeiro is affiliated with the Department of Industrial Engineering at the University of Minho, Portugal. Her research focuses on AI-enabled logistics analytics, industrial operations, and data-driven supply chain coordination.

Diogo Fernandes is affiliated with the Department of Informatics at the University of Évora, Portugal. His work examines reinforcement learning, intelligent agents, and applied AI systems for operational decision-making.

Helena Costa is affiliated with the School of Technology and Management at the Polytechnic Institute of Leiria, Portugal. Her research interests include supply chain analytics, digital operations, and human-centered AI governance.

Pedro Almeida is affiliated with the Department of Electromechanical Engineering at the University of Beira Interior, Portugal. His research explores robotics, cyber-physical systems, and engineering applications of intelligent control.

REFERENCES

Gunasekaran, A., & Ngai, E. W. T. (2004). Information systems in supply chain integration and management. European Journal of Operational Research, 159(2), 269-295. https://doi.org/10.1016/j.ejor.2003.08.016

Min, H. (2010). Artificial intelligence in supply chain management: Theory and applications. International Journal of Logistics Research and Applications, 13(1), 13-39. https://doi.org/10.1080/13675560902736537

Pournader, M., Ghaderi, H., Hassanzadegan, A., & Fahimnia, B. (2021). Artificial intelligence applications in supply chain management. International Journal of Production Economics, 241, 108250. https://doi.org/10.1016/j.ijpe.2021.108250

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215-234. https://doi.org/10.1007/s10796-021-10221-w

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70, 356-365. https://doi.org/10.1016/j.jbusres.2016.08.009

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70, 308-317. https://doi.org/10.1016/j.jbusres.2016.08.004

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868-1883. https://doi.org/10.1111/poms.12838

Lu, Y. (2019a). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365

Boute, R. N., Gijsbrechts, J., van Jaarsveld, W., & Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. European Journal of Operational Research, 298(2), 401-412. https://doi.org/10.1016/j.ejor.2021.07.016

Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis. Transportation Research Part E: Logistics and Transportation Review, 136, 101922. https://doi.org/10.1016/j.tre.2020.101922

Dolgui, A., Ivanov, D., & Sokolov, B. (2018). Ripple effect in the supply chain: An analysis and recent literature. International Journal of Production Research, 56(1-2), 414-430. https://doi.org/10.1080/00207543.2017.1387680

Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles toward survivability. International Journal of Production Research, 58(10), 2904-2915. https://doi.org/10.1080/00207543.2020.1750727

Queiroz, M. M., Telles, R., & Bonilla, S. H. (2020). Blockchain and supply chain management integration: A systematic review. Supply Chain Management: An International Journal, 25(2), 241-254. https://doi.org/10.1108/SCM-03-2018-0143

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715-1729. https://doi.org/10.1007/s10796-022-10248-7

Lu, Y. (2017a). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452-10473. https://doi.org/10.1109/JIOT.2021.3060508

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y. (2017b). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1-10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2019b). The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration, 15, 80-90. https://doi.org/10.1016/j.jii.2019.04.002

Lu, Y., Sigov, A. S., Ratkin, L., Ivanov, L. A., & Zuo, M. (2023). Quantum computing and industrial information integration: A review. Journal of Industrial Information Integration, 35, 100511. https://doi.org/10.1016/j.jii.2023.100511

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8, 279-292. https://doi.org/10.1007/BF00992698

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3, 9-44. https://doi.org/10.1023/A:1022633531479

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8, 229-256. https://doi.org/10.1007/BF00992696

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285. https://doi.org/10.1613/jair.301

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518, 529-533. https://doi.org/10.1038/nature14236

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 2094-2100. https://doi.org/10.1609/aaai.v30i1.10295

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. International Journal of Robotics Research, 32(11), 1238-1274. https://doi.org/10.1177/0278364913495721

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 34(6), 26-38. https://doi.org/10.1109/MSP.2017.2743240

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971. https://doi.org/10.48550/arXiv.1509.02971

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015). Trust region policy optimization. arXiv preprint arXiv:1502.05477. https://doi.org/10.48550/arXiv.1502.05477

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. Journal of Machine Learning Research, 17(39), 1-40. https://doi.org/10.48550/arXiv.1504.00702

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400. https://doi.org/10.48550/arXiv.1703.03400

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290. https://doi.org/10.48550/arXiv.1801.01290

Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. arXiv preprint arXiv:1802.09477. https://doi.org/10.48550/arXiv.1802.09477

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. arXiv preprint arXiv:1706.02275. https://doi.org/10.48550/arXiv.1706.02275

Zhang, K., Yang, Z., & Basar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Handbook of Reinforcement Learning and Control (pp. 321-384). Springer. https://doi.org/10.1007/978-3-030-60990-0_12

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. arXiv preprint arXiv:1803.11485. https://doi.org/10.48550/arXiv.1803.11485

Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. arXiv preprint arXiv:1705.08926. https://doi.org/10.48550/arXiv.1705.08926

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575, 350-354. https://doi.org/10.1038/s41586-019-1724-z

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529, 484-489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., et al. (2017). Mastering the game of Go without human knowledge. Nature, 550, 354-359. https://doi.org/10.1038/nature24270

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35-45. https://doi.org/10.1115/1.3662552

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics, 38(2), 325-339. https://doi.org/10.1214/aoms/1177698950

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24. https://doi.org/10.1109/TNNLS.2020.2978386

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. arXiv preprint arXiv:1705.10528. https://doi.org/10.48550/arXiv.1705.10528

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Perez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4909-4926. https://doi.org/10.1109/TITS.2021.3054625

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213-237. https://doi.org/10.1016/j.ymssp.2018.05.050

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. P., Basto, J. P., & Alcala, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

Theissler, A., Perez-Velazquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety, 215, 107864. https://doi.org/10.1016/j.ress.2021.107864

Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18-23. https://doi.org/10.1016/j.mfglet.2014.12.001

Monostori, L., Kadar, B., Bauernhansl, T., Kondoh, S., Kumara, S., Reinhart, G., Sauer, O., et al. (2016). Cyber-physical systems in manufacturing. CIRP Annals, 65(2), 621-641. https://doi.org/10.1016/j.cirp.2016.06.005

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of Industry 4.0: A review. Engineering, 3(5), 616-630. https://doi.org/10.1016/j.eng.2017.05.015

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. Production & Manufacturing Research, 4(1), 23-45. https://doi.org/10.1080/21693277.2016.1192517

Lu, Y., Liu, C., Wang, K. I. K., Huang, H., & Xu, X. (2020). Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robotics and Computer-Integrated Manufacturing, 61, 101837. https://doi.org/10.1016/j.rcim.2019.101837

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. IEEE Access, 8, 108952-108971. https://doi.org/10.1109/ACCESS.2020.2998358

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36-52. https://doi.org/10.1016/j.cirpj.2020.02.002

Onaji, I., Tiwari, D., Soulatiantork, P., Song, B., & Tiwari, A. (2022). Digital twin in manufacturing: Conceptual framework and case studies. International Journal of Computer Integrated Manufacturing, 35(8), 831-858. https://doi.org/10.1080/0951192X.2022.2027014

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F. J. Kahlen, S. Flumerfelt, & A. Alves (Eds.), Transdisciplinary Perspectives on Complex Systems (pp. 85-113). Springer. https://doi.org/10.1007/978-3-319-38756-7_4

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 51(11), 1016-1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. International Journal of Advanced Manufacturing Technology, 94, 3563-3576. https://doi.org/10.1007/s00170-017-0233-1

Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), 2405-2415. https://doi.org/10.1109/TII.2018.2873186

Negri, E., Fumagalli, L., & Macchi, M. (2017). A review of the roles of digital twin in CPS-based production systems. Procedia Manufacturing, 11, 939-948. https://doi.org/10.1016/j.promfg.2017.07.198

Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wang, L., & Nee, A. Y. C. (2021). Enabling technologies and tools for digital twin. Journal of Manufacturing Systems, 58, 3-21. https://doi.org/10.1016/j.jmsy.2019.10.001

Barricelli, B. R., Casiraghi, E., & Fogli, D. (2019). A survey on digital twin: Definitions, characteristics, applications, and design implications. IEEE Access, 7, 167653-167671. https://doi.org/10.1109/ACCESS.2019.2953499

Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers. IEEE Access, 8, 21980-22012. https://doi.org/10.1109/ACCESS.2020.2970143

Schleich, B., Anwer, N., Mathieu, L., & Wartzack, S. (2017). Shaping the digital twin for design and production engineering. CIRP Annals, 66(1), 141-144. https://doi.org/10.1016/j.cirp.2017.04.040

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and Industry 4.0: 360 degree comparison. IEEE Access, 6, 3585-3593. https://doi.org/10.1109/ACCESS.2018.2793265

Rosen, R., von Wichert, G., Lo, G., & Bettenhausen, K. D. (2015). About the importance of autonomy and digital twins for the future of manufacturing. IFAC-PapersOnLine, 48(3), 567-572. https://doi.org/10.1016/j.ifacol.2015.06.141

Boschert, S., & Rosen, R. (2016). Digital twin-the simulation aspect. In P. Hehenberger & D. Bradley (Eds.), Mechatronic Futures (pp. 59-74). Springer. https://doi.org/10.1007/978-3-319-32156-1_5

Soderberg, R., Warmefjord, K., Carlson, J. S., & Lindkvist, L. (2017). Toward a digital twin for real-time geometry assurance in individualized production. CIRP Annals, 66(1), 137-140. https://doi.org/10.1016/j.cirp.2017.04.038

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems, 48, 144-156. https://doi.org/10.1016/j.jmsy.2018.01.003