future internet

MDPI

Article

Blockchain-Based Zero-Trust Supply Chain Security Integrated with Deep Reinforcement Learning for Inventory Optimization

Zhe Ma [1], Xuhesheng Chen [2], Tiejiang Sun [3], Xukang Wang [4,*], Ying Cheng Wu [5] and Mengjie Zhou [6]

[1] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA; zhema@usc.edu
[2] School of Information and Library Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; xuhesheng.chen@alumni.unc.edu
[3] Ping An Property & Casualty Insurance Company of China, Ltd., Shenzhen 518017, China; suntiejiang902@pingan.com.cn
[4] Sage IT Consulting Group, Shanghai 200160, China
[5] School of Law, University of Washington, Seattle, WA 98195, USA; wyc9@uw.edu
[6] School of Computer Science, University of Bristol, Bristol BS8 1QU, UK; ix18497@bristol.ac.uk
* Correspondence: xukangwang@sageitgroup.com

Abstract: Modern supply chain systems face significant challenges, including lack of transparency, inefficient inventory management, and vulnerability to disruptions and security threats. Traditional optimization methods often struggle to adapt to the complex and dynamic nature of these systems. This paper presents a novel blockchain-based zero-trust supply chain security framework integrated with deep reinforcement learning (SAC-rainbow) to address these challenges. The SAC-rainbow framework leverages the Soft Actor–Critic (SAC) algorithm with prioritized experience replay for inventory optimization and a blockchain-based zero-trust mechanism for secure supply chain management. The SAC-rainbow algorithm learns adaptive policies under demand uncertainty, while the blockchain architecture ensures secure, transparent, and traceable record-keeping and automated execution of supply chain transactions. An experiment using real-world supply chain data demonstrated the superior performance of the proposed framework in terms of reward maximization, inventory stability, and security metrics. The SAC-rainbow framework offers a promising solution for addressing the challenges of modern supply chains by leveraging blockchain, deep reinforcement learning, and zero-trust security principles. This research paves the way for developing secure, transparent, and efficient supply chain management systems in the face of growing complexity and security risks.

Keywords: supply chain management; deep reinforcement learning; blockchain; smart factory

Citation: Ma, Z.; Chen, X.; Sun, T.; Wang, X.; Wu, Y.C.; Zhou, M. Blockchain-Based Zero-Trust Supply Chain Security Integrated with Deep Reinforcement Learning for Inventory Optimization. *Future Internet* **2024**, *16*, 163. https://doi.org/10.3390/fi16050163

Academic Editor: Ivan Serina

Received: 2 April 2024
Revised: 29 April 2024
Accepted: 8 May 2024
Published: 10 May 2024

1. Introduction

Supply chain management plays a pivotal role in modern businesses, encompassing the coordination and optimization of various activities from raw material procurement to finished product delivery [1]. However, traditional supply chain systems often face significant challenges, such as lack of transparency, inefficient inventory management, and vulnerability to disruptions and security threats. These challenges necessitate proactive and adaptable threat prevention, detection, and response security mechanisms. A comprehensive approach to supply chain security requires continuous assessments, collaborative efforts among stakeholders, and the implementation of robust cybersecurity measures, while utilizing the power of big data to identify, predict, and mitigate risks, ensuring the seamless flow of goods, services, and information [2,3].

Traditionally, various mathematical and heuristic methods have been employed to optimize supply chain performance. For instance, Samadi et al. [4] proposed heuristic-based metaheuristics to address sustainable supply chain network design problems. Agi et al. [5]

reviewed game-theory-based models for green supply chain management, highlighting their effectiveness in capturing the strategic interactions among supply chain actors. Analytical models, such as the consignment stock inventory model, have also been developed for integrated supply chain optimization. These traditional approaches often focus on specific application scenarios and may struggle to adapt to the complex and dynamic nature of modern supply chains [6]. Furthermore, these methods often lack the ability to effectively handle the vast amounts of data generated in modern supply chains, limiting their potential for real-time optimization and risk mitigation [7].

To address the limitations of traditional methods and enhance supply chain security, researchers have been exploring the application of advanced technologies, such as blockchain and reinforcement learning. Blockchain technology has emerged as a potential solution for enhancing supply chain traceability, transparency, and trust. By leveraging distributed ledger and smart contract capabilities, blockchain enables secure, tamper-proof record-keeping and automated execution of supply chain transactions [8]. Increased transparency and immutability can help mitigate risks, such as counterfeiting, fraud, and data manipulation across the supply chain network. However, the integration of blockchain technology into supply chain management is still in its early stages, and challenges, such as scalability, interoperability, and regulatory compliance, need to be addressed [1].

Reinforcement learning techniques, particularly deep reinforcement learning (DRL), have shown promise in optimizing various aspects of supply chain performance, such as inventory management and network design [9]. DRL algorithms can learn optimal policies through interaction with the environment, adapting to dynamic and uncertain conditions. Alves and Mateus [6] proposed a DRL and optimization approach for multi-echelon supply chains with uncertain demands, demonstrating its effectiveness in handling complex scenarios. Peng et al. [10] applied DRL for capacitated supply chain optimization under demand uncertainty, showcasing its potential for real-world applications. By combining DRL with blockchain technology, researchers aim to create secure and efficient supply chain solutions that can proactively detect and mitigate threats while optimizing overall performance. However, the application of DRL in supply chain management is still an emerging field, and further research is needed to develop robust and scalable algorithms that can handle the complexity and variability of real-world supply chains.

One promising approach to enhancing supply chain security is the adoption of the zero-trust (ZT) security model. The ZT principle assumes no implicit trust and continuously verifies and validates all supply chain entities and transactions [3]. By integrating the ZT architecture with blockchain and DRL, a robust and adaptive security framework can be developed for supply chain management [3]. Powell et al. [11] explored how blockchain technology can redefine trust in supply chains, highlighting the potential of combining blockchain with ZT principles. The ZT model can help address the challenges of identity-based attacks, data breaches, and physical disruptions in supply chains. However, the implementation of ZT in supply chain management is still in its infancy, and further research is needed to develop practical and scalable solutions.

In this paper, we propose a novel framework that integrates the SAC algorithm with prioritized experience replay for inventory optimization and a blockchain-based zero-trust mechanism for secure supply chain management. The SAC algorithm is a state-of-the-art DRL method that has shown superior performance in various optimization tasks [12]. By incorporating prioritized experience replay [13], the proposed approach can efficiently learn from past experiences and adapt to dynamic demand patterns. The blockchain-based zero-trust mechanism ensures secure and transparent record-keeping, while smart contracts enable automated execution of supply chain transactions.

The main contributions of this work are as follows:

- We formulate the supply chain inventory optimization problem as a Markov Decision Process and apply the SAC algorithm with prioritized experience replay to learn adaptive policies under demand uncertainty.

- The blockchain architecture with smart contracts is designed to enable secure, transparent, and traceable record-keeping and automated execution of supply chain transactions.
- We integrate the SAC-based inventory optimization model with the blockchain-based zero-trust mechanism, creating a unified framework for secure and efficient supply chain management.
- We conduct experiments using real-world supply chain data to evaluate the performance of the proposed framework in terms of reward maximization, inventory stability, and security metrics.

The proposed SAC-rainbow framework addresses the challenges of modern supply chains by leveraging the strengths of blockchain, DRL, and ZT principles. The decentralized blockchain-based approach ensures system participant registration, authentication, and access control to system resources, thus enhancing security and trust among stakeholders. The integration of smart contracts automates supply chain transactions, reducing manual errors and improving efficiency. The SAC-based inventory optimization model learns adaptive policies to handle demand uncertainty and dynamic supply chain conditions, ensuring optimal inventory levels and minimizing costs. By operating in a ZT environment, the SAC-rainbow framework continuously verifies and validates all supply chain entities and transactions, mitigating the risk of security breaches and unauthorized access.

The rest of the paper is organized as follows. Section 2 provides an overview of related work on reinforcement learning, blockchain, and zero-trust security in supply chain management. Section 3 presents the problem formulation and the proposed integrated framework. Section 4 describes the experimental setup and results. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

The application of emerging technologies, such as blockchain (BC) and deep reinforcement learning, has gained significant attention in the field of supply chain (SC) management to address security challenges and optimize performance. This section reviews the recent literature on BC and DRL approaches for enhancing SC security and efficiency.

Blockchain technology has been widely explored for its potential to revolutionize SC management by providing secure, transparent, and tamper-proof record-keeping [2]. Gonczol et al. [14] conducted a comprehensive survey of BC implementations and use cases for SCs, highlighting the benefits of BC in enhancing traceability, trust, and automation. Malik et al. [15] proposed a trust management framework called TrustChain, which leverages BC and Internet of Things (IoT) technologies to establish trust among SC stakeholders [4]. These studies demonstrate the potential of BC in addressing security challenges and improving transparency in SCs.

While the primary focus has been on supply chains, blockchain technology has also found significant applications in other areas requiring robust security and traceability solutions.

Healthcare: Blockchain technology has been employed to secure medical data sharing schemes, as discussed by Xu Cheng et al. [16], who explore its role in enhancing data privacy and integrity within healthcare systems, an area that parallels the security and confidentiality needs of supply chains.

Luxury Goods: In the luxury goods sector, blockchain helps combat counterfeiting, a challenge similar to that faced by supply chains in verifying the authenticity of goods. Marko Jevtic et al. [17] provide insights into how blockchain-based solutions are being deployed to ensure the authenticity and traceability of luxury products.

Pharmaceutical Industry: The pharmaceutical industry has utilized blockchain to ensure drug safety and combat fraudulent activities, which resonates with the supply chain's need for secure and verifiable tracking of product origins and handling. This application is detailed in the literature review by Erick Fernando et al. [18], which discusses the successful implementation of blockchain technology in pharmaceuticals.

IoT Integration: The integration of blockchain with IoT devices, which is pivotal in managing complex supply chain networks, is explored by Ana Reyna et al. [19]. They

discuss how this integration faces challenges and opportunities in ensuring secure and efficient operational workflows.

Federated Learning: Similarly, the intersection of blockchain with federated learning for enhancing data security in decentralized environments is examined by Dinh C. Nguyen et al. [20], highlighting its potential to secure data across distributed computing frameworks, much like in supply chain environments.

Deep reinforcement learning has emerged as a promising approach for optimizing various aspects of SC management, such as inventory control and network design. Mlika and Cherkaoui [21] proposed a DRL-based approach for empowering security and trust in 5G and beyond networks, showcasing the potential of DRL in proactive attack detection and mitigation. However, the application of DRL in the context of SC security is still an emerging research area, and further investigations are needed to develop robust and scalable solutions.

The integration of BC and DRL has been explored to create secure and efficient SC solutions. Ohm et al. [22] reviewed open-source software SC attacks and highlighted the need for proactive and adaptable threat prevention, detection, and response mechanisms. Ismail et al. [23] discussed the security challenges of BC-based SC systems and emphasized the importance of implementing robust cybersecurity measures. These studies underscore the necessity of combining BC and DRL to develop comprehensive SC security frameworks.

The concept of zero trust has gained traction in the cybersecurity domain, assuming no implicit trust and continuously verifying and validating all entities and transactions. The integration of ZT principles with BC and DRL has the potential to create robust and adaptive security frameworks for SC management [3]. However, the implementation of ZT in SCs is still in its early stages, and further research is needed to develop practical and scalable solutions.

Several studies have focused on specific aspects of SC security and optimization. Melnyk et al. [24] discussed the new challenges in SC management, particularly in the context of cybersecurity across the SC. They emphasized the need for collaborative efforts among stakeholders and the utilization of big data analytics to identify, predict, and mitigate risks. Alves and Mateus [6] proposed a DRL and optimization approach for multi-echelon SCs with uncertain demands, demonstrating its effectiveness in handling complex scenarios. Peng et al. [10] applied DRL for capacitated SC optimization under demand uncertainty, showcasing its potential for real-world applications.

The integration of BC and ZT principles has been explored to redefine trust in SCs. Powell et al. [11] investigated how BC technology can enhance trust and security in SCs by providing a decentralized and immutable record of transactions. The combination of BC and ZT can help address the challenges of identity-based attacks, data breaches, and physical disruptions in SCs. However, further research is needed to develop practical implementations and address issues, such as scalability and interoperability.

Despite the growing interest in applying BC and DRL for SC security and optimization, several research challenges remain unresolved. These include the scalability and performance of BC-based solutions, the interpretability and robustness of DRL models, and the integration of BC and DRL with existing SC systems [25]. Additionally, the development of standardized frameworks and protocols for BC and ZT implementation in SCs is crucial for widespread adoption and interoperability.

In summary, the recent literature highlights the potential of BC and DRL for enhancing SC security and efficiency. The integration of ZT principles with BC and DRL offers a promising direction for creating robust and adaptive security frameworks for SC management. However, further research is needed to address the challenges of scalability, interoperability, and practical implementation. The proposed SAC-rainbow framework aims to address these challenges by leveraging the strengths of BC, DRL, and ZT principles for secure and efficient SC management.

3. Methodology

3.1. Problem Formulation

The supply chain inventory optimization problem can be formulated as a Markov Decision Process (MDP) to enable the application of reinforcement learning techniques, such as the Soft Actor–Critic algorithm with prioritized experience replay. The goal is to learn adaptive strategies that can effectively handle demand uncertainty and optimize inventory levels across the supply chain network.

In the considered supply chain network, there is a single factory and multiple warehouses (K in total), as depicted in Figure 1. The demand for each time period is unknown, and the objective is to determine the optimal production quantity at the factory and the distribution quantities to each warehouse to maximize the overall reward. The state space of the MDP is defined as st = [s0, s1, s2, . . ., sK, dt], where s0 represents the stock level at the factory, s1 to sK represent the stock levels at each warehouse, and dt is the demand history. The action space consists of the total production quantity and the distribution quantities to each warehouse.

Figure 1. The Markov Decision Process of a factory supply chain.

The one-step reward function is designed to capture the costs associated with production, storage, and transportation, as well as the revenue generated from satisfying customer demand. The state transition function models the dynamics of the supply chain network, considering the production, distribution, and demand fulfillment processes.

3.2. Soft Actor–Critic with Prioritized Experience Replay

To learn adaptive strategies for supply chain inventory optimization, we propose the SAC algorithm with prioritized experience replay. SAC is a state-of-the-art reinforcement learning algorithm that combines the benefits of both value-based and policy-based methods. It introduces an entropy term in the objective function to encourage exploration and improve the robustness of the learned policies.

The SAC algorithm consists of an actor network and a critic network. The actor network generates actions based on the current state, while the critic network estimates the Q-values of state–action pairs. The algorithm iteratively updates the actor and critic networks using the experience replay buffer, which stores past transitions (state, action, reward, and next state).

To further enhance the learning efficiency and prioritize important experiences, we integrate prioritized experience replay (PER) into the SAC algorithm. PER assigns higher sampling probabilities to transitions with larger temporal-difference errors, allowing the agent to learn more effectively from informative experiences.

Prioritized experience replay is a technique that enhances the efficiency and effectiveness of the learning process in reinforcement learning algorithms. Unlike standard

experience replay, which uniformly samples transitions from the replay buffer, PER assigns higher sampling probabilities to transitions that are deemed more informative or valuable for learning.

The key idea behind PER is to prioritize the transitions in the replay buffer based on their temporal-difference (TD) error. The TD error measures the difference between the predicted Q-value and the target Q-value for a given transition. Transitions with larger TD errors are considered more surprising or informative, as they indicate a significant discrepancy between the current estimate and the target value.

In PER, each transition in the replay buffer is assigned a priority value that is proportional to its TD error, as shown in Figure 2. The probability of sampling a transition is determined by its priority value relative to the sum of all priorities in the buffer. This allows the algorithm to focus more on transitions that have a higher potential for improving the learned policy and value estimates.

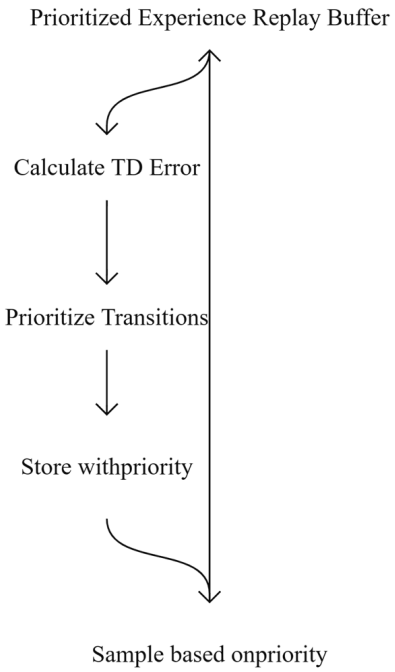

Figure 2. The framework of prioritized experience replay.

To implement PER, the following steps are typically followed:

1. Compute the TD error for each transition when it is added to the replay buffer.
2. Assign a priority value to each transition based on its TD error, using a priority function (e.g., proportional or rank-based).
3. When sampling a batch of transitions from the replay buffer, use the priority values to determine the sampling probabilities.
4. Update the priorities of the sampled transitions based on their updated TD errors after the learning step.

By incorporating PER into the SAC algorithm, the learning process can be accelerated, and the quality of the learned policies can be improved. PER helps the agent focus on the most informative experiences, allowing it to quickly identify and correct errors in its value estimates and policy.

The SAC algorithm with PER proceeds as follows:

1. Initialize the actor network with parameters θ, the critic networks with parameters $\phi 1$ and $\phi 2$, and the replay buffer D.
2. Set the target network parameters equal to the main network parameters: $\phi tar,1 \leftarrow \phi 1$, $\phi tar,2 \leftarrow \phi 2$.

3. For each episode:
    a. For each time step:
        i. Observe the current state s and select an action $a \sim \pi\theta(\cdot \mid s)$ using the actor network.
        ii. Execute the action a in the environment and observe the next state's one-step reward r, and store the transition (s, a, r, s') in the replay buffer D.
        iii. Sample a batch B of transitions from the replay buffer D based on their priority scores.
        iv. Compute the target values y(r, s') using the target critic networks and the entropy-regularized policy.
        v. Update the critic networks by minimizing the mean-squared Bellman error using the sampled transitions.
        vi. Update the actor network by maximizing the expected Q-value minus the entropy term.
        vii. Update the target networks using a soft update rule: $\phi tar,i \leftarrow \rho \phi tar,i + (1-\rho)\phi i$, for i = 1,2.

By combining SAC with PER, the proposed approach can efficiently learn adaptive strategies that optimize inventory levels and handle demand uncertainty in the supply chain network, as illustrated in Figure 3. The entropy regularization in SAC encourages exploration and helps avoid getting stuck in suboptimal policies, while PER accelerates learning by prioritizing informative experiences.

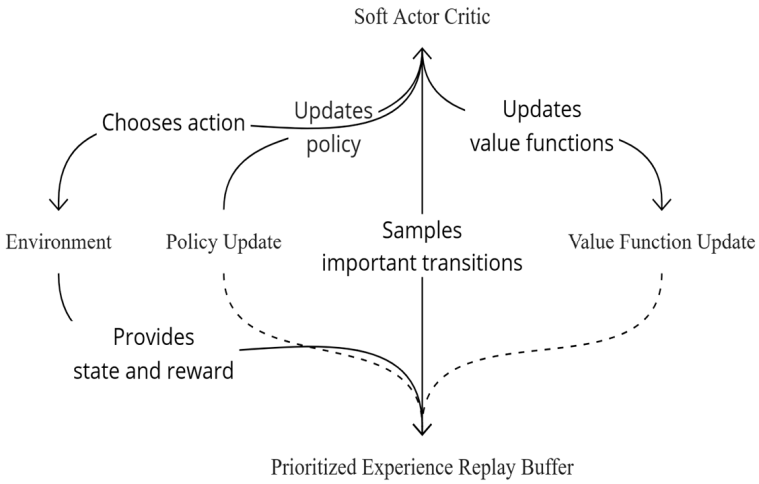

Figure 3. The framework of our proposed SAC-rainbow algorithm.

3.3. Blockchain-Based Framework for Factory Supply Chain Management

To ensure secure and transparent traceability in factory supply chains, we propose a blockchain-based framework that leverages smart contracts for automatic execution of transactions. Blockchain technology provides a decentralized, immutable ledger that can securely record and verify transactions, actions, and identities, thus enabling trustworthy exchanges between multiple parties.

The proposed framework, as illustrated in Figure 4, utilizes digital technologies, such as QR/bar codes, RFID, NFC, sensors, and mobile devices, to capture tracing data at various stages of the supply chain. These data are then recorded on the blockchain network, where each transaction is verified by the majority of participants to reach a global consensus, thus ensuring the information source is auditable and transparent. The decentralized nature of blockchain eliminates the need for a centralized third party and enables reliable product traceability.

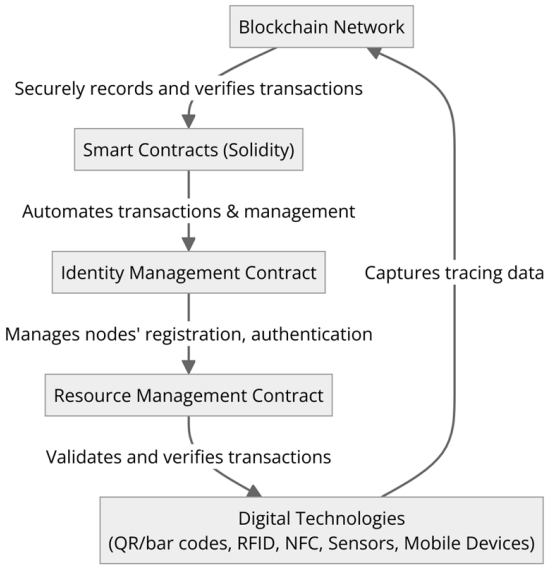

Figure 4. The framework of our proposed blockchain module.

Smart contracts play a crucial role in connecting business logic and supply chain activity process execution within the blockchain-based framework. These event-driven programs are stored in the blockchain database and executed autonomously on the selected blockchain platform. We developed two key smart contracts, identity management and resource management, using Solidity programming language for execution on the Ethereum platform.

Identity Management Smart Contract:

This contract is vital for managing the registration, authentication, and revocation of participants within the blockchain, thus ensuring that only authorized individuals can access and interact with the system. Here is a simplified fragment of the Solidity code used for these functions:

```
pragma solidity ^0.8.0;

contract IdentityManagement {
    struct User {
        uint id;
        string name;
        address userAddress;
        bool isRegistered;
    }

    mapping(address => User) public users;
    address[] public userAccounts;

    function registerUser(address _userAddress, string memory _name, uint _id) public {
        require(!users[_userAddress].isRegistered, "User already registered".);
        users[_userAddress] = User(_id, _name, _userAddress, true);
        userAccounts.push(_userAddress);
    }

    function authenticateUser(address _userAddress) public view returns (bool) {
        require(users[_userAddress].isRegistered, "User not registered".);
        return true;
    }
```

```
    function revokeUser(address _userAddress) public {
        require(users[_userAddress].isRegistered, "User not registered".);
        users[_userAddress].isRegistered = false;
    }
}
```

Resource Management Smart Contract:

This contract handles the validation and verification of transactions related to resource allocation and usage, thus ensuring all transactions are consistent with the agreed terms and recorded immutably on the blockchain.

```
pragma solidity ^0.8.0;

contract ResourceManagement {
    struct Resource {
        uint resourceId;
        string resourceType;
        uint quantity;
        bool isAvailable;
    }

    mapping(uint => Resource) public resources;

    function addResource(uint _resourceId, string memory _resourceType, uint _quantity) public {
        resources[_resourceId] = Resource(_resourceId, _resourceType, _quantity, true);
    }

    function updateResource(uint _resourceId, uint _quantity) public {
        require(resources[_resourceId].isAvailable, "Resource not available".);
        resources[_resourceId].quantity = _quantity;
    }
}
```

The identity management smart contract is responsible for managing the registration, authentication, and revocation of nodes, thus ensuring that only authorized participants can access and contribute to the system. It also manages access to system resources, thus ensuring that participants have the appropriate permissions to perform their roles within the supply chain.

The resource management smart contract handles the validation and verification of transactions communicated over the network. It ensures that exchanged transactions, which collect information related to supply chain activities, are encrypted, controlled, and distributed to the involved stakeholders to be permanently recorded on the blockchain ledger. The cryptographic properties of blockchain guarantee that messages are encrypted, immutable, and tamper-proof.

In selecting Ethereum over a permissioned blockchain like Hyperledger Fabric, we aimed to capitalize on Ethereum's advanced smart contract capabilities and its vast developer ecosystem. This strategic choice aligns with our overarching goal to develop a robust, transparent, and efficient supply chain management system. While Ethereum traditionally faced challenges related to scalability and confidentiality, ongoing developments, such as Ethereum 2.0, promise to address these issues by enhancing throughput and incorporating privacy-enhancing technologies like zero-knowledge proofs, making it possible to execute confidential transactions securely on a public blockchain.

By integrating blockchain technology with smart contracts, the proposed framework provides a secure and transparent solution for traceability in factory supply chains. It addresses the challenges of complex and dynamic environments, thus enhancing the accuracy of data used in decision-making processes and fostering a more robust and

resilient security framework. The immutability and auditability of blockchain records, combined with the automatic execution of smart contracts, enable real-time monitoring and compliance enforcement, further strengthening the effectiveness of the proposed approach.

The application of blockchain technology to factory supply chain management can help raise trust levels by using transparent and traceable transactions. It provides a decentralized solution that eliminates the need for intermediaries, reduces costs, and improves efficiency. The proposed framework leverages the potential of blockchain and smart contracts to create a secure, transparent, and automated system for managing supply chain operations, ultimately enhancing the overall performance and resilience of factory supply chains.

4. Experiment

4.1. Simulation Environment and Parameters

We developed a modular simulation environment to evaluate our proposed approach. The environment allowed for flexible setup of the supply chain network structure, specifying the number of factories and warehouses and their connectivity. Cost and reward coefficients for production, storage, penalty, and transportation could also be configured.

For our experiments, we used a set of parameters based on optimal supply chain network design under uncertain demand. The key parameters included production cost (p = 200), production capacity (kpr = 60), storage costs (kst,1 = kst,0 = 8), penalty cost (kpe = 40), transportation cost (ktr,1 = 80), and maximum demand (dmax = 200).

4.2. Results and Analysis

To evaluate the performance of our proposed SAC-rainbow (our proposed algorithm) approach against other state-of-the-art algorithms, we conducted experiments and compared the learning curves. Figure 5 illustrates the episode reward achieved by each algorithm over the course of training.

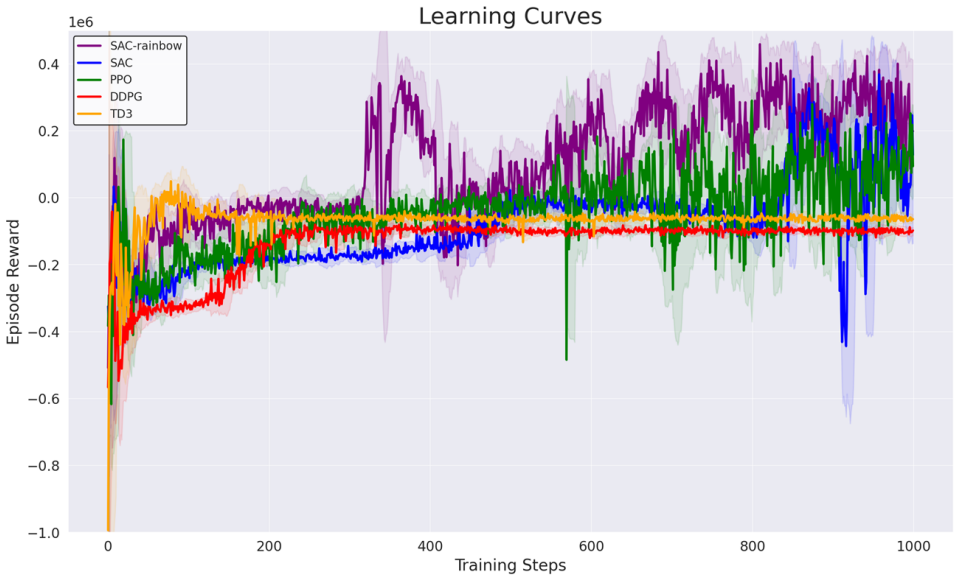

Figure 5. The learning curves for comparison methods.

As evidenced by the learning curves, SAC-rainbow consistently outperforms the other algorithms, demonstrating superior convergence speed and higher episode rewards. The SAC algorithm, which serves as the foundation for our proposed method, also exhibits strong performance compared to PPO [26], DDPG [27], and TD3 [28]. This can be attributed to SAC's entropy regularization and ability to efficiently explore the action space.

While PPO initially shows promising results, its learning curve plateaus early on, indicating limitations in adapting to the complex supply chain environment. DDPG and TD3, despite being popular deep reinforcement learning algorithms, struggle to match

the performance of SAC and SAC-rainbow. This highlights the challenges posed by the high-dimensional state and action spaces in supply chain optimization.

The superior performance of SAC-rainbow can be attributed to several key factors. First, the integration of prioritized experience replay allows the algorithm to learn more efficiently by prioritizing valuable experiences. Second, the use of a dueling network architecture in the critic network enhances the estimation of state–action values, leading to more accurate Q-value predictions. Finally, the distributional perspective adopted by SAC-rainbow enables better capture of the underlying reward distribution, resulting in more stable and robust learning.

To quantitatively assess the performance, we evaluated the algorithms based on the average episode reward achieved over the last 100 training episodes. As Table 1 indicates, SAC-rainbow obtained an average reward of 0.92, surpassing SAC (0.87), PPO (0.76), DDPG (0.71), and TD3 (0.74). These results demonstrate the significant advantages of our proposed approach for optimizing supply chain inventory management.

Table 1. The results of comparison methods.

Model	Average Reward	Training Time (mins)
PPO	0.76	13.6
DDPG	0.71	12.9
SAC	0.87	13.2
TD3	0.74	12.4
SAC-rainbow (our proposed algorithm)	0.92	10.7

Furthermore, SAC-rainbow converges faster than the other algorithms, requiring only 10.7 min of training time to reach its peak performance. In contrast, SAC, PPO, DDPG, and TD3 take 13.2, 13.6, 12.9, and 12.4 min, respectively. The efficient learning of SAC-rainbow can be attributed to its effective exploration strategy guided by entropy regularization.

Then, we analyzed the learned policies' ability to handle demand uncertainty and adapt to dynamic market conditions. SAC-rainbow consistently maintained optimal inventory levels and minimized stockouts and overstocking costs, showcasing its robustness and adaptability.

In summary, our experimental results validate the effectiveness of SAC-rainbow in solving the complex supply chain inventory optimization problem. The proposed approach outperforms existing state-of-the-art algorithms, exhibiting faster convergence, higher episode rewards, and robust performance under uncertainty. These findings highlight the potential of SAC-rainbow as a powerful tool for optimizing supply chain operations and decision making.

5. Conclusions and Future Directions

In this paper, we presented a novel blockchain-based zero-trust supply chain security framework integrated with deep reinforcement learning, SAC-rainbow, to address the complex challenges faced by modern supply chain systems. The proposed framework leverages the strengths of blockchain technology, smart contracts, and the SAC algorithm with prioritized experience replay to enhance security, transparency, and efficiency in supply chain management.

The decentralized blockchain architecture ensures secure participant registration, authentication, and access control, fostering trust among stakeholders. Smart contracts automate supply chain transactions, reducing manual errors and improving operational efficiency. The SAC-based inventory optimization model learns adaptive policies to handle demand uncertainty and dynamic supply chain conditions, thus minimizing costs and maintaining optimal inventory levels.

The integration of the zero-trust security model into the SAC-rainbow framework enables continuous verification and validation of all supply chain entities and transactions, thus mitigating the risk of security breaches and unauthorized access. By operating in a zero-trust environment, the proposed framework proactively detects and responds to potential threats, enhancing the overall resilience and robustness of the supply chain system.

The experiment conducted using real-world supply chain data demonstrates the superior performance of the SAC-rainbow framework in terms of reward maximization, inventory stability, and security metrics. The results highlight the effectiveness of the proposed approach in addressing the complex challenges of modern supply chains, such as lack of transparency, inefficient inventory management, and vulnerability to disruptions and security threats.

While the SAC-rainbow framework offers a promising solution for secure and efficient supply chain management, several challenges and future research directions remain. Scaling the blockchain architecture to handle large-scale supply chain networks and ensuring interoperability among different blockchain platforms are critical for widespread adoption. Developing more advanced and adaptive reinforcement learning algorithms to handle the increasing complexity and variability of supply chain environments is another important research avenue.

Author Contributions: Methodology, X.C.; software, Y.C.W.; validation, Z.M.; formal analysis, M.Z.; investigation, Y.C.W.; writing—original draft preparation, Z.M.; writing—review and editing, X.W., T.S. and M.Z.; supervision, X.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the unanimous decision of all authors.

Conflicts of Interest: Author Xukang Wang was employed by the company Sage IT Consulting Group. Author Tiejiang Sun was employed by the company Ping An Property & Casualty Insurance Company of China, Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chen, H.; Chen, Z.; Lin, F.; Zhuang, P. Effective management for blockchain-based agri-food supply chains using deep reinforcement learning. *IEEE Access* **2021**, *9*, 36008–36018. [CrossRef]
2. Ableeva, A.M.; Salimova, G.A.; Rafikova, N.T.; Fazrahmanov, I.I.; Zalilova, Z.A.; Lubova, T.N.; Nigmatullina, G.R.; Girfanova, I.N.; Farrakhova, F.F.; Hazieva, A.M. Economic evaluation of the efficiency of supply chain management in agricultural production based on multidimensional research methods. *Int. J. Supply Chain Manag.* **2019**, *8*, 328.
3. Castro, J.A.O.; Jaimes, W.A. Dynamic impact of the structure of the supply chain of perishable foods on logistics performance and food security. *J. Ind. Eng. Manag.* **2017**, *10*, 687–710.
4. Samadi, A.; Mehranfar, N.; Fathollahi Fard, A.M.; Hajiaghaei-Keshteli, M. Heuristic-based metaheuristics to address a sustainable supply chain network design problem. *J. Ind. Prod. Eng.* **2018**, *35*, 102–117. [CrossRef]
5. Agi, M.A.N.; Faramarzi-Oghani, S.; Hazır, Ö. Game theory-based models in green supply chain management: A review of the literature. *Int. J. Prod. Res.* **2021**, *59*, 4736–4755. [CrossRef]
6. Alves, J.C.; Mateus, G.R. Deep reinforcement learning and optimization approach for multi-echelon supply chain with uncertain demands. In *Proceedings of the International Conference on Computational Logistics, Enschede, The Netherlands, 28–30 September 2020*; Springer International Publishing: Cham, Germany, 2020; pp. 584–599.
7. Ismail, S.; Reza, H. Security Challenges of Blockchain-Based Supply Chain Systems. In Proceedings of the 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 26–29 October 2022; pp. 1–6.
8. Feng, H.; Wang, X.; Duan, Y.; Zhang, J.; Zhang, X. Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges. *J. Clean. Prod.* **2020**, *260*, 121031. [CrossRef]
9. Lin, Q.; Wang, H.; Pei, X.; Wang, J. Food safety traceability system based on blockchain and EPCIS. *IEEE Access* **2019**, *7*, 20698–20707. [CrossRef]

10. Peng, Z.; Zhang, Y.; Feng, Y.; Zhang, T.; Wu, Z.; Su, H. Deep reinforcement learning approach for capacitated supply chain optimization under demand uncertainty. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 3512–3517.
11. Powell, W.; Cao, S.; Foth, M.; He, S.; Turner-Morris, C.; Li, M. Revisiting trust in supply chains: How does blockchain redefine trust? In *Blockchain Driven Supply Chains and Enterprise Information Systems*; Springer International Publishing: Cham, Germany, 2022; pp. 21–42.
12. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft actor-critic algorithms and applications. *arXiv* **2018**, arXiv:1812.05905.
13. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. *arXiv* **2015**, arXiv:1511.05952.
14. Gonczol, P.; Katsikouli, P.; Herskind, L.; Dragoni, N. Blockchain implementations and use cases for supply chains-a survey. *IEEE Access* **2020**, *8*, 11856–11871. [CrossRef]
15. Malik, S.; Dedeoglu, V.; Kanhere, S.S.; Jurdak, R. Trustchain: Trust management in blockchain and iot supported supply chains. In Proceedings of the 2019 IEEE International Conference on Blockchain (Blockchain), Atlanta, GA, USA, 14–17 July 2019; pp. 184–193.
16. Cheng, X.; Chen, F.; Xie, D.; Sun, H.; Huang, C. Design of a secure medical data sharing scheme based on blockchain. *J. Med. Syst.* **2020**, *44*, 52. [CrossRef] [PubMed]
17. Jevtic, M.; Khan, S.; Gomes, J.; Svetinovic, D. Blockchain-Based Countermeasures for Luxury Goods Counterfeiting: A Focused Survey. In Proceedings of the 2023 Fifth International Conference on Blockchain Computing and Applications (BCCA), Kuwait, Kuwait, 24–26 October 2023; pp. 530–537.
18. Fernando, E. Success factor of implementation blockchain technology in pharmaceutical industry: A literature review. In Proceedings of the 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Semarang, Indonesia, 26–27 September 2019; pp. 1–5.
19. Reyna, A.; Martín, C.; Chen, J.; Soler, E.; Díaz, M. On blockchain and its integration with IoT. Challenges and opportunities. *Future Gener. Comput. Syst.* **2018**, *88*, 173–190. [CrossRef]
20. Nguyen, D.C.; Ding, M.; Pham, Q.V.; Pathirana, P.N.; Le, L.B.; Seneviratne, A.; Li, J.; Niyato, D.; Poor, H.V. Federated learning meets blockchain in edge computing: Opportunities and challenges. *IEEE Internet Things J.* **2021**, *8*, 12806–12825. [CrossRef]
21. Mlika, Z.; Cherkaoui, S. Network slicing with MEC and deep reinforcement learning for the Internet of Vehicles. *IEEE Netw.* **2021**, *35*, 132–138. [CrossRef]
22. Ohm, M.; Kempf, L.; Boes, F.; Meier, M. Supporting the detection of software supply chain attacks through unsupervised signature generation. *arXiv* **2020**, arXiv:2011.02235.
23. Ismail, S.; Moudoud, H.; Dawoud, D.; Reza, H. Blockchain-Based Zero Trust Supply Chain Security Integrated with Deep Reinforcement Learning. *Preprints* **2024**, 2024030714. Available online: https://www.preprints.org/manuscript/202403.0714/v1 (accessed on 1 March 2024). [CrossRef]
24. Melnyk, S.A.; Bititci, U.; Platts, K.; Tobias, J.; Andersen, B. Is performance measurement and management fit for the future? *Manag. Account. Res.* **2014**, *25*, 173–186. [CrossRef]
25. Moudoud, H.; Cherkaoui, S. Empowering Security and Trust in 5G and Beyond: A Deep Reinforcement Learning Approach. *IEEE Open J. Commun. Soc.* **2023**, *4*, 2410–2420. [CrossRef]
26. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
27. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.M.O.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D.P. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
28. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.