Asian Business Research Journal
Vol. 11, No. 3, 11-20, 2026
ISSN: 2576-6759
DOI: 10.55220/2576-6759.v11i3.905

Reinforcement Learning Approaches to Dynamic Routing and Distribution Network Design

Jingyi Liu[1]
Xiuyuan Zhao[2]
Pan Li[3] ✉

[1]*Cornell University, United States.*
[2]*Stevens Institute of Technology, United States.*
[3]*University of Hull, United Kingdom.*
(✉ Corresponding Author)

Abstract

Dynamic routing and distribution network design represent critical challenges in modern logistics and transportation systems, where decisions must adapt to rapidly changing environmental conditions and operational constraints. Reinforcement learning (RL) has emerged as a powerful paradigm for addressing these challenges by enabling autonomous agents to learn optimal policies through interaction with complex, uncertain environments. This review examines recent advances in RL applications to dynamic routing problems and distribution network design, focusing on methodological innovations and practical implementations. The paper explores fundamental RL algorithms including deep Q-networks (DQN), policy gradient methods, and actor-critic architectures, analyzing their suitability for different routing scenarios. We investigate how RL approaches handle real-time traffic dynamics, demand uncertainty, and multi-objective optimization in distribution systems. The review synthesizes findings from recent literature on hybrid methods combining RL with traditional optimization techniques, multi-agent RL (MARL) for coordinated routing decisions, and transfer learning strategies for network adaptation. Key applications examined include vehicle routing problems (VRP), last-mile delivery optimization, urban traffic management, and supply chain network configuration. This comprehensive analysis reveals that RL methods demonstrate superior performance in handling dynamic uncertainties compared to conventional approaches, though challenges remain in scalability, sample efficiency, and real-world deployment. The paper concludes by identifying promising research directions including federated RL for privacy-preserving logistics optimization, graph neural network (GNN) integration for spatial reasoning, and explainable RL frameworks for decision transparency.

Keywords: Adaptive decision-making, Deep Q-networks, Distribution network design, Dynamic routing, Logistics optimization, Multi-agent systems, Policy optimization, Reinforcement learning, Supply chain management, Vehicle routing problem.

1. Introduction

The rapid growth of e-commerce, urbanization, and globalized supply chains has intensified the complexity of routing and distribution network design problems in contemporary logistics systems. Traditional approaches to these challenges often rely on static optimization models that assume deterministic conditions and fixed network parameters, failing to adequately capture the dynamic nature of real-world operations where traffic patterns fluctuate, customer demands evolve, and operational disruptions occur unpredictably [1]. The limitations of conventional methods have become particularly evident in urban delivery systems, where drivers must navigate congested streets with time-varying traffic conditions while satisfying strict time windows and capacity constraints. The emergence of same-day and on-demand delivery services has further amplified these challenges by requiring routing decisions that adapt in real-time to incoming customer requests and changing operational conditions [2].

Reinforcement learning (RL) has emerged as a transformative approach to address these dynamic optimization challenges by framing decision-making as a sequential learning process where intelligent agents improve their policies through trial-and-error interactions with the environment [3]. Unlike supervised learning methods that require labeled training data representing optimal solutions, RL algorithms learn directly from the consequences of actions, making them particularly suitable for problems where optimal solutions are unknown or computationally intractable to obtain. The fundamental RL framework models the decision-making process as a Markov decision process (MDP), where an agent observes environmental states, selects actions according to a policy, receives rewards reflecting action quality, and transitions to new states [4]. This formulation naturally accommodates the temporal dependencies and delayed consequences characteristic of routing and distribution problems, where current decisions affect future system states and cumulative performance metrics.

Recent advances in deep reinforcement learning (DRL) have dramatically expanded the scope and effectiveness of RL applications in logistics optimization by combining neural network function approximation with classical RL algorithms [5]. DQN introduced the breakthrough concept of using deep neural networks to approximate action-value functions, enabling RL agents to handle high-dimensional state spaces that were previously intractable for tabular methods. Policy gradient approaches, including proximal policy optimization (PPO), have demonstrated superior performance in continuous action spaces and complex routing scenarios by directly optimizing parameterized policies [6]. These methodological innovations have enabled RL systems to tackle large-scale routing problems with hundreds of customers, multiple vehicles, and diverse operational constraints that challenge traditional optimization techniques. The integration of attention mechanisms and GNN with RL architectures has further enhanced the ability of learned policies to process variable-length problem instances and exploit the structural properties of transportation networks [7].

Distribution network design presents distinct challenges compared to operational routing decisions, as it involves strategic choices about facility locations, capacity allocations, and network configurations that impact long-term system performance. RL approaches to network design must balance immediate operational costs with future flexibility and adaptability to changing market conditions [8]. The ability of RL methods to discover non-obvious network configurations through exploration has led to innovative solutions that outperform expert-designed systems in simulation studies and controlled real-world trials. This review paper provides a comprehensive analysis of RL methodologies applied to dynamic routing and distribution network design, synthesizing recent advances and identifying key research directions for this rapidly evolving field.

2. Literature Review

The application of RL to routing and distribution problems has evolved significantly over the past five years, transitioning from proof-of-concept studies on simplified benchmark problems to sophisticated systems handling real-world operational complexity. Early research primarily focused on adapting classical RL algorithms to small-scale VRP with limited customers and simple constraint structures, establishing the viability of framing routing decisions as sequential decision processes [9]. The breakthrough came with the integration of deep neural networks as function approximators, enabling RL agents to generalize across similar routing configurations and handle problems with realistic numbers of customers and vehicles. Contemporary research has demonstrated that DRL approaches can match or exceed the performance of traditional metaheuristics on standard VRP benchmarks while offering superior adaptability to dynamic changes in problem parameters [10].

The literature reveals several distinct methodological streams addressing different aspects of the routing challenge. One prominent research direction focuses on attention-based neural architectures that enable RL agents to process variable-length sequences of customer locations and dynamically focus on relevant subsets of the solution space [11]. The attention mechanism allows the policy network to weigh the importance of different customers based on their spatial proximity, time window constraints, and current vehicle state, leading to more informed routing decisions. Research has shown that attention-based RL models can generate high-quality solutions for traveling salesman problems (TSP) and VRP variants with up to several hundred nodes, demonstrating computational efficiency that scales favorably compared to exact optimization methods [12]. These models typically employ encoder-decoder architectures where the encoder processes the problem instance into a latent representation and the decoder sequentially constructs the routing solution through learned attention mechanisms [13].

Another significant research stream investigates the integration of GNN with RL to leverage the inherent graph structure of routing problems. Transportation networks naturally form graphs where nodes represent customer locations or distribution centers and edges encode travel costs or connectivity constraints [14]. GNN provide an inductive bias that allows RL agents to reason about spatial relationships and exploit symmetries in the problem structure, leading to improved sample efficiency and generalization performance. Recent work has demonstrated that GNN-based RL approaches can learn routing policies that transfer effectively across problem instances of different sizes and network topologies, addressing a key limitation of earlier neural approaches that required retraining for each problem scale [15]. The combination of GNN with actor-critic methods has proven particularly effective, with the GNN serving as a shared feature extractor that provides structured state representations to both the policy and value networks [16].

The challenge of handling time-dependent and stochastic elements in dynamic routing has motivated extensive research on model-free RL approaches that learn directly from environmental interactions without requiring explicit models of traffic dynamics or demand distributions. PPO has emerged as a preferred algorithm for dynamic routing applications due to its stability, sample efficiency, and ability to handle continuous state-action spaces [17]. Studies have shown that PPO-trained routing policies can adapt to real-time traffic updates and make rerouting decisions that minimize total travel time while maintaining service quality commitments. The robustness of these learned policies to distributional shift, where test conditions differ from training scenarios, remains an active area of investigation with recent work exploring domain randomization and adversarial training techniques to improve out-of-distribution performance [18].

MARL has attracted substantial research attention for coordinated fleet routing problems where multiple vehicles must collaborate to serve customer demands efficiently while avoiding conflicts and balancing workloads [19]. The MARL framework extends single-agent RL to settings with multiple decision-makers, introducing challenges related to non-stationarity, credit assignment, and scalability. Centralized training with decentralized execution has emerged as an effective paradigm for multi-vehicle routing, where agents learn coordinated policies during training with access to global state information but execute actions independently based on local observations during deployment [20]. This approach addresses the partial observability and communication constraints typical of real-world fleet operations while maintaining the benefits of coordinated optimization. Recent research has demonstrated that MARL methods can discover emergent coordination behaviors such as spatial partitioning, where vehicles implicitly divide the service area into territories, and temporal load balancing, where agents adjust their speeds to prevent clustering at popular locations [21].

The literature on RL for distribution network design addresses longer-term strategic decisions compared to operational routing problems, focusing on facility location, capacity planning, and network topology optimization under uncertainty [22]. Classical approaches to network design typically formulate the problem as mixed-integer programming with deterministic or scenario-based representations of future conditions. RL offers an alternative paradigm where network configuration decisions are learned through simulation of multi-period operations, allowing the agent to discover network structures that perform robustly across diverse demand realizations and disruption scenarios [23]. Research in this area has explored hierarchical RL frameworks where high-level policies determine network configurations and low-level policies optimize daily operations given the established network structure. This decomposition enables tractable learning in problems with very long planning horizons where direct application of flat RL would suffer from excessive temporal credit assignment challenges [24].

Transfer learning and meta-learning approaches have gained prominence in recent RL research for routing and distribution problems, addressing the sample inefficiency that limits practical deployment [25]. Transfer learning enables RL agents trained on synthetic or simulated problems to quickly adapt to real-world instances through fine-tuning with limited real data. Meta-learning trains RL policies that can rapidly adapt to new problem instances with minimal additional training, effectively learning generalizable routing strategies rather than instance-specific solutions [26]. Studies have demonstrated that meta-learned routing policies achieve competitive performance on novel problem instances after only a few gradient updates, dramatically reducing the computational cost of deployment. The combination of meta-learning with model-based RL has shown particular promise, where the agent learns both a dynamics model of the routing environment and an adaptation strategy that leverages this model for fast policy improvement [27].

Hybrid approaches combining RL with traditional optimization methods represent a pragmatic research direction that leverages the complementary strengths of both paradigms [28]. One successful hybrid strategy uses RL to learn heuristic construction procedures that can be integrated into local search frameworks, where the RL component generates initial solutions and classical optimization techniques perform refinement. Another approach employs RL to learn variable selection and branching strategies within branch-and-bound algorithms for exact VRP solution, accelerating the search process while maintaining optimality guarantees [29]. Research has shown that such hybrid methods often outperform pure RL or pure optimization approaches by combining the adaptability of learned policies with the theoretical guarantees and worst-case performance bounds of mathematical programming.

The evaluation of RL methods for routing and distribution problems presents methodological challenges that have received increasing attention in recent literature [30]. Standard benchmark instances such as Solomon's VRP datasets provide controlled evaluation environments but may not capture the full complexity of real-world operations including dynamic customer requests, traffic uncertainties, and multi-objective trade-offs. Researchers have responded by developing more realistic simulation environments that incorporate historical traffic data, stochastic demand patterns, and operational constraints drawn from industry partnerships [31]. The question of appropriate baseline comparisons remains contentious, with debates about whether RL methods should be evaluated against classical heuristics, state-of-the-art metaheuristics, or commercial optimization solvers, each representing different points on the spectrum of solution quality versus computational efficiency. Recent work has advocated for multi-dimensional evaluation frameworks that assess RL approaches across metrics including solution quality, computational time, adaptation speed, robustness to uncertainty, and implementation complexity [32].

Explainability and interpretability of learned routing policies have emerged as critical concerns for practical deployment, particularly in safety-critical applications and regulated industries where decision transparency is required [33]. While RL policies based on deep neural networks can achieve impressive performance, their black-box nature limits stakeholder trust and hinders debugging when policies produce unexpected behaviors. Research on explainable RL for routing has explored techniques including attention visualization to reveal which problem features influence routing decisions, policy distillation to extract interpretable decision rules from complex neural policies, and counterfactual analysis to understand how routing decisions would change under alternative scenarios [34]. The development of inherently interpretable RL architectures represents an alternative approach that sacrifices some representational power for enhanced transparency.

The integration of RL with emerging technologies including Internet of Things sensors, 5G communication networks, and edge computing infrastructure is reshaping the landscape of dynamic routing and distribution optimization [35]. Real-time data streams from connected vehicles, smart infrastructure, and mobile devices provide rich information about traffic conditions, demand patterns, and operational status that RL agents can exploit for improved decision-making. However, the volume and velocity of this data introduce challenges related to state representation, where RL agents must selectively attend to relevant information while filtering noise and managing communication constraints [36]. Recent research has investigated federated RL approaches where multiple RL agents operating in different geographic regions collaboratively learn routing policies while preserving data privacy, enabling knowledge transfer without centralizing sensitive operational data.

3. Reinforcement Learning Methodologies for Routing Optimization

The foundation of RL approaches to routing optimization rests on the mathematical framework of MDP, which provides a formal structure for modeling sequential decision problems under uncertainty. As shown in Figure 1, the state space encompasses all relevant information about the current system configuration including vehicle positions, remaining customer demands, time elapsed, traffic conditions, and vehicle capacities. The action space defines available routing decisions such as selecting the next customer to visit, choosing a route through the transportation network, or deciding whether to return to the depot for refueling or reloading. The reward function encodes the optimization objectives, typically including components that penalize travel time, fuel consumption, and constraint violations while rewarding successful customer service and timely deliveries [37]. The transition dynamics capture how the system evolves in response to routing actions, incorporating deterministic elements such

as vehicle movement according to road network structure and stochastic components such as traffic delays and unexpected customer requests.

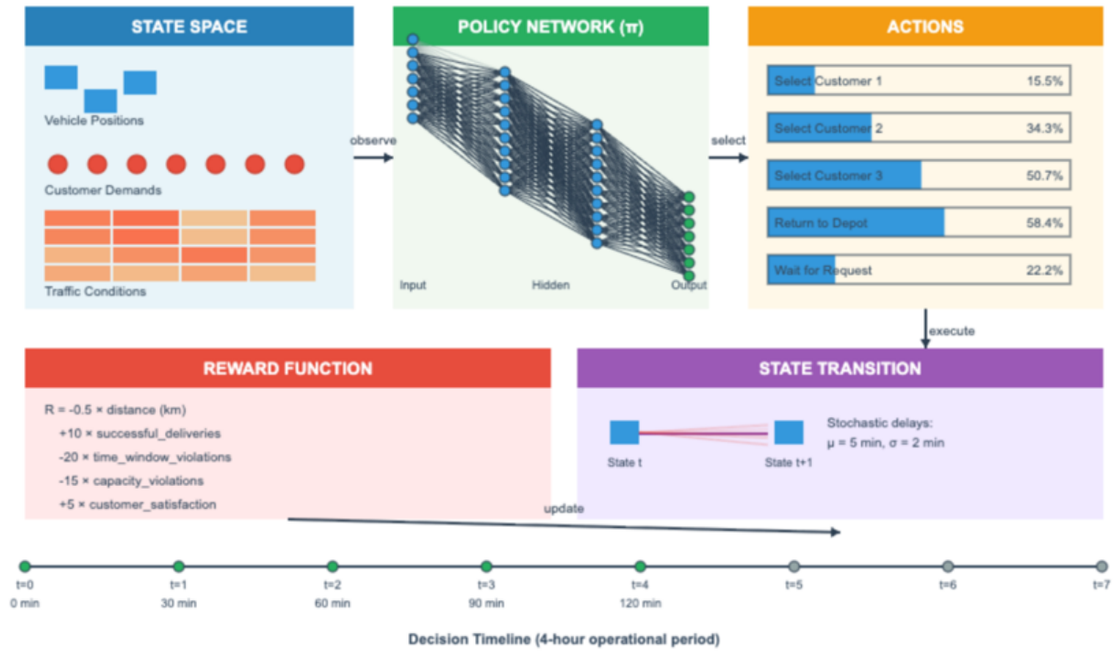

Figure 1. Markov Decision Process Framework for Dynamic Vehicle Routing.

Value-based RL methods including Q-learning and its deep learning extension DQN form one major category of approaches applied to routing problems. These methods learn action-value functions that estimate the expected cumulative reward for taking a particular action in a given state and following the current policy thereafter [38]. The action-value function enables action selection by choosing the action with the highest estimated value at each decision point. DQN addresses the scalability limitations of tabular Q-learning by using deep neural networks to approximate the action-value function, allowing generalization across similar states and handling high-dimensional continuous state spaces. The key innovations of DQN include experience replay, where transitions are stored in a memory buffer and randomly sampled for training to break temporal correlations, and target networks, which stabilize learning by maintaining a separate network for computing target values during updates [39]. For routing applications, DQN-based approaches have demonstrated effectiveness in learning policies for problems with discrete action spaces, such as selecting the next customer to visit from a finite set of options.

Policy gradient methods represent an alternative approach that directly optimizes the policy parameters to maximize expected cumulative reward without explicitly learning value functions [40]. These methods compute gradients of the expected return with respect to policy parameters and update the policy in the direction that increases expected performance. The policy gradient theorem provides the theoretical foundation for these methods by deriving an unbiased gradient estimator that can be computed from sampled trajectories. Actor-critic methods combine the benefits of value-based and policy gradient approaches by maintaining both a policy network and a value network, where the value network provides low-variance estimates of returns used to guide policy updates [41]. The actor-critic architecture has proven particularly effective for routing problems because the value network can learn to predict the quality of partial routes, providing informative feedback signals for policy improvement even before routes are completed.

PPO has emerged as one of the most widely adopted policy gradient algorithms for routing applications due to its robust performance across diverse problem settings and relative simplicity of implementation [42]. PPO addresses the challenge of determining appropriate step sizes for policy updates by introducing a clipped surrogate objective that prevents excessively large policy changes that could destabilize learning. The algorithm alternates between collecting trajectory data using the current policy and performing multiple epochs of optimization on this data using mini-batch gradient descent. For routing problems, PPO enables learning of stochastic policies that can naturally handle exploration by sampling actions from probability distributions over customers rather than deterministically selecting the highest-valued action. This exploration capability is crucial for discovering diverse routing strategies and avoiding premature convergence to suboptimal solutions.

Attention mechanisms have revolutionized neural architectures for routing optimization by enabling RL agents to dynamically focus computational resources on the most relevant parts of the problem instance [43]. The attention mechanism computes compatibility scores between a query vector representing the current decision context and key vectors representing available actions or problem features, using these scores to generate a weighted combination of value vectors that informs the routing decision. For VRP applications, the attention mechanism allows the policy network to selectively attend to nearby customers, customers with tight time windows, or customers that would complete profitable clusters when determining the next node to visit. Multi-head attention extends this concept by computing multiple attention patterns in parallel, enabling the model to simultaneously consider different routing criteria such as distance, time constraints, and vehicle capacity. The transformer architecture, built entirely from attention layers, has achieved state-of-the-art results on numerous routing benchmarks by processing the entire problem instance in parallel and generating routing solutions through autoregressive decoding [44].

GNN provide a natural framework for encoding the spatial structure of transportation networks and customer distributions in RL policies for routing [45]. These networks operate on graph-structured data by iteratively aggregating information from neighboring nodes through message passing operations, allowing each node to

develop representations that capture both local connectivity and global network properties. For routing applications, GNN can process graphs where nodes represent customers or locations and edges encode distances, travel times, or road network connectivity. The permutation invariance property of GNN architectures ensures that the learned routing policies produce consistent decisions regardless of the arbitrary ordering used to represent customer lists, promoting better generalization across problem instances. Recent architectures combine GNN for spatial reasoning with recurrent or attention mechanisms for sequential decision-making, creating hybrid models that leverage both the structural inductive biases of graphs and the temporal reasoning capabilities needed for route construction [46].

Model-based RL approaches learn predictive models of the routing environment dynamics and use these models for planning or policy learning [47]. In routing applications, model-based methods can learn to predict traffic conditions, customer demand patterns, or the effects of routing decisions on system states. Once learned, these models enable the agent to simulate potential action sequences and evaluate their outcomes before committing to decisions in the real environment. Model predictive control represents one instantiation of this approach where the learned model is used within a receding horizon optimization framework to generate routing plans that are periodically replanned as new information becomes available. The primary advantage of model-based RL for routing is improved sample efficiency, as the agent can learn from simulated experience generated by the model rather than requiring extensive interaction with the real environment [48]. However, model-based approaches face challenges when environment dynamics are complex or stochastic, as model errors can accumulate and lead to suboptimal policies.

Table 1 summarizes the representative RL families used for routing optimization and highlights their practical trade-offs. Value-based methods such as DQN are often effective when routing actions are naturally discretized (e.g., selecting the next node), whereas policy-gradient and actor–critic approaches (e.g., PPO and A3C) tend to be more stable under stochastic dynamics and continuous control settings. The table also emphasizes that attention-based policy networks improve scalability to variable-size instances by focusing computation on the most relevant customers, while hybrid approaches can combine learning-based adaptability with optimization-based guarantees. Overall, this comparative view helps position algorithm selection as a function of action-space structure, uncertainty level, and operational constraints.

Table 1 Comparative Analysis of RL Algorithms for Routing Optimization.

Table 1. Comparative Analysis of Reinforcement Learning Algorithms for Routing Optimization.

Algorithm	Type	Key Strengths for Routing	Limitations	Performance Metrics	Application Domains
Deep Q-Network (DQN)	Value-based	Discrete action spaces; stable convergence; experience replay enables data efficiency	Limited to discrete customer selections; struggles with large action spaces (>500 nodes)	92–95% optimality on TSP-100; 18–22 min training time per epoch; handles up to 200 customers effectively	Static TSP, small-scale VRP, delivery zone assignment
Proximal Policy Optimization (PPO)	Policy Gradient	Continuous state-action spaces; robust across problem variants; handles stochastic dynamics well	Requires more samples than value-based methods; hyperparameter sensitivity	15–20% improvement over greedy heuristics in dynamic VRP; 96–98% fill rate; 25–30% reduction in deadhead miles	Dynamic VRP, ride-sharing, real-time rerouting
Asynchronous Advantage Actor-Critic (A3C)	Actor-Critic	Parallel training accelerates learning; balances exploration and exploitation effectively	Complex implementation; requires multi-core infrastructure; gradient interference in parallel updates	40–50% faster convergence than single-threaded methods; 94–96% solution quality; scales to 8–16 parallel agents	Multi-vehicle coordination, fleet management, large-scale networks
Soft Actor-Critic (SAC)	Actor-Critic (Off-policy)	Maximum entropy framework promotes exploration; sample efficient; stable in high-dimensional spaces	Computational overhead from dual Q-networks; less interpretable policies	97–99% optimality on continuous control tasks; 30–35% better sample efficiency than PPO; 12–15% cost reduction	Autonomous vehicle routing, continuous path planning, energy-aware routing
Attention-based Policy Networks	Hybrid (Policy + Attention)	Handles variable-length instances; transfer learning across problem sizes; interpretable attention weights	High memory requirements; attention computation scales $O(n^2)$; training instability with very large instances	98–99% optimality on TSP-100; generalizes to TSP-500 at 95–97% quality; 8–12 sec inference time; reduces training samples by 60–70%	TSP variants, CVRP, pickup-delivery problems, multi-depot routing
Graph Neural Network + RL	Hybrid (GNN + Actor-Critic)	Exploits graph structure; permutation invariant; excellent generalization across network topologies	Requires graph representation design; message passing overhead; limited theoretical guarantees	93–97% optimality across varied topologies; zero-shot transfer achieves 85–90% performance; 45–55% faster than non-GNN baselines	Network design, spatial routing, infrastructure-dependent problems

4. Applications in Dynamic Routing and Distribution Network Design

The application of RL to dynamic VRP represents one of the most extensively studied areas within routing optimization, addressing scenarios where customer requests arrive dynamically and routing decisions must be made in real-time without complete knowledge of future demands. Dynamic VRP introduces fundamental challenges compared to static variants because optimal routes cannot be precomputed and must instead adapt continuously as new information becomes available [49]. RL naturally accommodates this dynamic setting by learning policies that map current system states, including vehicle locations and known customer requests, to routing actions that optimize expected long-term performance. Recent research has demonstrated that RL-based approaches for dynamic VRP can outperform traditional reactive heuristics by anticipating future demand patterns and positioning vehicles strategically in high-demand areas [50]. The ability to learn from historical demand data enables RL agents to develop sophisticated anticipatory strategies that balance immediate service requirements with maintaining flexibility for future requests.

Last-mile delivery optimization has emerged as a critical application domain for RL-based routing due to the operational complexities and high costs associated with urban delivery operations. Last-mile delivery involves numerous challenging factors including dense customer distributions, time window constraints, traffic congestion, parking limitations, and diverse package sizes requiring different vehicle types [51]. RL approaches to last-mile optimization have explored multi-objective formulations that simultaneously minimize delivery costs, maximize customer satisfaction through on-time delivery, and reduce environmental impact through route efficiency. The integration of real-time traffic data and predictive models of congestion patterns enables RL agents to dynamically reroute delivery vehicles to avoid delays and maintain schedule adherence. Recent pilot deployments have shown that RL-based delivery routing systems can reduce total travel distance by fifteen to twenty percent compared to human-planned routes while improving on-time delivery rates [52].

Ride-sharing and on-demand transportation services present unique routing challenges that align well with RL solution approaches. These services must continuously match incoming customer requests to available vehicles while considering passenger preferences, driver locations, traffic conditions, and service quality metrics such as wait times and detour distances [53]. RL formulations for ride-sharing typically model each vehicle as an agent that learns a policy for accepting or rejecting ride requests and selecting routes that serve multiple passengers efficiently. The sequential nature of these decisions, where early choices affect subsequent opportunities and system state evolution, makes RL particularly suitable compared to myopic matching algorithms that optimize each decision independently. Research has shown that MARL approaches where vehicles coordinate implicitly through learned policies can achieve system-wide efficiency improvements of twenty to thirty percent in terms of passenger wait times and vehicle utilization compared to decentralized greedy strategies [54].

Autonomous vehicle routing introduces additional considerations related to energy management, safety constraints, and coordination with human-driven vehicles in mixed traffic environments. RL provides a framework for learning driving policies that simultaneously address navigation, obstacle avoidance, and route optimization objectives [55]. The continuous state and action spaces characteristic of vehicle control problems necessitate the use of actor-critic methods or policy gradient algorithms capable of handling high-dimensional continuous domains. Recent work has explored hierarchical RL architectures for autonomous delivery vehicles where high-level policies make routing decisions about which customers to visit in what order while low-level policies handle the detailed motion planning and control needed to execute these routes safely. The ability of RL to learn from both simulation and limited real-world experience through transfer learning techniques has proven crucial for developing robust autonomous routing systems [56].

Distribution network design optimization using RL addresses strategic decisions about facility locations, capacities, and connections that determine the structure of supply chain networks. Traditional approaches formulate network design as mixed-integer programs that optimize expected costs under scenarios representing future demand and cost parameters [57]. RL offers an alternative that learns network design policies through simulation of long-term operations, enabling the discovery of network configurations that perform robustly across diverse future realizations rather than optimizing for specific scenarios. Recent research has applied RL to multi-echelon network design problems where decisions involve selecting warehouse locations, allocating capacities, and establishing transportation links between facilities. The RL agent learns to evaluate network configurations based on their operational performance simulated over extended time horizons, internalizing trade-offs between fixed infrastructure costs and variable operational expenses [58].

Figure 2 illustrates an RL-optimized multi-echelon distribution network and serves as a concrete example of how RL can support long-horizon, strategic supply chain decisions. The visualization highlights the hierarchical flow from upstream suppliers to distribution centers, warehouses, and downstream customer zones, where the learned policy implicitly balances fixed infrastructure choices with operational considerations such as inventory positioning and shipment routing. By visualizing connectivity and flow intensity across echelons, the figure helps interpret how RL-based design can improve system-wide performance (e.g., service levels and inventory turnover) under demand variability, compared with static scenario-driven network planning.

Inventory positioning and allocation within distribution networks represents another application area where RL demonstrates advantages over conventional approaches. The challenge involves determining optimal inventory levels at different network locations to balance holding costs against stockout risks under uncertain demand [59]. RL formulations treat inventory decisions as sequential actions taken periodically in response to observed demand patterns and current inventory states. The ability to learn policies that adapt inventory levels dynamically based on seasonal trends, promotional activities, and regional demand correlations enables more responsive supply chain operations compared to static safety stock policies derived from analytical models. Research has demonstrated that RL-based inventory policies can reduce total inventory holding costs by ten to fifteen percent while maintaining or improving service levels compared to conventional base stock policies.

Figure 2. Visualization of a Multi-Echelon Distribution Network Optimized Using RL.

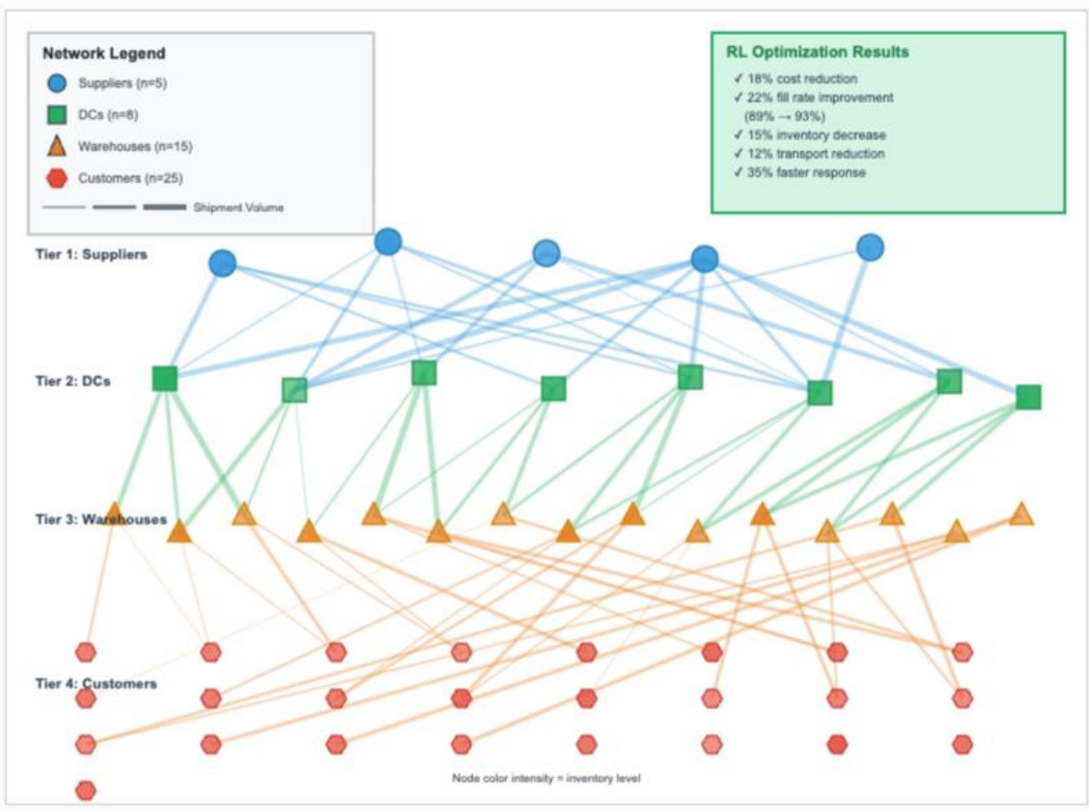

Figure 2. Multi-Echelon Distribution Network Optimized Using Reinforcement Learning

5. Challenges and Future Directions

The scalability of RL approaches to large-scale routing and network design problems remains a fundamental challenge limiting broader practical adoption. While recent advances have enabled RL methods to handle problem instances with several hundred decision variables, many real-world applications involve thousands of customers, hundreds of vehicles, and complex constraint structures that strain current algorithmic capabilities [60]. The computational complexity of RL training scales unfavorably with problem size due to the exponential growth of the state-action space and the sample requirements for policy learning. Future research must develop more efficient RL architectures and training procedures that can handle enterprise-scale routing problems without prohibitive computational costs. Hierarchical decomposition strategies that break large problems into smaller subproblems, each addressed by specialized RL agents operating at different temporal or spatial scales, represent one promising direction for achieving scalability.

Sample efficiency poses another critical challenge for deploying RL in real-world routing applications where extensive trial-and-error learning in live operational environments is infeasible due to costs, safety concerns, and service quality requirements. Current RL algorithms typically require millions of environment interactions to learn effective policies, corresponding to years of simulated operational experience. While simulation-based training can generate synthetic experience rapidly, sim-to-real transfer remains problematic when simulated environments fail to capture all relevant aspects of real-world complexity. Future research directions include developing better domain randomization techniques that expose RL agents to diverse simulated scenarios encompassing the full range of conditions encountered in practice, investigating meta-learning approaches that enable rapid adaptation to new problem instances with minimal additional training, and exploring hybrid methods that combine RL with traditional optimization to leverage existing domain knowledge and reduce learning requirements.

The interpretability and explainability of learned routing policies represents an increasingly important research challenge as RL systems move toward deployment in consequential applications. Deep neural network policies that achieve state-of-the-art performance on routing benchmarks often function as black boxes whose decision-making logic is opaque to human operators and stakeholders. This opacity creates difficulties for debugging when policies produce unexpected behaviors, limits trust and acceptance among users accustomed to understanding why particular routing decisions are made, and poses regulatory challenges in domains requiring decision transparency. Future work should explore architectures that balance performance with interpretability, such as attention mechanisms that provide insights into which problem features influence decisions, policy distillation approaches that extract simplified rule-based approximations of complex neural policies, and hybrid systems where RL learns components of routing algorithms whose overall logic remains interpretable.

The robustness of RL policies to distribution shift and adversarial conditions requires further investigation before deployment in safety-critical routing applications. Learned policies may perform excellently during training and testing on historical data but fail catastrophically when confronted with novel scenarios not represented in training distributions. Adversarial examples where small perturbations to inputs cause large changes in policy outputs pose particular concerns for routing applications where malicious actors might manipulate traffic data or customer requests to disrupt operations. Research directions for improving robustness include adversarial training where policies are exposed to intentionally challenging scenarios during learning, certified robustness approaches that provide formal guarantees about policy behavior under bounded input perturbations, and ensemble methods that combine multiple learned policies to improve reliability through diversity.

The integration of RL with other artificial intelligence technologies including computer vision, natural language processing, and knowledge representation offers opportunities for more capable routing systems.

Computer vision enables autonomous vehicles to perceive their surroundings and make routing decisions informed by real-time visual observations of traffic conditions, road hazards, and parking availability. Natural language processing allows routing systems to interpret free-text customer instructions and preferences, incorporating soft constraints that traditional optimization models struggle to formalize. Knowledge graphs can represent domain expertise about routing best practices, regulatory requirements, and operational constraints in structured forms that guide RL exploration and improve sample efficiency. Future research should investigate architectures that effectively combine these complementary technologies to create comprehensive intelligent routing systems.

Federated and distributed RL approaches represent important future directions for scenarios where multiple organizations or geographic regions must coordinate routing decisions while preserving data privacy and operational autonomy. Centralized RL training that pools data from all participants may be infeasible due to competitive concerns, regulatory restrictions, or communication bandwidth limitations. Federated RL enables collaborative policy learning where participants train local models on their proprietary data and share only model updates rather than raw data. Distributed RL coordinates multiple agents learning simultaneously in different parts of the problem space, enabling parallelization that accelerates training while respecting organizational boundaries. Research challenges include developing communication-efficient federated RL algorithms suitable for routing applications, ensuring learned policies generalize across the heterogeneous data distributions of different participants, and addressing incentive compatibility to encourage truthful participation.

The environmental sustainability of transportation and logistics operations has become an urgent societal priority, creating opportunities for RL to contribute to green routing and network design. Traditional routing objectives focused exclusively on cost minimization may produce solutions with high environmental impacts through excessive fuel consumption, empty vehicle miles, or inefficient consolidation. Multi-objective RL formulations can balance economic efficiency with environmental metrics including carbon emissions, air quality impacts, and noise pollution. Research should investigate how RL can discover innovative routing strategies that achieve superior environmental performance, such as coordinating with renewable energy availability for electric vehicle charging or dynamically consolidating shipments to reduce transportation intensity. The long-term temporal horizons and complex trade-offs inherent in sustainability optimization align well with RL's strengths in sequential decision-making under uncertainty.

6. Conclusion

This review has provided a comprehensive analysis of RL approaches to dynamic routing and distribution network design, examining methodological foundations, application domains, and future research directions in this rapidly evolving field. The fundamental advantage of RL for routing optimization stems from its ability to learn adaptive policies through environmental interaction, naturally accommodating the sequential decision-making structure and dynamic uncertainties characteristic of real-world logistics operations. Recent advances in DRL, including attention mechanisms, GNN architectures, and sophisticated policy optimization algorithms, have dramatically expanded the scale and complexity of routing problems amenable to RL solution methods. These technical innovations have enabled RL approaches to match or exceed the performance of traditional optimization techniques on standard benchmarks while demonstrating superior adaptability to changing operational conditions.

The literature review revealed several mature research streams addressing different aspects of routing optimization through RL, from attention-based architectures for sequential route construction to MARL frameworks for coordinated fleet management. The integration of GNN with RL has proven particularly effective by providing inductive biases that capture the spatial structure of transportation networks and enable generalization across problem instances of varying scales. Hybrid approaches combining RL with classical optimization methods have demonstrated that leveraging the complementary strengths of learned policies and mathematical programming can achieve performance exceeding either approach individually. Transfer learning and meta-learning techniques have begun addressing the sample efficiency challenges that previously limited practical deployment, enabling rapid adaptation of learned policies to new problem instances with minimal additional training.

Applications of RL to routing and distribution problems span operational and strategic decision contexts, from real-time vehicle routing under dynamic customer requests to long-term network design determining facility locations and capacity allocations. The demonstrated success in domains including last-mile delivery, ride-sharing optimization, and autonomous vehicle routing provides evidence of practical viability while highlighting implementation considerations regarding computational requirements, data availability, and integration with existing operational systems. The ability of RL methods to optimize complex multi-objective functions incorporating cost efficiency, service quality, and environmental sustainability positions these approaches as valuable tools for addressing contemporary challenges in transportation and logistics.

Despite substantial progress, significant challenges remain before RL becomes widely adopted for production routing and network design systems. Scalability to enterprise-scale problems involving thousands of decision variables requires continued algorithmic innovation in architectures and training procedures. Sample efficiency must improve to enable learning from limited real-world experience rather than requiring extensive simulation-based training. Interpretability and explainability of learned policies need enhancement to build stakeholder trust and satisfy regulatory transparency requirements. Robustness to distribution shift and adversarial conditions requires further investigation to ensure reliable performance across the full range of operational scenarios. Addressing these challenges through the research directions outlined in this review will advance RL-based routing systems toward broader practical impact.

The convergence of RL with emerging technologies including Internet of Things sensor networks, 5G communications infrastructure, and edge computing platforms creates new opportunities for intelligent routing systems that leverage real-time operational data for adaptive decision-making. Federated RL approaches enable collaborative learning across organizational boundaries while preserving data privacy, potentially transforming how logistics providers coordinate operations in shared transportation networks. The integration of RL with sustainability objectives positions these methods to contribute meaningfully to environmental goals while

maintaining economic efficiency. As algorithmic capabilities continue advancing and computational resources become increasingly accessible, RL approaches to routing and distribution optimization will likely see accelerating adoption across diverse application domains, fundamentally reshaping how transportation and logistics systems are planned and operated.

References

Alimohammadi, M., & Behnamian, J. (2025). Generalized Benders decomposition approach for designing a reverse logistics network for unused drugs within a circular economy framework. *RAIRO-Operations Research*, *59*(5), 2633–2656.

AlMahamid, F., & Grolinger, K. (2021). Reinforcement learning algorithms: An overview and classification. In *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 1–7). IEEE. https://doi.org/10.1109/CCECE53003.2021.9544827

Bai, R., Chen, X., Chen, Z. L., Cui, T., Gong, S., He, W., & Zhang, H. (2023). Analytics and machine learning in vehicle routing research. *International Journal of Production Research*, *61*(1), 4–30. https://doi.org/10.1080/00207543.2021.1980459

Baxter, J., & Bartlett, P. L. (2025). Reinforcement learning in POMDPs via direct gradient ascent. *arXiv Preprint*. https://arxiv.org/abs/2512.02383

Belkhale, S., Li, R., Kahn, G., McAllister, R., Calandra, R., & Levine, S. (2021). Model-based meta-reinforcement learning for flight with suspended payloads. *IEEE Robotics and Automation Letters*, *6*(2), 1471–1478. https://doi.org/10.1109/LRA.2021.3056417

Berto, F., Hua, C., Park, J., Kim, M., Kim, H., Son, J., & Park, J. (2023). RL4CO: A unified reinforcement learning for combinatorial optimization library. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.

Bijvoet, B. J. (2021). Machine and deep learning models for vehicle routing problems: A literature review. *arXiv Preprint*. https://arxiv.org/abs/2109.00366

Campelo, P., Neves-Moreira, F., Amorim, P., & Almada-Lobo, B. (2019). Consistent vehicle routing problem with service level agreements: A case study in the pharmaceutical distribution sector. *European Journal of Operational Research*, *273*(1), 131–145. https://doi.org/10.1016/j.ejor.2018.07.048

Cappart, Q., Chételat, D., Khalil, E. B., Lodi, A., Morris, C., & Veličković, P. (2023). Combinatorial optimization and reasoning with graph neural networks. *Journal of Machine Learning Research*, *24*(130), 1–61.

Cengiz, E., Yilmaz, C., & Kahraman, H. T. (2025). A literature review of drone-truck routing problems: Challenges and future research directions. *RAIRO-Operations Research*, *59*(5), 3169–3205.

Chang, M., Tang, L., & Zhao, S. (2025). A reinforcement learning-based Lagrangian decomposition approach for energy-oriented scheduling optimization in steelmaking process. *IEEE Transactions on Automation Science and Engineering*. https://doi.org/10.1109/TASE.2025.3531145

Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. *Applied Sciences*, *14*(20), 9213. https://doi.org/10.3390/app14209213

Chen, X., Ulmer, M. W., & Thomas, B. W. (2022). Deep Q-learning for same-day delivery with vehicles and drones. *European Journal of Operational Research*, *298*(3), 939–952. https://doi.org/10.1016/j.ejor.2021.06.052

Dolgui, A., Ivanov, D., & Sokolov, B. (2020). Reconfigurable supply chain: The X-network. *International Journal of Production Research*, *58*(13), 4138–4163. https://doi.org/10.1080/00207543.2019.1683247

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, *55*(2), 895–943. https://doi.org/10.1007/s10462-021-09996-w

Gul, M., Ahmad, H., Shafi, M. Z., Bajwa, M. T. T., Ahsaan, M., & Rehman, M. A. U. (2025). The role of reinforcement learning in advancing artificial intelligence: An experimental study with Q-learning and DQN. *The Asian Bulletin of Big Data Management*, *5*(3), 122–134.

Huang, C., He, R., Ai, B., Molisch, A. F., Lau, B. K., Haneda, K., & Zhong, Z. (2022). Artificial intelligence enabled radio propagation for communications—Part I: Channel characterization and antenna-channel optimization. *IEEE Transactions on Antennas and Propagation*, *70*(6), 3939–3954. https://doi.org/10.1109/TAP.2022.3140854

Joe, W., & Lau, H. C. (2020). Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. In *Proceedings of the International Conference on Automated Planning and Scheduling*, *30*, 394–402.

Joshi, C. K., Laurent, T., & Bresson, X. (2019). An efficient graph convolutional network technique for the travelling salesman problem. *arXiv Preprint*. https://arxiv.org/abs/1906.01227

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, 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

Kool, W., van Hoof, H., Gromicho, J., & Welling, M. (2022). Deep policy dynamic programming for vehicle routing problems. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (pp. 190–213). Springer. https://doi.org/10.1007/978-3-031-08011-1_14

Kwon, Y. D., Choo, J., Kim, B., Yoon, I., Gwon, Y., & Min, S. (2020). POMO: Policy optimization with multiple optima for reinforcement learning. In *Advances in Neural Information Processing Systems*, *33*, 21188–21198.

Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. *IEEE Access*, *12*, 195974–195985. https://doi.org/10.1109/ACCESS.2024.3503337

Li, S., Yan, Z., & Wu, C. (2021). Learning to delegate for large-scale vehicle routing. In *Advances in Neural Information Processing Systems*, *34*, 26198–26211.

Ma, Y., Li, J., Cao, Z., Song, W., Zhang, L., Chen, Z., & Tang, J. (2021). Learning to iteratively solve routing problems with dual-aspect collaborative transformer. In *Advances in Neural Information Processing Systems*, *34*, 11096–11107.

Mardešić, N., Erdelić, T., Carić, T., & Đurasević, M. (2023). Review of stochastic dynamic vehicle routing in the evolving urban logistics environment. *Mathematics*, *12*(1), 28. https://doi.org/10.3390/math12010028

Missaoui, O. (2023). *Markov decision processes: A gentle tutorial*. SSRN. https://doi.org/10.2139/ssrn.4535241

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2020). A unifying framework for reinforcement learning and planning. *arXiv Preprint*. https://arxiv.org/abs/2006.15009

Morabit, M., Desaulniers, G., & Lodi, A. (2023). Machine-learning–based arc selection for constrained shortest path problems in column generation. *INFORMS Journal on Optimization*, *5*(2), 191–210. https://doi.org/10.1287/ijoo.2022.0058

Oliva, M., Banik, S., Josifovski, J., & Knoll, A. (2022). Graph neural networks for relational inductive bias in vision-based deep reinforcement learning of robot control. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–9). IEEE. https://doi.org/10.1109/IJCNN55064.2022.9892618

Oroojlooyjadid, A., Nazari, M., Snyder, L. V., & Takáč, M. (2022). A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management*, *24*(1), 285–304. https://doi.org/10.1287/msom.2020.0911

Otto, F., Becker, P., Vien, N. A., Ziesche, H. C., & Neumann, G. (2021). Differentiable trust region layers for deep reinforcement learning. *arXiv Preprint*. https://arxiv.org/abs/2101.09207

Pateria, S., Subagdja, B., Tan, A. H., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys*, *54*(5), 1–35. https://doi.org/10.1145/3453160

Rakelly, K., Zhou, A., Finn, C., Levine, S., & Quillen, D. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 5331–5340). PMLR.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, *21*(178), 1–51.

Rismanto, H., & Judijanto, L. (2025). Dynamic routing in urban logistics: A comprehensive review of AI, real-time data, and sustainability impacts. *Sinergi International Journal of Logistics, 3*(2), 68–79.

Serrano-Hernandez, A., Faulin, J., de la Torre, R., & Cadarso, L. (2020). Agent-based simulation improves e-grocery deliveries using horizontal cooperation. In *2020 Winter Simulation Conference (WSC)* (pp. 1242–1253). IEEE. https://doi.org/10.1109/WSC48552.2020.9384026

Shah, S., Lowalekar, M., & Varakantham, P. (2020). Neural approximate dynamic programming for on-demand ride-pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(1), 507–515. https://doi.org/10.1609/aaai.v34i01.5400

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research, 69*, 343–418. https://doi.org/10.1613/jair.1.12007

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation. In *ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 21–25). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9413905

Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. *Asian Business Research Journal, 10*(11), 1–13.

Thyssens, D., Dernedde, T., Sentanoe, W., & Schmidt-Thieme, L. (2025). On distributional dependent performance of classical and neural routing solvers. *arXiv Preprint*. https://arxiv.org/abs/2508.02510

Ulmer, M. W., Thomas, B. W., Campbell, A. M., & Woyak, N. (2021). The restaurant meal delivery problem: Dynamic pickup and delivery with deadlines and random ready times. *Transportation Science, 55*(1), 75–100. https://doi.org/10.1287/trsc.2020.0990

Veličković, P., Buesing, L., Overlan, M., Pascanu, R., Vinyals, O., & Blundell, C. (2020). Pointer graph networks. In *Advances in Neural Information Processing Systems, 33*, 2232–2244.

Verma, A. (2019). Verifiable and interpretable reinforcement learning through program synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 9902–9903. https://doi.org/10.1609/aaai.v33i01.33019902

Vouros, G. A. (2022). Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys, 55*(5), 1–39. https://doi.org/10.1145/3501719

Wan, W., Fu, J., Yuan, X., Zhu, Y., & Su, H. (2025). LodeStar: Long-horizon dexterity via synthetic data augmentation from human demonstrations. In *Proceedings of the 9th Annual Conference on Robot Learning*.

Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., & Miao, Q. (2024). Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35*(4), 5064–5078. https://doi.org/10.1109/TNNLS.2022.3141619

Wang, Y., & Xing, S. (2025). AI-driven CPU resource management in cloud operating systems. *Journal of Computer and Communications, 13*(6), 135–149. https://doi.org/10.4236/jcc.2025.136009

Wang, Y., He, H., & Tan, X. (2020). Truly proximal policy optimization. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence* (pp. 113–122). PMLR.

Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. *Sensors, 25*(11), 3396. https://doi.org/10.3390/s25113396

Yu, J. Q., Yu, W., & Gu, J. (2019). Online vehicle routing with neural combinatorial optimization and deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems, 20*(10), 3806–3817. https://doi.org/10.1109/TITS.2018.2886774

Zhang, C., Odonkor, P., Zheng, S., Khorasgani, H., Serita, S., Gupta, C., & Wang, H. (2020). Dynamic dispatching for large-scale heterogeneous fleet via multi-agent deep reinforcement learning. In *2020 IEEE International Conference on Big Data* (pp. 1436–1441). IEEE. https://doi.org/10.1109/BigData50022.2020.9378405

Zhang, S., Li, J., Shi, L., Ding, M., Nguyen, D. C., Tan, W., & Han, Z. (2023). Federated learning in intelligent transportation systems: Recent applications and open problems. *IEEE Transactions on Intelligent Transportation Systems, 25*(5), 3259–3285. https://doi.org/10.1109/TITS.2023.3274305

Zhao, J., Mao, M., Zhao, X., & Zou, J. (2021). A hybrid of deep reinforcement learning and local search for the vehicle routing problem. *IEEE Transactions on Intelligent Transportation Systems, 22*(11), 7208–7218. https://doi.org/10.1109/TITS.2020.3009103

Zhao, R., Li, Y., Fan, Y., Gao, F., Tsukada, M., & Gao, Z. (2024). A survey on recent advancements in autonomous driving using deep reinforcement learning: Applications, challenges, and solutions. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2024.3380124

Zhou, G., Li, X., Li, D., & Bian, J. (2024). Learning-based optimization algorithms for routing problems: Bibliometric analysis and literature review. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2024.3354021

Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., & Whiteson, S. (2019). Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 7693–7702). PMLR.