Research Annals of Industrial and Systems Engineering

www.raise.reapress.com

Res. Ann. Ind. Syst. Eng. Vol. 3, No. 2 (2026) 76–85.

Paper Type: Original Article

Designing a Bioterrorism Response Network Considering Transportation: A Reinforcement Learning Approach

Mahdie Sadeghian[1], Fereshteh Parvaresh[1,*]

Department of Industrial and Systems Engineering, Isfahan University of Technology, Isfahan, Iran; mahdie.sadeqian@in.iut.ac.ir; parvaresh@iut.ac.ir.

Citation:

Received: 19 December 2025 *Revised: 13 February 2026* *Accepted: 19 April 2026*	Sadeghian, M., & Parvaresh, F. (2026). Designing a bioterrorism response network considering transportation: A reinforcement learning approach. *Research Annals of Industrial and Systems Engineering*, *3*(2), 76-85.

Abstract

Bioterrorism attacks are among the most serious national security threats, requiring the design of rapid and effective response networks. This study presents a three-level hybrid model based on the Defender–Attacker–Defender (DAD) framework, in which strategic stockpiling, tactical distribution, and operational transportation decisions are addressed simultaneously. The novelty of this research lies in integrating the Vehicle Routing Problem (VRP) with capacity, time-window, and demand uncertainty considerations into the context of biological attacks. To overcome computational complexity and ensure real-time responsiveness to environmental changes, a Reinforcement Learning (RL) module is developed that learns optimal drug distribution routes through interaction with the environment, while balancing storage, transportation, and human loss costs. Numerical experiments and sensitivity analyses show that the proposed algorithm outperforms traditional methods by reducing both casualties and total costs, and it achieves fast convergence and high efficiency under uncertainty. Ultimately, this approach can serve as a foundation for developing real-scale response systems to biological crises.

Keywords: Bioterrorism, Biological attacks, Vehicle routing problem, Crisis logistics, Mathematical modeling Reinforcement learning.

1 | Introduction

In recent years, biological attacks have emerged as a significant threat to national security, as even a small quantity of pathogenic agents can pose a widespread risk to public health. Experiences from terrorist incidents and bioterrorism simulation exercises conducted in countries such as the United States have demonstrated that effective preparedness against such threats requires robust prevention and control systems, as well as well-designed medical countermeasure supply chains [1]. Existing studies have primarily addressed response supply chain design through three-stage Defender–Attacker–Defender (DAD) models, in which strategic decisions regarding the location of pharmaceutical stockpiles are followed by operational decisions

Corresponding Author: parvaresh@iut.ac.ir

https://doi.org/10.22105/raise.v3i2.88

concerning drug distribution [1], [2]. Although robust optimization techniques have commonly been employed to solve these models [3], their applicability to large-scale problems is often limited by computational complexity. Furthermore, many studies simplify the transportation component and overlook practical considerations such as vehicle capacity constraints, delivery time windows, and uncertainties associated with demand and transportation routes. In practice, incorporating transportation decisions corresponds to the classical Vehicle Routing Problem (VRP), which is known to be computationally challenging [4], [5]. To address these limitations, the present study extends the traditional DAD framework by explicitly integrating the transportation layer, resulting in a DAD–VRP structure. Moreover, Reinforcement Learning (RL) is adopted as the solution approach to manage the increased complexity of the proposed model. RL algorithms, including policy-gradient methods, REINFORCE, Actor–Critic, and Q-learning, have been widely recognized as effective techniques for solving dynamic decision-making problems under uncertainty [6–9]. In this study, the REINFORCE algorithm is combined with a recurrent neural network architecture to identify efficient drug distribution routes. This integration enables the model to achieve superior performance compared with conventional optimization-based approaches [8].

2 | Problem Definition

The present study extends the three-stage DAD framework proposed in [10]. Within this framework, following decisions regarding warehouse location and inventory allocation, and after the occurrence of an attack, the third stage focuses on the distribution of medical countermeasures. In many previous studies, this stage has been simplified by assuming direct transportation of pharmaceuticals from warehouses to demand centers. However, in real-world emergency response operations, transportation activities are subject to various practical constraints, including limited vehicle capacities, delivery time requirements, and uncertainties in demand and transportation routes.

To better capture these operational realities, the transportation layer is incorporated into the conventional DAD framework, resulting in a new DAD–VRP structure. In addition to addressing demand coverage and inventory management, the proposed model explicitly considers transportation-related constraints such as vehicle capacity limitations, delivery scheduling requirements, and the prioritization of high-risk regions. Consequently, the resulting problem is formulated as a multi-stage nonlinear mixed-integer optimization model. Owing to its computational complexity, solving this model using conventional optimization approaches becomes challenging, thereby motivating the adoption of advanced solution techniques such as RL.

3 | Mathematical Formulation

The development of the proposed model requires a set of assumptions that define the operational characteristics and scope of the problem. Accordingly, this section first presents the assumptions adopted in the study, followed by the notation and parameters used in the mathematical formulation. Subsequently, the problem is formulated as a mathematical optimization model that captures the interactions among facility location decisions, inventory allocation, attack scenarios, and transportation operations within the proposed DAD–VRP framework.

3.1 | Assumptions

Due to the integration of the VRP into the framework proposed in [10] and the adoption of RL as the solution methodology, the following assumptions are considered in this study:

Assumption 1. Each vehicle has a limited carrying capacity, and transportation time is assumed to be a function of the travel distance.

Assumption 2. Only a finite fleet of vehicles is available for resource distribution operations.

Assumption 3. Warehouse locations, target regions, demand levels, remaining inventory, and response times are represented as the states of the RL environment.

Assumption 4. The selection of delivery routes and vehicle assignments for resource distribution constitutes the actions of the learning agent.

Assumption 5. The reward function is defined based on a combination of demand coverage, delivery delay minimization, and transportation cost reduction.

3.2 | Notations

The notation used throughout this study is summarized in *Table 1*.

Table 1. Notations.

Sets			
V	Set of all nodes.	$E \subset V \times V$	Set of edges connecting the nodes.
$V_D \subset V$	Set of demand nodes.	K_i	Set of vehicles available at the warehouse i.
$V_s \subset V$	Set of warehouse (storage) nodes.	D	Set of attack scenarios.
Parameters			
h_i	Inventory holding cost per unit of pharmaceutical stock at warehouse node i (i ∈ V_S).	b	Cost associated with the loss of a human life.
t_{pq}	Travel time from node p to node q.	c_i	Storage capacity of warehouse i (i ∈ V_S).
ρ_t	Survival rate as a function of treatment time (a decreasing function).	T_{max}	Maximum allowable delivery time.
$\rho_0 \in [0,1]$	Survival rate in the absence of treatment.	Γ_i	Attack budget associated with the descendants of node i (i ∈ V).
$\hat{d}_j$	Maximum possible demand at demand node j (j ∈ V_D).	Q_i	Capacity of each vehicle assigned to warehouse i.
M	A sufficiently large positive constant.		
Decision variables			
x_i	Quantity of pharmaceutical inventory stored at node i (i ∈ V_S).	$z_j^{i,k}(d) \in \{0,1\}$	Equals 1 if vehicle k from warehouse i visits demand node j under scenario d; 0 otherwise (k ∈ K_i, i ∈ V_S, j ∈ V_D).
$s_j(d)$	Unmet treatment demand (shortage) at demand node *j* under attack scenario d (j ∈ V_D).	$q_j^{i,k}(d)$	Quantity of pharmaceuticals delivered to demand node j by vehicle k departing from warehouse i under scenario d.
$y_{pq}^{i,k}(d) \in \{0,1\}$	Equals 1 if vehicle k from warehouse i travels directly from node p to node q under scenario d; 0 otherwise (k ∈ K_i, i ∈ V_S, p,q ∈ V_D).	$\omega_p^{i,k}(d)$	Arrival time of vehicle k from warehouse i at demand node p under scenario d.

3.3 | Mathematical Model

The mathematical formulation proposed in this study extends the LLC model presented in [10] and is represented by *Eqs. (1)-(12)*. The model consists of three interconnected components: Initial inventory allocation, demand generation following an attack, and the transportation stage formulated as a VRP.

The proposed problem is formulated as a minimization model. As shown in *Eq. (1)*, the objective function consists of two primary components. The first component minimizes inventory holding costs, while the second minimizes the maximum number of fatalities resulting from insufficient or delayed treatment. The model incorporates inventory capacity and demand coverage constraints, and its overall structure is adapted from the framework introduced in [10]. *Constraint (2)* enforces the storage capacity limitation of each warehouse, whereas *Constraint (3)* ensures that the quantity of pharmaceuticals transported from a warehouse does not exceed its available inventory. *Constraint (4)* requires each vehicle to depart from its designated origin warehouse.

Since the VRP component is modeled dynamically in this study, the departure warehouse and the return warehouse of a vehicle are not necessarily identical. Therefore, the index i denotes the departure warehouse, while i_in represents the arrival warehouse. *Constraint (6)* maintains flow conservation at each demand node by balancing vehicle arrivals and departures whenever a node is visited. *Constraint (7)* determines the quantity of pharmaceuticals delivered to each demand node.

Demand satisfaction is modeled in *Constraint (8)*, where the level of demand coverage is determined by the amount of delivered pharmaceuticals and any remaining shortage. *Constraint (9)* represents the vehicle capacity limitation. *Constraints (10)* and *(11)* correspond to the delivery time restriction and the recursive time relationship associated with vehicle movements between nodes, respectively. Finally, *Constraint (12)* defines the domains of the decision variables.

The attack scenarios considered in this study are generated according to the methodology presented in [10].

$$\min \sum_{i\in V_s} h_i x_i + \max_{d\in D} b[\sum_{j\in V_D}(1-\rho_0)s_j(d) + \sum_{j\in V_D}\sum_{i\in V_s}\sum_{k\in K_i}(1-\rho(\omega_j^{i,k}(d))q_j^{i,k}(d)], \tag{1}$$

$$x_i \le c_i \ , \quad \forall i \in V_s, \tag{2}$$

$$\sum_{j\in V_D}\sum_{k\in K_i} q_j^{i,k} \le x_i, \quad \forall i \in V_s, \forall d \in D, \tag{3}$$

$$\sum_{q:(i,q)\in E} y_{iq}^{i,k}(d) = 1, \quad \forall i \in V_s, \forall d \in D, \forall k \in K_i, \tag{4}$$

$$\sum_{p:(p,i_{in})\in E} y_{p,i_{in}}^{i,k}(d) = 1, \quad \forall i_{in} \in V_s, \forall d \in D, \forall k \in K_i, \forall i \in V_s, \tag{5}$$

$$\sum_{p:(p,r)\in E} y_{p,r}^{i,k}(d) = \sum_{q:(r,q)\in E} y_{r,q}^{i,k}(d), \tag{6}$$

$$q_j^{i,k}(d) \le M z_j^{i,k}(d),$$
$$\sum_{p:(p,j)\in E} y_{p,j}^{i,k}(d) = z_j^{i,k}(d), \sum_{q:(j,q)\in E} y_{j,q}^{i,k}(d) = z_j^{i,k}(d), \quad \forall j \in V_D, \forall d \in D, \forall k \in K_i, \forall i \in V_s \tag{7}$$

$$\sum_{i\in V_s}\sum_{k\in K_i} q_j^{i,k}(d) + s_j(d) = \hat{d}_j(d), \quad \forall j \in V_D, \forall d \in D \tag{8}$$

$$\sum_{i\in V_D} q_j^{i,k}(d) \le Q_i, \quad \forall d \in D, \forall k \in K_i, \forall i \in V_s, \tag{9}$$

$$\omega_j^{i,k}(d) \le T_{max}, \quad \forall j \in V_D, \forall d \in D, \forall k \in K_i, \forall i \in V_s, \tag{10}$$

$$\omega_q^{i,k}(d) \ge \omega_p^{i,k}(d) + t_{pq} - M(1 - y_{pq}^{i,k}(d)), \quad \forall (p,q) \in E, \forall d \in D, \forall k \in K_i, \forall i \in V_s, \tag{11}$$

$$x_i \geq 0, s_j(d) \geq 0, q_j^{i,k}(d) \geq 0, y_{pq}^{i,k} \in \{0,1\}, z_j^{i,k}(d) \in \{0,1\}. \quad (12)$$

4 | Solution Methodology

Given the inherent characteristics of pharmaceutical distribution problems in emergency situations, including nonlinearity, environmental dynamics, and uncertainty, conventional approaches such as linear programming and heuristic algorithms often exhibit limited effectiveness due to their restricted generalization capability and inability to adapt to real-time environmental changes. Consequently, this study adopts an RL approach, a branch of machine learning in which an agent learns through interaction with its environment and improves its decision-making process via trial-and-error experience [11], [12]. The objective of the agent is to maximize the cumulative reward over time while progressively learning an effective decision policy from observed outcomes.

To implement this approach, a simulation environment was developed consisting of a central warehouse with limited storage capacity, a set of demand nodes, and a fleet of vehicles with predefined capacities. At each decision step, the state of the environment is represented by the remaining demand at each node, the current locations of the vehicles, their residual capacities, and the elapsed time since the beginning of the episode. Based on the observed state, the agent selects an action, which may include traveling to a demand node, returning to the warehouse, or remaining idle. The reward function is designed to encourage successful pharmaceutical delivery, shorter delivery times, and efficient resource utilization. Conversely, delayed responses, unmet demand, and increased operational costs result in negative rewards.

The REINFORCE algorithm is employed to train the agent in the proposed environment. As one of the fundamental policy-gradient methods, REINFORCE directly updates the policy without requiring an approximation of the value function, making it suitable for problems characterized by large state and action spaces [13]. At the end of each episode, the total operational cost, including inventory holding costs, transportation costs, and costs associated with casualties, is calculated and reported as the primary performance measure.

To maintain computational tractability, an initial implementation of the proposed model was conducted using a limited-scale problem instance. The input parameters used in the experiments are summarized in *Table 2*, where parameter values were selected to ensure a balanced representation of the problem environment. Furthermore, the pseudocode of the proposed solution procedure is presented in *Fig. 1*.

Table 2. Parameters of the simulation environment.

Parameter	Value	Parameter	Value
Number of nodes	10	Transportation cost	Uniform distribution over [0, 1]
Number of available vehicles	3	Treatment survival rate parameter	Uniform distribution over [0.5, 1]
Vehicle capacity	100	Maximum demand	Uniform distribution over [10, 50]
Maximum allowable time	50	Attack parameter	Uniform distribution over [0, 1]
Baseline survival rate	0.7	Distance between nodes	Uniform distribution over [1, 10]
Number of neurons in the neural network	16	Cost of loss of life	Uniform distribution over [0, 1]
		Inventory holding cost	Uniform distribution over [0, 1]

```
Initialize policy network parameters θ randomly
For each episode = 1 to M do
    Reset environment with initial state s_0
    Initialize empty trajectory τ = []
    For t = 0 to T-1 do
        Sample action a_t ~ π_θ(a_t|s_t)
        Execute action a_t, observe reward r_t and next state s_{t+1}
        Store (s_t, a_t, r_t) in trajectory τ
        s_t ← s_{t+1}
    End For
    For each step t in trajectory τ do
        Compute return G_t
    Update policy parameters
    End For
    Evaluate performance (optional)
End For
```

Fig. 1. REINFORCE sodu code.

5 | Numerical Results and Sensitivity Analysis

The proposed model was evaluated using the parameter settings presented in *Table 2* and trained over 5,000 episodes. Since fixed values were not assigned to inventory holding costs and several other cost parameters, different runs of the model may produce different objective function values. Therefore, the results reported in this section correspond to a representative execution of the proposed framework.

During the initial training phase with 1,000 episodes, significant fluctuations were observed in the total operational cost, cumulative reward, number of casualties, and inventory holding cost. However, as the number of training episodes increased, the learning process gradually stabilized and converged toward improved solutions. When the number of episodes was increased to 5,000, convergence occurred more rapidly, and the magnitude of fluctuations was substantially reduced, as illustrated in *Fig. 2*.

Overall, the results indicate that increasing the number of training episodes enables the agent to learn more effective decision policies, leading to lower operational costs and fewer casualties. These findings demonstrate that the REINFORCE algorithm, despite its relatively simple structure, is capable of learning high-quality decision policies in complex and uncertain environments. The observed convergence behavior further confirms the suitability of RL for solving dynamic pharmaceutical distribution problems under bioterrorism-related emergency conditions. Finally, the optimal inventory allocation obtained for each node is reported in *Table 3*.

Table 4 compares the performance of the proposed algorithm under two learning settings: with exploration and without exploration. In the absence of exploration, the algorithm converges more rapidly to a stable policy and achieves a lower total operational cost. However, the resulting policy exhibits limited diversity, increasing the likelihood of convergence to locally optimal solutions.

In contrast, incorporating exploration encourages a broader search of the solution space, resulting in more diverse decision policies and improved performance with respect to critical humanitarian measures, particularly casualty reduction. Although the total operational cost is slightly higher in this setting, the exploration mechanism enables the agent to identify more effective long-term strategies that would otherwise remain undiscovered.

These findings highlight the importance of maintaining a balanced exploration–exploitation trade-off in RL -based emergency response systems. While exploitation promotes rapid convergence and cost reduction, controlled exploration can ultimately produce more robust, adaptive, and socially desirable policies, especially in environments characterized by uncertainty and dynamic operational conditions.

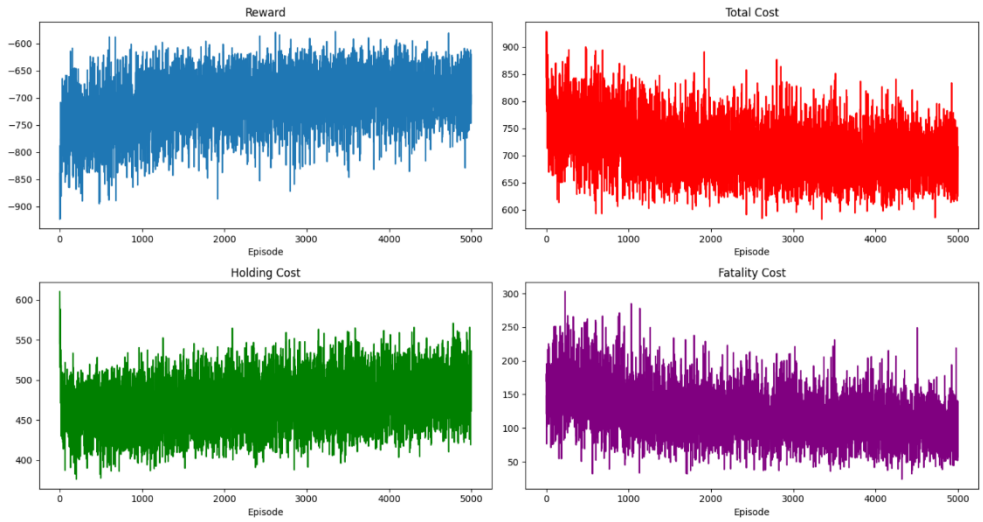

Fig. 2. Evolution of total cost during the 5,000-episode training process.

Table 3. Inventory allocation across nodes after 5,000 episodes (without exploration).

Node	Inventory Level (Units)	Node	Inventory Level (Units)
Node 1	29	Node 6	29
Node 2	36	Node 7	28
Node 3	25	Node 8	29
Node 4	30	Node 9	36
Node 5	22	Node 10	25

Table 4. Performance comparison between exploration and non-exploration policies.

Performance Metric	Without Exploration	With Exploration	Performance Metric	Without Exploration	With Exploration
Inventory Holding Cost	727.49	592.79	Casualty Cost	65.88	94.07
Transportation Cost	264.92	132.62	Total Cost	1058.28	818.82

Figs. 3-5 illustrate the routes generated for each vehicle during different stages of the training process. In these figures, Node 0 represents the central warehouse, while the highlighted routes indicate the paths selected by the RL agent. The routes shown in red correspond to the decisions chosen by the learned policy. Presenting the routes at different training stages provides insight into the evolution of the agent's behavior and demonstrates how routing decisions become progressively more structured and efficient as learning proceeds.

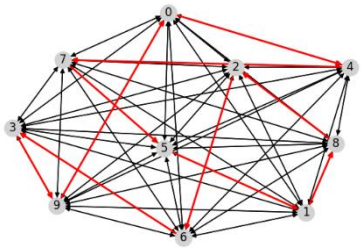

Fig. 3. Vehicle 1's route.

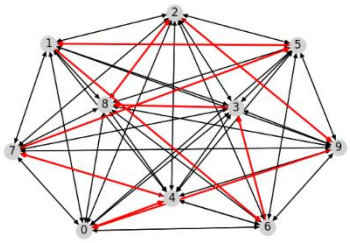

Fig. 4. Vehicle 2's route.

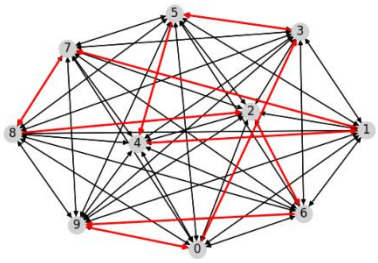

Fig. 5. Vehicle 3's route.

A disaggregated analysis of the cost components throughout the training process revealed that transportation and casualty costs decreased significantly, while inventory holding costs remained relatively stable. This finding indicates that the RL algorithm employed in this study reduced casualties by increasing the initial inventory levels at selected nodes, thereby lowering the contribution of higher-risk cost components. Furthermore, during the final 100 episodes, approximately 72% of the total cost was attributed to inventory holding, 16% to casualties, and 11% to transportation. The algorithm's reward also improved from initially negative values to values approaching the optimal level, indicating a reduction in total cost and an improvement in decision-making quality. Overall, under uncertain conditions, the algorithm was able to learn a near-optimal policy that achieved a balance among inventory, transportation, and casualty costs. In addition, a comparison was conducted to identify the difference between the optimal solution and the solution obtained by the algorithm. All of these results are presented in *Fig. 6*.

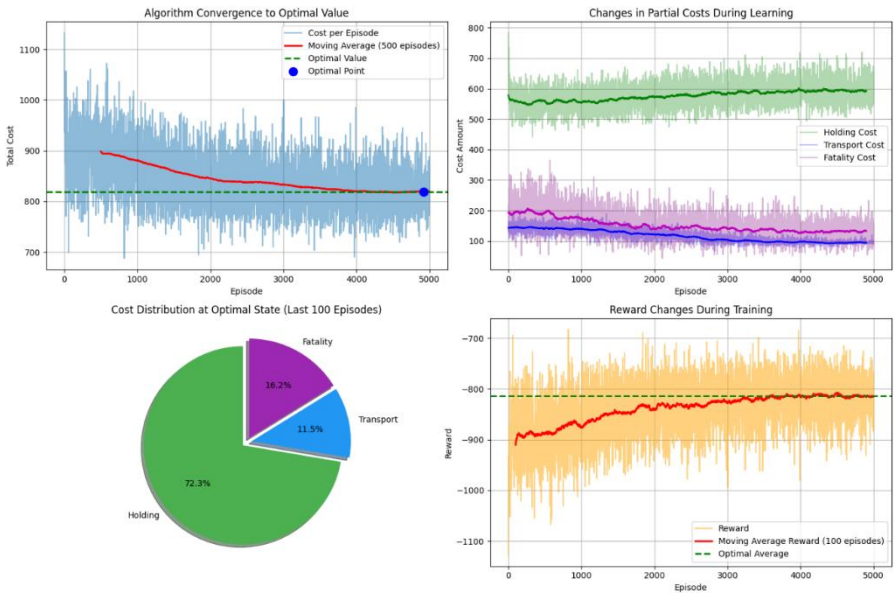

Fig. 6. Comparison between the optimal solution and the solution obtained by the REINFORCE algorithm.

6 | Conclusion

This study presents a novel framework for responding to biological attacks through the integration of the DAD model and the VRP. By employing RL to optimize pharmaceutical distribution, the proposed problem was solved without requiring model linearization, and the corresponding results were obtained. The findings demonstrate the capability of the proposed framework to respond dynamically to environmental changes while maintaining an effective balance among different cost components. Furthermore, the results indicate that the RL algorithm employed in this study achieved satisfactory convergence behavior.

For future research, more advanced RL algorithms, such as PPO and DQN, may be applied and compared with the proposed approach. In addition, the model can be extended to larger-scale problem instances in order to further evaluate its effectiveness, scalability, and computational performance.

Author Contributaion

Conceptualization, M. S. and F. P.; methodology, M. S. and F. P.; validation, M. S.; formal analysis, M. S.; writing-creating the initial design, M. S.; writing-reviewing and editing, F. P.; monitoring, F. P.; project management, F. P.

Conflicts of Interest

No conflict of interest.

References

[1] Herrmann, J., Kahn, B., Wollek, S., & Nicholson, A. (2016). *The nation's medical countermeasure stockpile: Opportunities to improve the efficiency, effectiveness, and sustainability of the CDC strategic national stockpile: Workshop summary*. National Academies Press. https://library.nationalmuseum.gov.ph/digital-collection/digi/NA02/2016/23532.pdf

[2] Brown, G., Carlyle, M., Salmerón, J., & Wood, K. (2006). Defending critical infrastructure. *Interfaces, 36*(6), 530–544. https://doi.org/10.1287/inte.1060.0252

[3] Ben-Tal, A., Ghaoui, L. El, & Nemirovski, A. (2009). *Robust optimization*. Robust optimization. https://doi.org/10.1515/9781400831050

[4] Toth, P., & Vigo, D. (2002). *The vehicle routing problem*. SIAM. https://doi.org/10.1137/1.9780898718515

[5] Laporte, G. (2009). Fifty years of vehicle routing. *Transportation science, 43*(4), 408–416. https://doi.org/10.1287/trsc.1090.0301

[6] Kool, W., Van Hoof, H., & Welling, M. (2018). *Attention, learn to solve routing problems*! https://doi.org/10.48550/arXiv.1803.08475

[7] Nazari, M., Oroojlooy, A., Snyder, L., & Takác, M. (2018). Reinforcement learning for solving the vehicle routing problem. *Advances in neural information processing systems, 31*. https://doi.org/10.48550/arXiv.1802.04240

[8] Lin, B., Ghaddar, B., & Nathwani, J. (2021). Deep reinforcement learning for the electric vehicle routing problem with time windows. *IEEE transactions on intelligent transportation systems, 23*(8), 11528–11538. https://doi.org/10.1109/TITS.2021.3105232

[9] Phiboonbanakit, T., Horanont, T., Huynh, V.N., & Supnithi, T. (2021). A hybrid reinforcement learning-based model for the vehicle routing problem in transportation logistics. *Ieee access, 9*, 163325–163347. https://doi.org/10.1109/ACCESS.2021.3131799

[10] Simchi-Levi, D., Trichakis, N., & Zhang, P. Y. (2019). Designing response supply chain against bioattacks. *Operations research, 67*(5), 1246–1268. https://doi.org/10.1287/opre.2019.1862

[11] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. J*ournal of artificial intelligence research, 4*, 237–285. https://doi.org/10.1613/jair.301

[12] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press. https://mitpress.mit.edu/9780262039246/reinforcement-learning/

[13] Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in neural information processing systems, 12*. https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf