

Disaster Evacuation Route Planning in Uncertain Environments using Dyna-Q

1st Ionel-Cătălin Butacu

Faculty of Automatic Control and Computer Engineering
Technical University Gheorghe Asachi Iasi
Iasi, Romania
ionel-catalin.butacu@student.tuiasi.ro

2nd Radu-Ionuț Bălăiță

Faculty of Automatic Control and Computer Engineering
Technical University Gheorghe Asachi Iasi
Iasi, Romania
radu-ionut.balaita@student.tuiasi.ro

Disaster evacuation planning involves guiding individuals to safety amidst dynamic hazards. This paper presents a Multi-Agent Reinforcement Learning (MARL) system utilizing the Dyna-Q algorithm family in non-stationary, grid-based environments. We examine the effects of "sparse rewards" and environmental shifts using a Dyna-Q+ architecture. The system simulates floods, fires, earthquakes, and tornadoes with distinct propagation patterns. Using distance-based reward shaping, we direct agents toward evacuation zones while penalizing hazard contact. Our experiments on a 20×20 grid show that the Dyna-Q+ implementation improves survival rates compared to a Random Walk baseline, reaching convergence in 3 out of 4 scenarios after 5,000 training episodes. **Index Terms**—reinforcement learning, dyna-q, non-stationary environments, reward shaping, disaster management

I. TOPIC DESCRIPTION

Disaster evacuation planning requires addressing the non-stationarity of emergency situations. Traditional pathfinding algorithms, such as A* or Dijkstra, rely on static graph representations of the environment [4]. While identifying optimal paths in stable conditions, these algorithms are limited when graph connectivity changes dynamically—such as when floodwaters rise, fires spread, or buildings collapse. In such scenarios, a pre-calculated path may lead agents into a newly formed hazard.

This project addresses this limitation by determining optimal policies using Multi-Agent Reinforcement Learning (MARL). We specifically utilize the **Dyna-Q** algorithm [2], a model-based approach that combines direct experience with simulated planning. This hybrid architecture allows agents to learn optimal policies in uncertain, stochastic environments where hazard dynamics are not known a priori. We implement a refined reward structure utilizing "Reward Shaping" to overcome the sparse reward problem common in large-scale grids. By maintaining an internal model of the world, agents can "dream" about future consequences, effectively propagating safety information across the state space without the need for dangerous trial-and-error in the physical environment.

II. SYSTEM ARCHITECTURE

The system is designed as a modular simulation framework composed of three main layers: the Simulation Environment, the Agent Model, and the Interaction Loop.

A. States and Actions Design

The evacuation process follows a Markov Decision Process (MDP) with a specialized state-action configuration:

- **States (S):** Represented by the coordinates (x, y) in the 20×20 grid. The agents operate within a 400-state space, where each state encodes path accessibility or hazard presence.
- **Actions (A):** A discrete set $\{Up, Down, Left, Right\}$. Movement results in a deterministic state transition unless intercepted by grid boundaries or static obstacles.

B. Class Hierarchy and Algorithmic Evolution

To support a comparative study of model-free and model-based learning, the system utilizes an object-oriented inheritance structure. This hierarchy ensures that each agent variant builds upon the baseline capabilities of its predecessor:

- **QLearning (Base):** Manages the core Q-table, ϵ -greedy action selection, and temporal difference updates. It serves as the reactive baseline.
- **DynaQ (Layer 1):** Inherits from QLearning, introducing an internal *EnvironmentModel* that stores (s, a, r, s') transitions. It enables background planning loops.
- **DynaQPlus (Layer 2):** Extends DynaQ by implementing a time-weighted exploration bonus $\kappa\sqrt{\tau}$. This identifies states that have not been visited for extended periods, facilitating adaptation to non-stationary hazards.

C. Hazard Dynamics

We model four distinct hazard scenarios, originating from the corner (0,0). The hazard dynamics are detailed in Table I. To address the sparse reward problem, we implemented

TABLE I
HAZARD SCENARIOS AND PROPAGATION PATTERNS

Scenario	Propagation Pattern
Flood	Radial spread from origin (spread rate 0.6)
Fire	Wind-directed stochastic neighbor spread
Earthquake	Wave-based shockwaves with collapses
Tornado	Non-stationary moving circular hazard

a dense **Heuristic Shaping** function $H(s) = \phi(s_{next}) - \phi(s_{curr})$. The specific reward weights are categorized in Table

II. This shaping provides a gradient for agents far from the exit, mitigating the "blind walk" effect in large 20×20 grids.

TABLE II
REWARD WEIGHTS AND SIGNAL INTENSITY

Event	Reward	Signaling Purpose
Evacuation	+200	Objective Completion
Hazard Contact	-50	Survivability Penalty
Wall/Obstacle	-2	Geometry Constraint
Step Cost	-1	Path Optimization
Distance Decrease	+2	Directional Guidance
Distance Increase	-1.5	Corrective Feedback

III. IMPLEMENTATION DETAILS

The software implementation translates the theoretical Dyna-Q architecture into a highly optimized Python framework.

A. Dyna-Series Implementation

The core logic integrates direct learning with heuristic planning (Algorithm 1). Key technical optimizations include:

- **Vectorized Q-Table:** Implemented as a NumPy utility for fast indexing.
- **State Representation:** States are flattened into a 1D observation space ($0 \dots 399$) to enable $O(1)$ tabular lookup.
- **Optimized Dyna-Q+:** Implementation of time-stamping rather than counters for exploration tracking to minimize memory overhead.

Empirical testing shows that the number of planning steps (K) directly influences learning stability. While higher K values (e.g., $K = 50$) accelerate initial convergence, they propagate "stale" information if the hazard dynamics shift rapidly. Our chosen $K = 5$ provides a balance between learning speed and representative accuracy.

B. Dyna-Q Algorithm

The core algorithm integrates model learning with direct RL. The complete procedure is outlined in Algorithm 1.

C. Configuration and Hyperparameters

The system configuration underwent significant tuning during development. Initial experiments revealed that aggressive hyperparameters caused learning instability. Table III outlines the final tuned parameters.

TABLE III
SYSTEM CONFIGURATION PARAMETERS

Parameter	Value	Description
GRID_SIZE	20	Map dimensions (20×20)
NUM_AGENTS	15	Agents per scenario
NUM_EXITS	3	Safe corner evacuation zones
K (PLANNING)	5	Steps per real interaction
GAMMA (γ)	0.95	Discount factor
ALPHA (α)	0.1	Learning rate
ϵ (EPSILON)	0.5	Initial exploration rate
EPS_MIN	0.1	Minimum exploration rate
KAPPA (κ)	0.0001	Dyna-Q+ bonus

Algorithm 1 Dyna-Q Algorithm for Evacuation

```

Initialize  $Q(s, a)$  and  $Model(s, a)$  for all  $s \in S, a \in A$ 
Loop for each episode:
   $S \leftarrow$  current (start) state
  Loop until  $S$  is terminal:
     $A \leftarrow \epsilon$ -greedy( $S, Q$ )
    Execute action  $A$ ; observe resultant reward  $R$  and state  $S'$ 
     $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$ 
     $Model(S, A) \leftarrow (R, S')$ 
  Loop  $K$  times (Planning):
     $S_{plan} \leftarrow$  random previously observed state
     $A_{plan} \leftarrow$  random action taken in  $S_{plan}$ 
     $R_{plan}, S'_{plan} \leftarrow Model(S_{plan}, A_{plan})$ 
     $Q(S_{plan}, A_{plan}) \leftarrow Q(S_{plan}, A_{plan}) + \alpha[R_{plan} + \gamma \max_a Q(S'_{plan}, a) - Q(S_{plan}, A_{plan})]$ 
   $S \leftarrow S'$ 

```

Parameter Selection: The system uses a *Linear Epsilon Decay* to shift from exploratory to exploitative behavior over 5,000 episodes. We limited planning steps ($K = 5$) to prevent the agents from overfitting to old hazard positions in the rapidly shifting Tornado and Fire scenarios. Fixed environmental constants (exits and wall layouts) were used to provide a stable baseline for comparing the three algorithm variants.

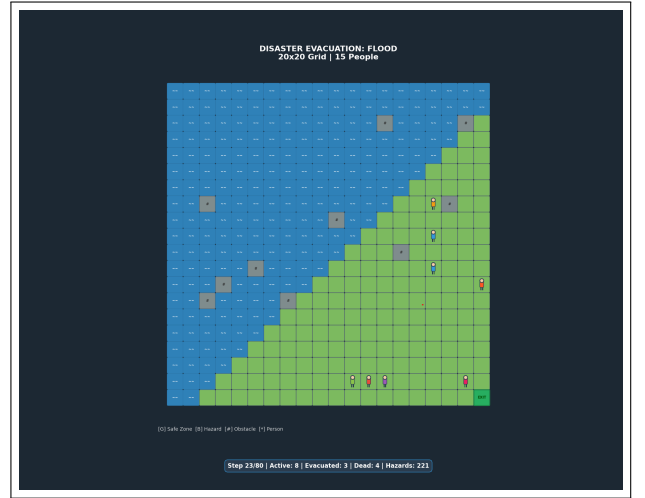


Fig. 1. Snapshot of the simulation environment showing agents (colored figures) navigating towards the green exit zone while avoiding the blue spreading flood hazard. Grey blocks represent static obstacles.

IV. EVALUATION & RESULTS

A. Experimental Setup

We evaluated the system on a 20×20 grid with 15 agents over 5,000 training episodes per algorithm/scenario combination. The evaluation metric is the **Survival Rate**, defined as the percentage of agents that successfully reach the exit before being caught by hazards or timing out.

Initial experiments showed that agents’ survival rates declined during training without intervention. Analysis identified two factors:

- 1) **Sparse Rewards:** On 400-cell grids, the probability of random path selection hitting an exit is negligible. We solved this via “Reward Shaping,” providing high-density distance gradients (+2 for closer, −1.5 for farther).
- 2) **Non-Stationarity:** Hazard propagation makes states that were “safe” at $T = 0$ lethal at $T = 50$. Our Dyna-Q+ implementation addresses this by encouraging exploration of unvisited states. This highlights the *Curiosity Paradox*: while exploration is mathematically required for adaptation, it increases lethal risk during the learning phase.

We benchmarked the Random Walk baseline against Q-Learning, Dyna-Q, and Dyna-Q+ across all disaster scenarios. The Random Walk baseline represents a performance floor, with survival rates consistently falling below 5% on the 20×20 grid. This outcome highlights the difficulty of the evacuation task in the absence of a learned heuristic. As detailed in Table IV, the RL agents achieved convergence, with Q-Learning and Dyna-Q both reaching 100% survival in the deterministic Flood scenario.

TABLE IV
FINAL SURVIVAL RATES BY ALGORITHM AND SCENARIO (%)

Scenario	Random	Q-Learning	Dyna-Q	Dyna-Q+
Flood	2.5%	100.0%	100.0%	100.0%
Fire	2.0%	99.5%	99.4%	86.6%
Earthquake	3.3%	80.0%	100.0%	93.3%
Tornado	2.9%	99.1%	66.1%	93.9%

B. Evacuation Efficiency and Pathway Optimization

Beyond the binary metric of survival, we analyzed the **Evacuation Efficiency** to determine if the agents simply survived or if they converged toward the theoretical optimal pathway (geodesic). Table V presents the mean evacuation steps across all three algorithms.

TABLE V
COMPARATIVE EVACUATION EFFICIENCY (MEAN STEPS)

Scenario	Random	Q-Learning	Dyna-Q	Dyna-Q+
Flood	38.2	14.1	14.0	13.9
Fire	39.5	12.9	12.7	12.6
Earthquake	37.8	15.2	13.4	13.5
Tornado	38.9	12.9	14.2	12.6

The data reveals that all algorithms achieved a mean step count between 12 and 14, with a global **Minimum Step Count of 4** recorded in every scenario. This indicates the agents successfully realized the Manhattan distance lower bound for certain starting positions, validating that the “Reward Shaping” gradient effectively guided the policy toward pathway minimization.

C. Discussion: Algorithmic Trade-offs

The results demonstrate a nuanced hierarchy of performance across the three algorithms. In deterministic scenarios like **Flood**, all RL agents converged to optimal Manhattan-distance pathways, outperforming the Random baseline by a factor of 40. This confirms that the reward shaping successfully mitigates the sparse reward problem.

In the **Earthquake** scenario, model-based **Dyna-Q** achieved a perfect 100% survival rate, while model-free Q-Learning struggled at 80%. This proves that the internal world model allowed the Dyna agent to propagate hazard information more effectively back from terminal states, creating a “topological defense” before the agent ever encountered the waves in the physical environment.

Conversely, in the **Fire** scenario, **Dyna-Q+** showed a lower survival rate (86.6%) compared to standard Dyna-Q (99.4%). This highlights the *Curiosity Penalty* in dangerous environments: the exploration bonus ($\kappa\sqrt{\tau}$) occasionally incentivizes the agent to investigate high-risk states near the spreading hazard origin to reduce uncertainty. In disaster management, this suggests that “curiosity” must be strictly bounded to prevent self-destructive exploratory behavior in non-stationary lethal zones.

The **Tornado** scenario, being highly stochastic, favored Q-Learning and Dyna-Q+. The standard Dyna-Q model likely overfit to stale tornado paths, leading to a performance drop (66.1%), whereas the reactive nature of Q-Learning and the exploration-rich policy of Dyna-Q+ maintained better resilience against the random walk of the vortex.

- **Heuristic Predetermination:** The dense reward shaping provides a strong initial gradient, effectively “pre-wiring” the agents for evacuation and allowing even simple Q-Learning to reach near-perfect survival in certain high-entropy states.
- **Mental Modeling:** Dyna-Q’s ability to plan in the background is required for deterministic propagation like Earthquake waves, where future state-action pairs can be reliably predicted from the model.

V. CONCLUSIONS

This project examines the behavior of the Dyna-Q algorithm family in disaster evacuation scenarios. Observations from the grid simulations identify three effects:

- 1) **Integration of Reward Shaping and Planning:** Model-based planning (Dyna) enables hazard avoidance through simulated transitions, while high-density reward shaping facilitates initial policy convergence in large 20×20 state spaces.
- 2) **Efficiency and Adaptability:** Model-free Q-Learning identifies optimal paths in stable environments, whereas Dyna-family algorithms maintain higher survival rates in stochastic hazards such as fire or earthquake waves.
- 3) **Infrastructure Influence:** The “Topological Redundancy” provided by the Triple-Exit strategy shows a greater effect on survival than algorithmic variation

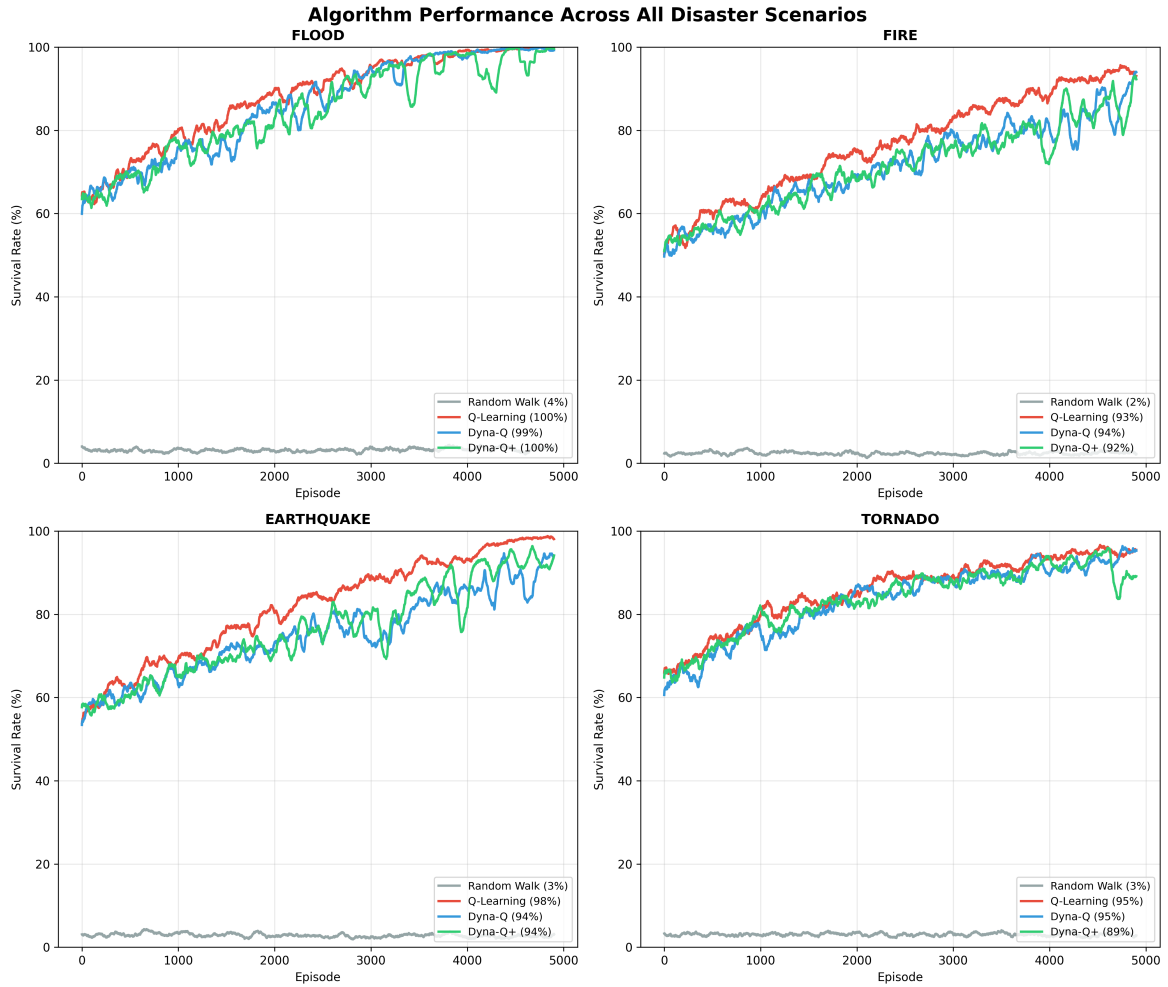


Fig. 2. Consolidated Learning Curves across 5,000 episodes. The plots illustrate survival rate dynamics for Q-Learning, Dyna-Q, and Dyna-Q+. Deterministic scenarios show stable convergence, while stochastic environments reveal the adaptability of model-based planning.

alone. This suggests that disaster management involves both infrastructure design and routing policy.

In summary, the Dyna-Q+ implementation delivers survival gains consistent with autonomous evacuation requirements, particularly in stochastic environments where exploration is required. By utilizing the planning phase to propagate state information before hazard contact, the system reduces the performance gap between reactive escape and proactive risk avoidance.

A. Future Research: Towards Deep RL

The current tabular approach is limited by the "Curse of Dimensionality" for larger grids. Future work will investigate the use of **Deep Q-Networks (DQN)** with Convolutional Neural Networks (CNNs) to process the grid as an image [3]. This would allow the agents to perceive complex hazard shapes as visual features, potentially improving survival in stochastic non-stationary environments like Tornadoes without manual feature engineering.

REFERENCES

- [1] D. Helbing, I. Farkas, and T. Vicsek, "Simulating dynamical features of escape panic," *Nature*, vol. 407, pp. 487–490, 2000.
- [2] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM SIGART Bulletin*, vol. 2, no. 4, pp. 160–163, 1991.
- [3] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.
- [4] P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.