Applied Computer Science 2026, vol. 22, no. 2, pp. 1–15
https://doi.org/10.35784/acs_8814

Submitted: 2025-12-01 | Revised: 2026-04-11 | Accepted: 2026-04-13

Keywords: SARSA, ACO, path planning, swarm robot, reinforcement learning

Aicha HAFID [1], *Riadh HOCINE* [1*], *Lahcene GUEZOULI* [1]
[1] University of Batna 2, Algeria, aicha.hafid@univ-batna2.dz, riadh.hocine@univ-batna2.dz, lahcene.guezouli@univ-batna2.dz
[*] Corresponding author: riadh.hocine@univ-batna2.dz

Path planning in swarm robotics exploration using SARSA and ACO algorithms

Abstract

Swarm robotics is a particularly promising approach for autonomous exploration in complex and uncertain environments, with applications ranging from environmental monitoring to hazardous-area inspection. A major challenge lies in optimising robot trajectories to minimise travel distance while ensuring comprehensive and effective coverage of the exploration area. In this context, we propose a hybrid path-planning framework that combines the SARSA Reinforcement Learning algorithm with the ACO approach, drawing inspiration from collective coordination mechanisms in nature, particularly the use of pheromones as a medium for self-organisation. This framework leverages both individual learning and swarm intelligence in a complementary manner, thereby enabling more robust, scalable, and efficient exploration. A comparative analysis of the two methods was conducted to identify the most effective approach for optimising robot trajectories while minimising energy consumption. In this process, robots take into account obstacle avoidance, whether obstacles are traversable, using either pheromone-based environmental marking or reinforcement learning strategies. Simulation results demonstrate the effectiveness of a hybrid model that integrates SARSA with ACO, significantly enhancing trajectory quality and exploration coverage. However, they also reveal that increasing the environment size substantially increases the total travel distance and slows SARSA convergence due to the expansion of the state space. To overcome this limitation, future work will explore neural network–based value function approximation, which is expected to improve generalisation and accelerate convergence in large-scale scenarios.

1. INTRODUCTION

Swarm robotics is an emerging and promising paradigm for addressing the challenges of autonomous navigation and exploration in complex, uncertain, and dynamic environments. Inspired by the collective behaviour of social insects, swarm systems leverage simple individual agents to achieve sophisticated global behaviours through local interactions and distributed decision-making. This approach has demonstrated high potential in a wide range of real-world applications, including post-disaster search and rescue operations (Hafid et al., 2024; Moosavi et al., 2024; Phadke & Medrano, 2024; Ganduri & Pathri, 2024) precision agriculture for large-scale crop monitoring (Abhang et al., 2024; Tan et al., 2024; Lytridis et al., 2021; Chiu et al., 2024), and planetary surface exploration, where robustness, scalability, and autonomy are critical.

Traditional robotics systems often rely on centralised control architectures, which can suffer from limited scalability, vulnerability to single points of failure, and reduced adaptability in highly dynamic or partially observable environments. In contrast, decentralised swarm approaches offer enhanced fault tolerance, better resource distribution, and real-time adaptability by enabling each robot to operate autonomously, with minimal reliance on global knowledge or continuous communication.

In this context, Reinforcement Learning (RL) plays a key role in enabling robots to adapt their behaviours through trial-and-error interactions with the environment. Among the various RL algorithms, SARSA (State–Action–Reward–State–Action) stands out as particularly well-suited for swarm robotics due to its on-policy learning strategy. Unlike Q-learning, which updates value estimates based on the best possible future action (off-policy), SARSA updates values based on the action actually taken, promoting more cautious and context-aware behaviours, an important characteristic for navigation in partially known, obstacle-rich, or cooperative multi-agent scenarios.

In this work, we propose a hybrid decision-making model that combines the learning capabilities of SARSA with a bio-inspired coordination mechanism derived from Ant Colony Optimisation (ACO). This approach enables each robot to learn optimal policies autonomously while indirectly coordinating with its peers through pheromone-like environmental cues. Our objective is to improve path-planning efficiency, reduce exploration redundancy, and maintain distributed control without requiring complex communication infrastructure. The proposed method aims to strike a balance between individual learning and collective intelligence, paving the way for more scalable and resilient swarm robotics systems in real-world deployments.

The remainder of this paper is organised as follows. Section 2 describes the architecture of the proposed swarm system. Section 3 presents the stochastic model used for the simulation. Section 4 introduces the SARSA and ACO algorithms used in this study. Section 5 describes the proposed hybrid SARSA–ACO approach. Section 6 presents the experimental setup, followed by simulation results in Section 8 and discussion in Section 9.

1.2. Related work

Classical swarm exploration strategies, primarily based on simple reactive behaviours, have shown promising results in structured environments but suffer from several limitations when deployed in more complex or dynamic settings. These limitations include restricted local perception, poor scalability, lack of memory or learning, and redundant path coverage due to insufficient coordination mechanisms among robots (Abdulsaheb & Kadhim, 2023; Alshammrei et al., 2022; Li et al., 2022; Yang et al., 2023; Badamasi et al., 2025). Such shortcomings can lead to inefficient exploration, increased energy consumption, and mission delays, especially in unknown or cluttered environments.

To overcome these challenges, recent research efforts have increasingly turned to RL techniques, enabling agents to learn optimal navigation policies based on interaction with the environment. Algorithms such as Q-learning (Maoudj & Hentout, 2020; Low et al., 2019; Li et al., 2022; Puente-Castro et al., 2024; Zhou et al., 2024) and SARSA (Mohan et al., 2021; Dong et al., 2024; Zaghbani et al., 2024; Habiba & Jahan, 2023; Peng et al., 2023) have been successfully applied to swarm robotic systems, offering the advantage of experience-based adaptation, which allows agents to refine their decisions over time through trial and error (Kakish, Elamvazhuthi & Berman, 2021; Liu et al., 2023; Chang et al., 2021). SARSA, in particular, has been recognised for promoting more conservative, safer policies, making it well-suited to cooperative scenarios where the environment is only partially observable.

In parallel, several bio-inspired approaches have emerged, especially those mimicking pheromone-based coordination mechanisms inspired by ant colony behaviours. In these models, agents deposit virtual pheromones in the environment, which are subsequently sensed by other agents to influence their movement decisions. This stigmergic communication promotes decentralised coordination and helps avoid redundant exploration and accelerate convergence in multi-agent tasks (Viseras et al., 2016; Singh et al., 2020; Li & Yang, 2024).

Beyond RL and bio-inspired methods, classical graph-based algorithms such as A* (Li et al., 2022; Du, 2024) and RRT* (Mao et al., 2025; Yildiz et al., 2024) continue to be widely used for robot path planning. A* is a deterministic, heuristic-driven algorithm that excels in static and fully known environments. At the same time, RRT* (Rapidly-exploring Random Tree), as a sampling-based method, is suited for planning in high-dimensional spaces and non-convex environments. RRT* has been particularly adapted to swarm robotics for homotopic path optimisation and collision avoidance. However, both algorithms typically require global knowledge of the map and do not inherently support learning or decentralised decision-making, which limits their scalability and adaptability in real-world exploration scenarios.

More recently, Deep Reinforcement Learning (DRL) has been explored to address the limitations of tabular RL in high-dimensional state and action spaces. Algorithms such as Proximal Policy Optimisation: PPO (Wan et al., 2025; Wei et al., 2024) have been applied to multi-robot systems for tasks requiring greater policy generalisation and dynamic adaptation in partially observable or uncertain environments. While these approaches offer high representational power and flexibility, they come with significant drawbacks, including high computational demand, complex training, and limited deployment feasibility on resource-constrained swarm platforms.

In summary, although substantial progress has been made through classical planning, bio-inspired coordination, and reinforcement learning, there remains a need for hybrid approaches that combine learning capability, decentralised control, and efficient collective coordination. The current work builds on this trend

by proposing an integration of SARSA with a pheromone-inspired mechanism to harness the strengths of both paradigms for scalable, adaptive swarm exploration.

1.3. Contribution

This work proposes a hybrid SARSA–ACO framework for decentralised swarm exploration in stochastic environments. The approach combines SARSA reinforcement learning with pheromone-based coordination inspired by Ant Colony Optimisation to improve exploration efficiency. Each robot learns an adaptive navigation policy through the SARSA update rule defined in Equation (1), where s, a, r, and s' denote the state, action, reward, and next state, respectively. A pheromone-based avoidance mechanism is introduced to reduce redundant exploration among robots. The performance of the proposed method is evaluated through a comparative analysis with standard SARSA and ACO algorithms.

$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha[r + \gamma Q(s',a') - Q_{old}(s,a)] \quad (1)$$

Unlike existing RL–ACO hybrid approaches, the proposed framework integrates pheromone-based avoidance directly into the SARSA learning process. This allows robots to dynamically adapt their navigation policy while simultaneously reducing redundant exploration through decentralised pheromone signalling.

For clarity, Table 1 summarises the main notations and variables used in the SARSA, ACO, and hybrid SARSA-ACO algorithms.

Tab. 1. Notation used in SARSA, ACO, Hybrid SARSA-ACO algorithms

Notation	Description
s	Current state of the environment
a	Action selected by the robot in state s
s'	Next state from executing action a
r	Reward received after executing the action
a'	Action selected in the new states
$Q(s,a)$	Q-function: expected benefit of taking action a in state s
α	Learning rate, controlling the influence of newly acquired information
γ	Discount factor, determining the importance of future rewards
ε	The exploration rate used in the ε-greedy policy balances exploration and exploitation
ε_{min}	Minimum exploration rate
$decay$	Multiplicative factor applied to ε to gradually reduce exploration over episodes
τij	Pheromone intensity associated with the transition from cell i to cell j
P	Pheromone evaporation rate
$\Delta\tau$	Quantity of pheromone deposited after target detection
$\tau 0$	Initial pheromone level
ηij	Heuristic desirability of moving from cell i to cell j

2. ARCHITECTURE OF THE PROPOSED SYSTEM

The proposed swarm-based exploration system is designed to support decentralised decision-making, local learning, and stigmergic coordination among robots. Each robot operates autonomously and is equipped with perception, decision-making, and motion control modules.

As illustrated in Fig. 1, robots use onboard sensors to detect obstacles, identify targets, and perceive pheromone information in the environment. Decision-making is performed using the SARSA reinforcement learning algorithm, where each robot maintains a local Q-table and selects actions according to an ε-greedy policy. Cooperation among robots is achieved through a shared pheromone matrix representing virtual traces deposited in the environment. These pheromone signals guide exploration and reduce redundant visits to previously explored areas. The decision process, shown in Fig. 2, consists of observing the current state, selecting and executing an action, receiving a reward from the environment, and updating the Q-table using the SARSA learning rule. When a target is detected, a pheromone marker is deposited to indirectly inform other robots.

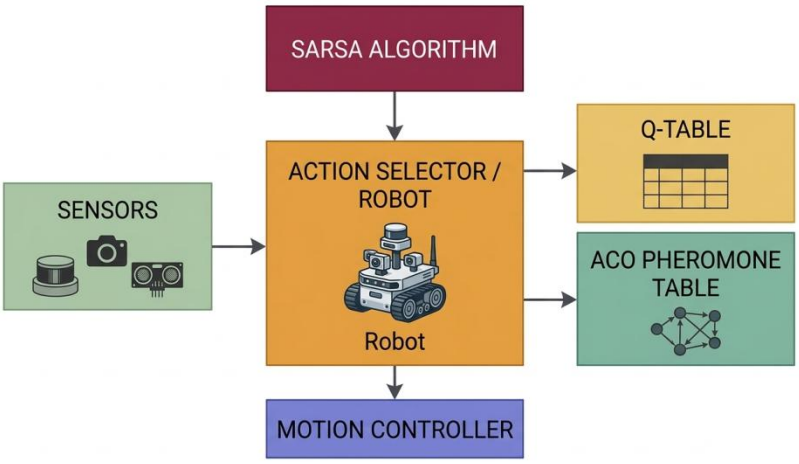

Fig. 1. System architecture of the proposed swarm robot

The proposed exploration system employs a hybrid coordination strategy that integrates direct inter-agent communication with indirect environmental signalling via pheromones. Each robot operates under a decentralised communication model, exchanging local perceptions—such as detected targets, obstacles, and partial maps—with neighbours within its sensing range. This localised approach minimises computational overhead and eliminates the single-point-of-failure risk inherent in centralised systems, thereby enhancing real-time adaptability.

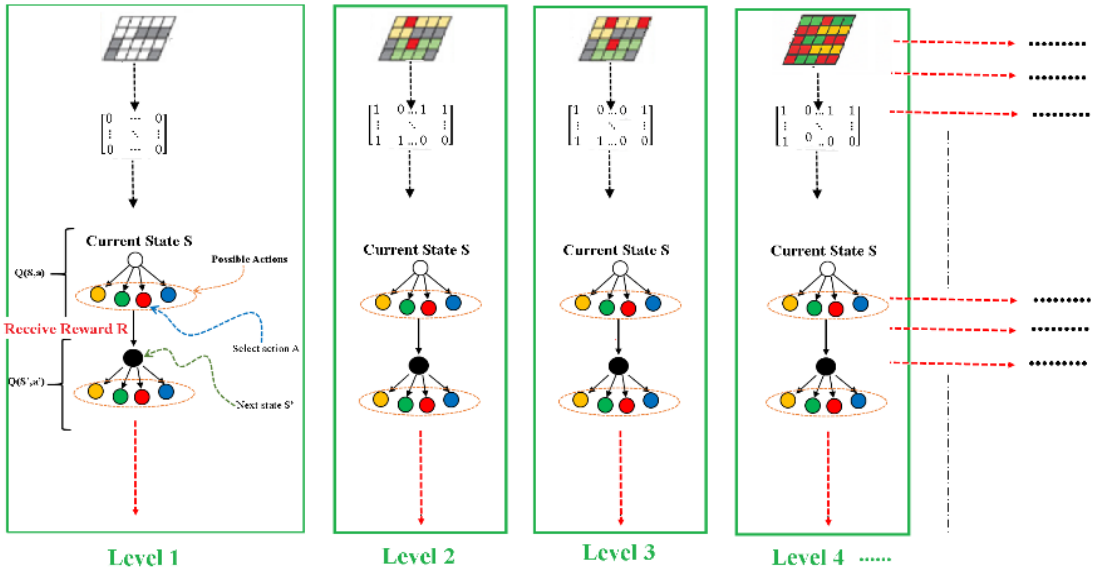

Fig. 2. Hierarchical representation of the SARSA algorithm procedure.

In parallel with this communication, each agent independently executes the SARSA reinforcement learning algorithm. By incorporating shared neighbourhood data into their local state representation, agents achieve context-aware decision-making while maintaining SARSA's on-policy requirements. This synergy ensures individualised learning that benefits from collective insights, accelerating convergence and improving behavioural robustness in dynamic environments.

3. STOCHASTIC MODEL FOR MULTI-ROBOT EXPLORATION

The simulation model considers a stochastic multi-robot system operating in a two-dimensional grid environment. A group of 12 autonomous robots is randomly deployed to explore the area and locate 30 targets distributed at random. The environment also includes two types of obstacles: traversable obstacles representing

difficult but navigable terrain, and non-traversable obstacles corresponding to impassable barriers. Each robot moves in a discrete space using four possible actions (up, down, left, right) and perceives its local state from its surroundings to guide its decision-making process.

During exploration, robots interact dynamically with the environment by observing obstacles, detecting targets, and updating their navigation policies through reinforcement learning. To ensure realistic experimentation, robots, targets, and obstacles are randomly generated at each simulation run. The system is designed to avoid deadlocks and maintain high fault tolerance, allowing the remaining robots to continue the mission even if one robot fails, thereby ensuring robust, decentralised exploration.

4. PROPOSED HYBRID SARSA-ACO METHOD

4.1. Standard SARSA algorithm

The SARSA algorithm is an on-policy reinforcement learning method that enables each robot to learn an optimal action policy through interaction with the environment. As shown in Algorithm 1, the process begins by initialising the environment, the Q-table Q(s, a), and the learning parameters, including the learning rate α, the discount factor γ, and the exploration rate ε. At each episode, the robot selects an action using an ε-greedy policy, executes the action, and observes the resulting reward and next state. A new action is then selected in the next state, and the Q-value is updated using the temporal-difference rule (Equation(1)). This process continues until a terminal state is reached, while the exploration rate gradually decreases to encourage convergence toward an efficient policy.

Algorithm 1: SARSA for Swarm Robot Exploration

Initialize environment (robots, obstacles), $Q(s, a)$
Set learning rate α, discount factor γ, exploration rate ϵ
for episode $= 1$ to MaxEpisodes **do**
  Reset environment and initialize starting state s
  Choose action a from s using ϵ-greedy policy
  for $t = 1$ to MaxSteps **do**
    Execute action a
    Observe reward r and next state s'
    Choose next action a' using ϵ-greedy policy
    Update:

$$Q_{\text{new}}(s, a) \leftarrow Q_{\text{old}}(s, a) + \alpha \left[r + \gamma Q(s', a') - Q_{\text{old}} \right]$$

    $s \leftarrow s'$, $a \leftarrow a'$
    if mission complete or s' is terminal **then**
      break
    end if
  end for
  $\epsilon \leftarrow \max(\epsilon \cdot decay, \epsilon_{min})$
end for

4.2. ACO coordination mechanism

The ACO is a metaheuristic inspired by the foraging behaviour of real ants. It is based on the principle of stigmergy, in which indirect communication occurs through environmental modifications, specifically through pheromone deposition and evaporation. In robotic swarms. ACO enables distributed decision-making without requiring any explicit communication between robots.

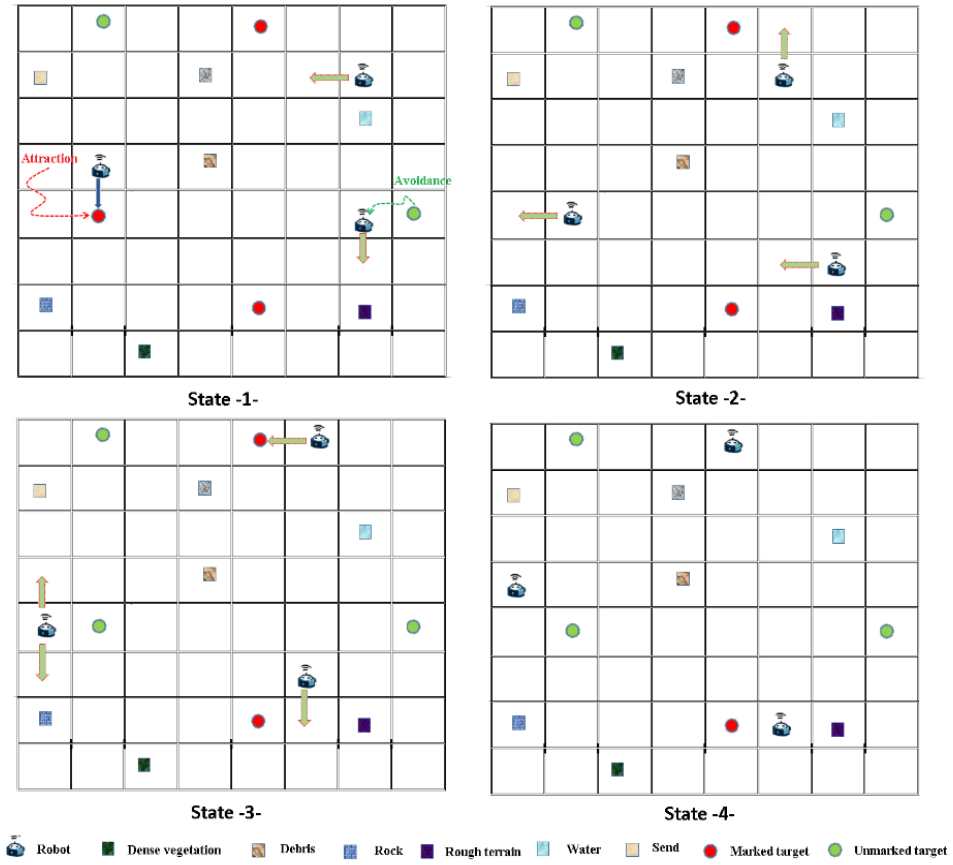

Fig. 3. Marking and avoidance principle in the modified ACO algorithm

Algorithm 2: ACO Detection and Marking Strategy

Initialize pheromone matrix $\tau(x, y) = \tau_0$ and distribute robots/targets.
Define influence parameters α (pheromone) and β (visibility).
while targets remain in the environment **do**
  for each robot **do**
    if robot is an exploration robot **then**
      Movement Selection: Select next cell (x', y') based on probability:
      $P_{ij} = \frac{[\tau_{ij}]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum [\tau_{ik}]^{\alpha} \cdot [\eta_{ik}]^{\beta}}$
      Explore cell and detect local features.
      if a target is detected **then**
        Update pheromone: $\tau(x, y) \leftarrow \tau(x, y) + \Delta\tau$ (Equation 3).
      end if
      Evaporation: Apply $\tau(x, y) \leftarrow (1 - \rho)\tau(x, y)$.
    end if
    if robot is a removal robot **then**
      Follow maximum pheromone gradient $\max(\tau)$ to reach targets.
      Remove target and reset $\tau(x, y)$ for that local area.
    end if
  end for
end while

In our implementation, a swarm of robots explores a complex environment to detect specific metallic targets such as gold or silver. When a robot detects such a target, it marks the position using a single type of virtual pheromone to indicate that the area has already been explored. Unlike classical ACO models, where pheromone trails evaporate over time, the marking here is permanent. This intentional design choice ensures

that once an area is explored, it remains marked, preventing redundant revisits by other robots. As a result, the swarm collectively focuses on unexplored regions, achieving efficient, non-overlapping exploration without centralised control (Viseras et al., 2016; Carr & Wang, 2022; Chunfeng & Fengqi, 2023). More generally, in ACO, the probability Pij of a robot moving from node i to node j depends on both the amount of pheromone τij and a heuristic value ηij (Lin, 2025), such as the inverse of the distance. The probability is defined by Equation (2):

$$P_{ij} = \frac{\tau_{ij}^{\alpha}.\eta_{ij}^{\beta}}{\sum_{k\epsilon N_i} \tau_{ik}^{\alpha}.\eta_{ik}^{\beta}} \tag{2}$$

where α and β are parameters that control the influence of the pheromone intensity and the heuristic factor, respectively. This pheromone-based approach leads to emergent collective behaviour, in which robots tend to converge on high-quality paths discovered by their peers. While highly scalable and robust to individual failures, ACO lacks an explicit memory or learning mechanism. Therefore, hybrid models that integrate ACO with reinforcement learning algorithms (SARSA in our simulation) offer a promising balance between reactive distributed coordination and adaptive learning. In our approach, once the target has been identified, the robots avoid previously marked positions while exploring the environment. This strategy helps optimise energy consumption while minimising the total trajectory length (Hafid et al., 2025).
Equation (3) formalises the mechanism by which robots avoid areas previously marked by pheromone signals. The transition probability of selecting a neighbouring cell (x,y) is defined as

$$p(x,y) = \frac{(1-D(x,y)).\eta(x,y)}{\sum_{(x',y')\epsilon N}(1-D(x',y')).\eta(x',y')} \tag{3}$$

where D(x,y) represents the avoidance function. This function returns 1 when the pheromone intensity at position (x, y) exceeds the predefined threshold τ0, indicating that the location has already been detected and marked by another robot, and returns 0 otherwise. Figure 3 illustrates the two navigation behaviours adopted by the robots during exploration. The first behaviour corresponds to attraction toward potential targets that have not yet been detected, represented by cells without pheromone traces (red circles). The second behaviour reflects avoidance of previously detected targets, where the pheromone intensity is equal to 1, and the locations are marked (green circles). In such cases, robots adapt their trajectories by selecting alternative neighbouring cells outside the marked regions. These navigation decisions are guided by the SARSA learning mechanism, which enables robots to progressively optimise their exploration trajectories while minimising redundant visits and reducing overall energy consumption.

4.3. Hybrid SARSA-ACO decision process

The proposed hybrid SARSA–ACO algorithm combines reinforcement learning with pheromone-based coordination to improve the swarm's exploration efficiency. In this framework, the Ant Colony Optimisation (ACO) mechanism provides probabilistic guidance for robot movements based on pheromone information. At the same time, the SARSA algorithm learns an optimal action-selection policy through interaction with the environment (Hu, 2023). The overall procedure of the hybrid learning and coordination process is summarised in Algorithm 3.

Let s denote the robot's current state and a the selected action. The state representation includes relevant environmental information, such as the robot's position, neighbouring obstacles, and pheromone intensity in surrounding cells. At each decision step, the robot evaluates the candidate neighbouring cells using the pheromone-based transition probability defined in Equation (3). This probability reflects the attractiveness of each neighbouring position based on the pheromone distribution and the available heuristic information in the environment.

Algorithm 3: Hybrid SARSA-ACO for Swarm Robots

Input: Learning rate α, Discount factor γ, Evaporation rate ρ, Influence factors β, η

Output: Optimized policy $Q(s,a)$ and Pheromone map $\tau(s,a)$

Initialize $Q(s,a)$ arbitrarily for all (s,a)

Initialize $\tau(s,a)$ to a small constant value τ_0

foreach *Exploration Episode* **do**

  Observe current state s

  Choose action a from s using the hybrid probability:

  $P(a|s) = \frac{[Q(s,a)]^{\beta} \cdot [\tau(s,a)]^{\eta}}{\sum_{b \in A} [Q(s,b)]^{\beta} \cdot [\tau(s,b)]^{\eta}}$

  while *s is not terminal* **do**

    Execute action a, observe reward r and next state s'

    Choose next action a' from s' using the same hybrid probability $P(a'|s')$

    `// SARSA Individual Update`

    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$

    `// ACO Collective Update (Stigmergy)`

    $\tau(s,a) \leftarrow (1-\rho)\tau(s,a) + \Delta\tau$ *// where $\Delta\tau$ depends on task success*

    $s \leftarrow s'$

    $a \leftarrow a'$

  end

end

The action selection process combines the SARSA learning mechanism with pheromone-guided exploration. Specifically, the robot applies an epsilon-greedy strategy to select its next action. With probability (1 − ε), the robot selects the action that maximises the Q-value associated with the current state. With probability ε, the robot selects an action according to the pheromone-based probability distribution P(x,y) defined in Equation (3). This hybrid strategy enables robots to exploit the learned policy while still benefiting from pheromone-guided exploration. The complete decision process is detailed in Algorithm 3.

After executing the selected action, the robot observes the immediate reward r and transitions to a new state s′. The next action a′ is selected according to the same hybrid policy. The state–action value is then updated using the SARSA learning rule:

$$Q(s,a) = Q(s,a) + \alpha \left[r + \gamma\, Q(s',a') - Q(s,a) \right] \tag{4}$$

Through this update process, the robot progressively refines its decision policy based on the rewards received during exploration. In parallel, when a robot detects a target, it deposits a pheromone marker in the corresponding cell of the environment. This pheromone signal serves as an indirect communication mechanism between robots and supports collective coordination within the swarm. The pheromone matrix, therefore, acts as a distributed memory that guides exploration and reduces redundant movements (Scharin & Jansson, 2024).

4.4. Algorithmic workflow

The proposed hybrid method integrates reinforcement learning with bio-inspired stigmergic coordination. At each decision step, a robot observes its local state, including position, surrounding obstacles, and pheromone intensity. The robot then selects an action using an ε-greedy SARSA policy while considering pheromone-based transition probabilities derived from the ACO mechanism. After executing the action, the robot receives a reward from the environment and updates its Q-table according to the SARSA update rule. If a target is detected, the robot deposits a pheromone signal in the environment to inform other robots and reduce redundant exploration.

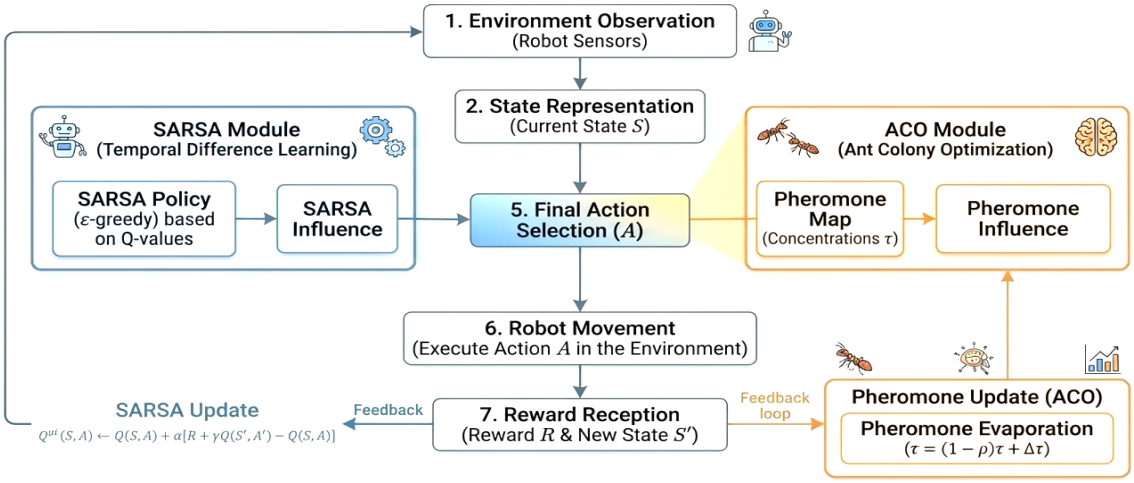

Fig. 4. Interaction scheme of SARSA and ACO algorithms

5. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed hybrid SARSA and ACO path planning method, all simulations were implemented in Python 3.12.3 using the Anaconda Navigator IDE, with a framework developed based on standard libraries such as NumPy, Pandas, TensorFlow, and Keras, supporting numerical processing and enabling future integration with Deep Reinforcement Learning (DRL) techniques; the simulation environment is designed as a 2D grid world where each robot navigates using discrete four directional movement while managing its own Q-table and interacting indirectly through a shared pheromone matrix. All experiments were conducted on a personal computer running Windows 10 Pro 64-bit, equipped with an Intel Core i5-4200M CPU at 3.40 GHz×4, 8 GB of RAM, and a 500 GB hard drive, allowing for stable execution of multiple exploration scenarios, including environments up to 2000×2000 cells.

6. SIMULATION RESULT

The simulation environment is modelled as a grid, where each cell represents an area of 1 square meter (1 m × 1m). A robot's movement from one cell to an adjacent one corresponds to a real-world displacement of 1 meter. Therefore, all trajectory distances reported in the results are expressed in meters, based on the total number of steps taken on the grid. This clear correspondence between grid steps and physical distance ensures practical interpretability and supports the model's applicability in real-world scenarios. To assess the performance and robustness of the proposed hybrid approach, we conducted extensive simulations in stochastically generated environments of varying complexity. Each simulation scenario involved a randomly assigned number and placement of static obstacles, a non-uniform initial distribution of swarm robots, and exploration targets positioned at arbitrary locations within the environment. This randomness reflects the unpredictability often encountered in real-world scenarios. The initial parameters used in all simulations are summarised in Tab. 2.

Tab. 2. Initial parameters for the simulation

Parameter	Value
Robot (N)	12
Obstacles (O)	8
Target (T)	30
Speed simulation (S)	5
Reward (R)	0

Learning rate (α)	0.001
Discount factor (γ)	0.9
Exploration rate (ε)	0.9
τ0	0.1
ρ	0.05
Δτ	1
α(pheromone)	1
β(pheromone)	2

The parameters for the pheromone mechanism used in the ACO component of the hybrid algorithm are summarised in Table 2. These parameters control pheromone deposition and evaporation and their influence on the robots' action-selection process.

For each environment size, we performed 30 independent simulation runs to ensure statistical relevance. Performance metrics included the average number of iterations required to complete the exploration task and the mean cumulative distance travelled by the robots throughout the mission. These indicators provide insights into both the system's efficiency and scalability. As reported in Tab. 3, a general trend was observed: both the number of iterations and the total path length increased with the environment's size. For example, in a 400×400 environment, the exploration task was completed in 47 iterations with an average path length of 30.07 meters. In contrast, a 1000×1000 environment required 362 iterations and a mean distance of 394.20 meters. Notably, even in a considerably larger 2000×2000 area, the system completed the exploration mission within 400 iterations, covering a total distance of 366.80 meters. The results clearly indicate that increasing the environment size significantly impacts SARSA convergence. For instance, when scaling from 400×400 to 1000×1000, the number of iterations increases from 47 to 362 (approximately +670%), while the average path length grows from 30.07 m to 394.20 m (approximately +1200%). Even in the largest scenario (2000×2000), the mission is still completed, but convergence requires 400 iterations and an average path length of 366.80 m. These results highlight the scalability of the proposed method. The integration of SARSA enables each robot to adapt its behaviour in real time, while the ACO-inspired pheromone-avoidance mechanism significantly reduces redundant exploration. This mechanism allows robots to autonomously avoid previously visited zones, improving coverage efficiency and limiting overlap in exploration paths. The curves in Fig. 4 illustrate the relationships among environment size, mission duration, and path length. The relatively smooth progression in both metrics suggests stable system behaviour as environmental complexity increases. Moreover, the system exhibits consistent convergence across multiple trials, confirming the reliability of the learning process under varying conditions.

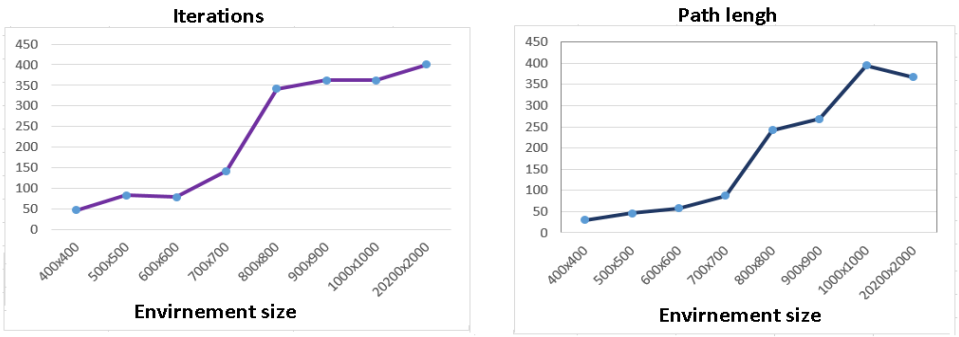

Fig. 5. Variation of iterations and path length vs environment size

In summary, the simulation outcomes demonstrate that the proposed hybrid approach not only adapts well to

increasing spatial scales but also maintains high levels of exploration efficiency. These findings reinforce the potential of SARSA combined with bio-inspired strategies for large-scale autonomous robotic deployment in unknown or dynamically evolving environments.

7. DISCUSSION

To quantitatively evaluate the proposed method, several performance metrics were considered, including average travel distance, learning convergence speed, and environmental exploration coverage.

Fig. 6 shows the evolution of the sum of Q-values over the number of iterations for each of the 12 robots in the swarm. A general upward trend can be observed during the early learning phases, indicating improved action policies as robots gain experience. However, some curves exhibit stabilisation or even decline, suggesting premature convergence to suboptimal policies or insufficient exploration. Moreover, the variability across robots highlights differences in local environments or interactions, reflecting an imbalance in learning progress. This differentiated behaviour underlines the importance of implementing coordination mechanisms or adaptive learning strategies to ensure consistent performance across the swarm. Fig. 8 and Fig. 9 clearly illustrate the robustness and efficiency of the proposed hybrid exploration strategy. In the two-dimensional visualisation (Fig. 8), the robot's trajectory demonstrates smooth navigation through a complex environment with several untraversable obstacles, including rocks, debris, and rough terrain.

The experimental system considered in this study consists of a swarm of autonomous robot agents operating in a discrete grid-based environment containing obstacles and target locations. Each robot acts as an independent agent that perceives its local state, selects actions using the SARSA policy, and dynamically interacts with the environment. Coordination among robots is achieved through a virtual pheromone mechanism inspired by ACO, enabling indirect communication between agents and reducing redundant exploration. This interaction between the agents and the environment allows the swarm to progressively learn efficient navigation strategies while ensuring effective coverage of the search space.

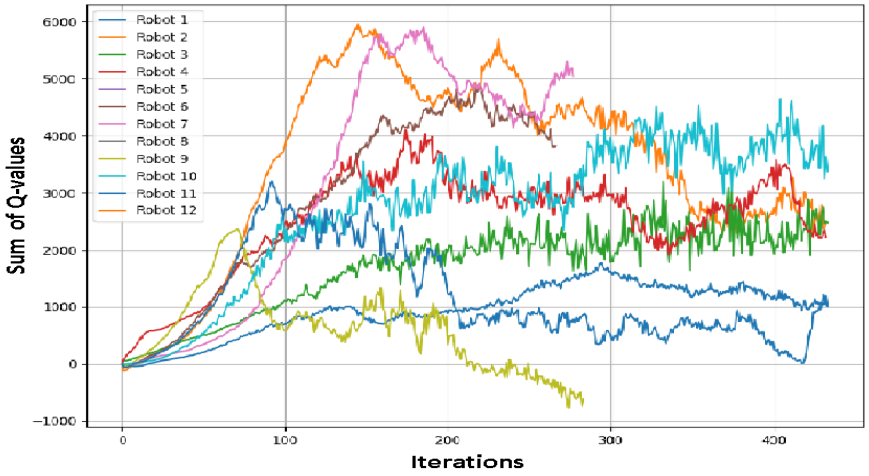

Fig. 6. Sum of Q-values for each robot

Tab. 3. Path evolution according to the environment size

Environment size	400×400	500×500	600×600	700×700	800×800	900×900	1000×1000	2000×2000
Iteration count	47	83	79	142	342	362	362	400
Path length	30.07m	46.20m	57.93m	87.80m	242.07m	268.07m	394.20m	366.80m

To better visualise the performance improvement of the proposed approach, a comparative analysis between the standard SARSA algorithm and the Hybrid SARSA-ACO method was conducted across different environment sizes, as illustrated in Fig. 7.

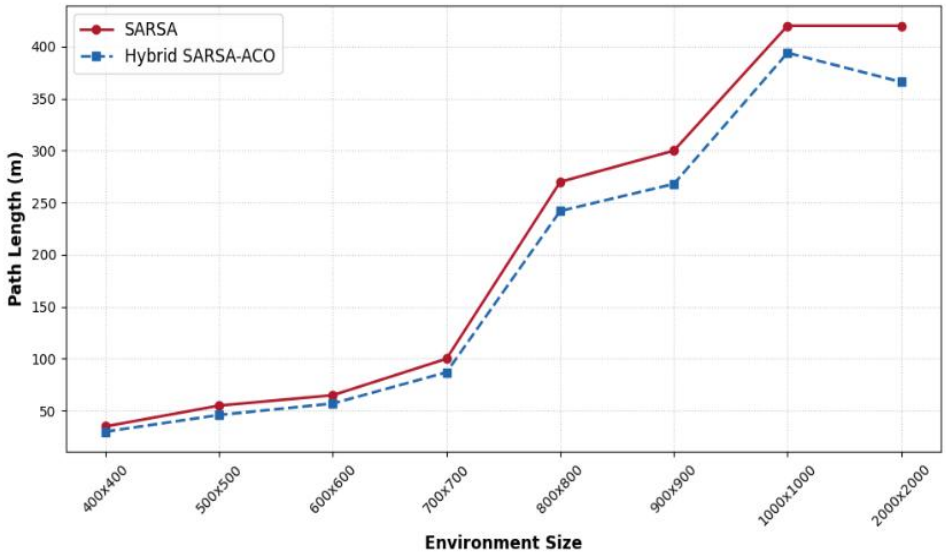

Fig. 7. Comparison between SARSA and Hybrid SARSA-ACO across different environment sizes

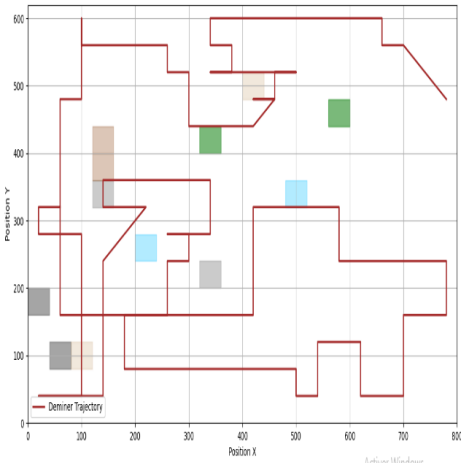

Fig. 8. Visualisation of Robot Path in 2D

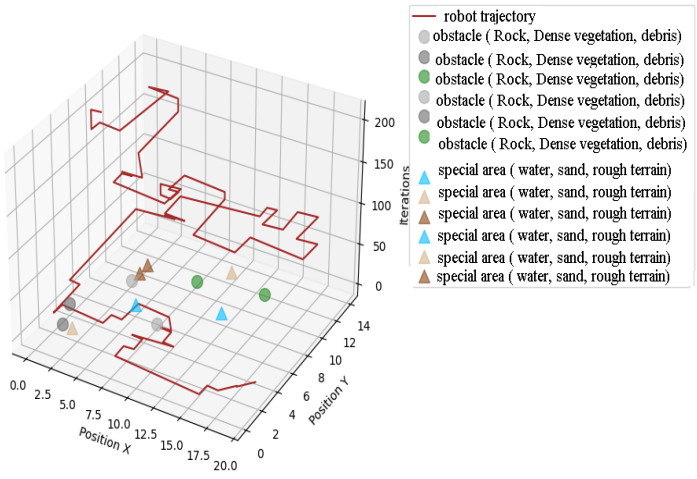

Fig. 9. Visualisation of Robot Path in 3D

The robot consistently avoids these restricted areas without encountering persistent blockages, highlighting the system's ability to anticipate environmental constraints and dynamically adapt its behaviour.

The three-dimensional representation (Fig.9) further reinforces these observations by introducing a temporal or evaluative dimension−such as cumulative *Q-values* or time steps. This perspective reveals the progressive improvement in the robot's policy, showing continuous learning and increasingly efficient decision-making. The absence of redundant movements or looping patterns emphasises a purposeful, optimised exploration process. This performance is made possible by the hybrid approach combining the SARSA reinforcement learning algorithm with a pheromone-based avoidance mechanism inspired by ACO. While SARSA enables robots to iteratively learn optimal policies through interaction with the environment, introducing virtual pheromone trails discourages revisiting previously explored areas, thereby reducing path overlap. The synergy between adaptive learning and bioinspired heuristics accelerates convergence while ensuring broad, effective coverage. It is worth noting that although most navigation proceeded smoothly, occasional deadlock situations were observed, temporarily hindering robot movement. In such cases, robots employ a local evasion strategy (Sing et al., 2020; Li & Yang, 2024) to escape the deadlock autonomously. This behaviour enhances exploration fluidity and prevents repetitive movement within the same areas, contributing to the overall efficiency and resilience of the multirobot system.

8. COMPARATIVE ANALYSIS WITH BASELINE METHODS

To further highlight the effectiveness of the proposed hybrid SARSA-ACO approach, we conducted additional experiments using two baseline strategies:

- SARSA-only: Each robot independently learns its navigation policy based on reinforcement learning without any stigmergic coordination.
- ACO-only: Robots coordinate through pheromone-based marking and avoidance, but without adaptive reinforcement learning.

Tab. 4 summarises the results obtained in environments of size 800×800, averaged over 30 independent runs. The comparison clearly demonstrates the complementary benefits of integrating adaptive learning with stigmergic coordination. Specifically, the proposed SARSA‑ACO hybrid reduces redundant trajectory overlap by approximately 22% compared to SARSA-only and achieves a 17% shorter average path length than ACO-only (Tab. 4). These improvements are consistent across multiple trials, underscoring the robustness and scalability of the proposed approach. This comparative analysis highlights the complementary nature of the two paradigms. On the one hand, SARSA allows each agent to progressively adapt its navigation strategy through direct interaction with the environment, which is particularly advantageous in partially observable or dynamic contexts.

Tab. 4. Comparison of the proposed SARSA‑ACO method with baseline approaches

Method	Avg. Path Length	Avg. Iterations	Redundancy (%)
SARSA-only	482.5	410	39%
ACO-only	436.7	368	33%
SARSA-ACO proposed	362.1	342	17%

However, when used in isolation, SARSA tends to generate redundant exploration patterns, since individual agents lack an explicit mechanism to share knowledge or avoid overlapping trajectories. On the other hand, ACO provides a robust, scalable mechanism for distributed coordination through stigmergic communication, enabling agents to implicitly share environmental information and reduce overlap. Nevertheless, ACO in its classical form does not incorporate adaptive policy refinement, which limits its ability to optimise long-term navigation decisions. By integrating the adaptive learning capability of SARSA with the decentralised coordination provided by ACO, the proposed hybrid method leverages both experience-driven learning and bio-inspired collective intelligence. This synergy results in a more efficient exploration process, characterised by reduced redundancy, improved spatial coverage, and faster convergence toward mission objectives.

9. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a hybrid approach that combines the SARSA-RL algorithm with a pheromone-based mechanism inspired by ACO to enhance path planning in swarm robotic exploration. This strategy enables each robot to autonomously learn adaptive navigation policies while benefiting from indirect stigmergic coordination through environmental marking. As a result, the system effectively reduces redundant trajectories, improves spatial coverage, and maintains a fully decentralised decision-making process. Simulation results demonstrated the robustness of the proposed method across environments of different sizes, with reliable performance even under stochastic conditions such as random obstacle placement and heterogeneous robot deployment. The proposed hybrid SARSA–ACO exploration framework also has potential applications in several real-world scenarios requiring efficient multi-robot coordination. For instance, it can be applied to environmental monitoring tasks where autonomous robots must explore large areas to collect data while avoiding redundant coverage. Similarly, in search-and-rescue missions, The proposed strategy can help robotic teams rapidly explore disaster-affected environments while maintaining decentralised coordination.

The quantitative analysis showed that larger environments significantly affect SARSA performance. For example, when scaling from 400×400 to 1000×1000, the number of iterations increased from 47 to 362 (approximately +670%), while the average path length grew from 30.07 m to 394.20 m (approximately +1200%). Although the mission was still completed in the largest environment (2000×2000), convergence required 400 iterations and an average distance of 366.80 m, confirming the scalability challenge of SARSA

in very large state‑action spaces. To mitigate these limitations, we plan to integrate neural network‑based function approximation in future work, thereby transitioning toward a DRL framework. This extension is expected to enhance generalisation, accelerate convergence, and improve adaptability in large-scale, dynamic, or partially observable environments. Furthermore, we intend to investigate advanced multi-robot coordination strategies, such as decentralised policy sharing, attention-based communication models, and cooperative task allocation. Finally, experimental validation on physical robotic platforms will be carried out to assess performance under real-world constraints, including energy consumption, sensor noise, and hardware failures. We also aim to explore real-time mapping, dynamic reward shaping, and exploration in probabilistic terrain structures. These extensions will further improve the autonomy, robustness, and efficiency of swarm robotic systems for critical missions such as search and rescue, environmental monitoring, and hazardous-area inspection.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

Abdulsaheb, J. A., & Kadhim, D. J. (2023). Classical and heuristic approaches for mobile robot path planning: A survey. *Robotics, 12*(4), Article 93. https://doi.org/10.3390/robotics12040093

Abhang, L., Gummadi, A., Changala, R., Vuyyuru, V., & Raj, A. I. I. (2024). Swarm intelligence for multi-robot coordination in agricultural automation. In *Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 455–460). IEEE. https://doi.org/10.1109/ICACCS60023.2024.10544520

Alshammrei, S., Boubaker, S., & Kolsi, L. (2022). Improved Dijkstra algorithm for mobile robot path planning and obstacle avoidance. *Computers, Materials & Continua*, *72*(3), 5939–5954. https://doi.org/10.32604/cmc.2022.028165

Badamasi, M. A., Kabir, I. K., Ahmed, G., & El-Ferik, S. (2025). Autonomous mobile robot path planning techniques, a review: Classical and heuristic techniques. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3579863

Carr, C., & Wang, P. (2022). Fast-spanning ant colony optimisation (FASACO) for mobile robot coverage path planning. *ArXiv, abs/2205.15691*. https://doi.org/10.48550/arXiv.2205.15691

Chang, L., Shan, L., Jiang, C., & Dai, Y. (2021). Reinforcement based mobile robot path planning with improved dynamic window approach in unknown environment. *Autonomous Robots, 45*(1), 51–76. https://doi.org/10.1007/s10514-020-09947-4

Chiu, D., Nagpal, R., & Haghighat, B. (2024). Optimization and evaluation of a multi robot surface inspection task through particle swarm optimization. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8996–9002). IEEE. https://doi.org/10.1109/ICRA57147.2024.10611661

Chunfeng, S., & Fengqi, W. (2023). Mobile robot path planning based on improved ant colony optimization. In *International Symposium on Artificial Intelligence and Robotics* (pp. 422–432). Springer. https://doi.org/10.1007/978-981-99-9109-9_40

Dong, H., Zhao, D., Huang, D., Yan, K., & Ren, W. (2024). Sarsa (lambda) reinforcement learning based path planning of unmanned aerial vehicles. In *Proceedings of the 2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS)* (pp. 1099–1105). IEEE. https://doi.org/10.1109/DDCLS61622.2024.10606895

Du, L. (2024). Path planning for robots integrating improved A* and DWA algorithms. *AIP Conference Proceedings, 3144*(1), Article 050004. https://doi.org/10.1063/5.0212345

Ganduri, K. V., & Pathri, B. P. (2024). Swarm intelligence in action: Particle swarm optimization and rendezvous algorithms for swarm robotics. *Journal of Field Robotics, 41*. https://doi.org/10.1002/rob.22456

Habiba, U., & Jahan, R. (2023). Path planning for UAV drones using Sarsa: Enhancing efficiency and performance. In *2023 International Conference on Drone and Robotics* (pp. 1–6). IEEE. https://doi.org/10.1109/ICICIS56802.2023.10430246

Hafid, A., Hocine, R., & Guezouli, L. (2024). Analyzing swarm robotics approaches in natural disaster scenarios: A comparative study. In *Proceedings of the 2024 1st International Conference on Innovative and Intelligent Information Technologies (IC3IT)* (pp. 1–6). IEEE. https://doi.org/10.1109/IC3IT63743.2024.10869410

Hafid, A., Hocine, R., Guezouli, L., & Momene, H. (2025). Federated reinforcement learning and Deep Q-Network: Improving fault tolerance and energy consumption in swarm robotics for mine prospection missions. *IEEE Access*, *13*, 189926-189958 https://doi.org/10.1109/ACCESS.2025.3626283

Hu, M. (2024). *Art of reinforcement learning*. Springer.

Kakish, Z., Elamvazhuthi, K., & Berman, S. (2021). Using reinforcement learning to herd a robotic swarm to a target distribution. In *Distributed Autonomous Robotic Systems* (pp. 401–414). Springer. https://doi.org/10.1007/978-3-030-92750-9_30

Li, J., & Yang, S. X. (2024). Bio-inspired neural network for real-time evasion of multi-robot systems in dynamic environments. *Biomimetics, 9*(3), Article 176. https://doi.org/10.3390/biomimetics9030176

Li, Y., Jin, R., Xu, X., Qian, Y., Wang, H., Xu, S., & Wang, Z. (2022). A mobile robot path planning algorithm based on improved A* algorithm and dynamic window approach. *IEEE Access, 10*, 57736–57747. https://doi.org/10.1109/ACCESS.2022.3179397

Li, Y., Wang, H., Fan, J., & Geng, Y. (2022). A novel Q-learning algorithm based on improved whale optimization algorithm for path planning. *PLOS One, 17*(12), Article e0279438. https://doi.org/10.1371/journal.pone.0279438

Lin, N. (2025). Path planning of library management robot based on PDOACO algorithm. *IEEE Access*, *13*, 78376-78390. https://doi.org/10.1109/ACCESS.2025.3565519

Liu, L., Wang, X., Yang, X., Liu, H., Li, J., & Wang, P. (2023). Path planning techniques for mobile robots: Review and prospect. *Expert Systems with Applications, 227*, Article 120254. https://doi.org/10.1016/j.eswa.2023.120254

Low, E. S., Ong, P., & Cheah, K. C. (2019). Solving the optimal path planning of a mobile robot using improved Q-learning. *Robotics and Autonomous Systems, 115*, 143–161. https://doi.org/10.1016/j.robot.2019.02.013

Lytridis, C., Kaburlasos, V., Pachidis, T., Manios, M., Vrochidou, E., Kalampokas, T., & Chatzistamatis, S. (2021). An overview of cooperative robotics in agriculture. *Agronomy, 11*(9), Article 1818. https://doi.org/10.3390/agronomy11091818

Mao, P., Lv, S., & Quan, Q. (2025). Tube RRT*: Efficient homotopic path planning for swarm robotics passing-through large-scale obstacle environments. *IEEE Robotics and Automation Letters, 10*(3), 2247–2254. https://doi.org/10.1109/LRA.2025.3531151

Maoudj, A., & Hentout, A. (2020). Optimal path planning approach based on Q-learning algorithm for mobile robots. *Applied Soft Computing, 97*, Article 106796. https://doi.org/10.1016/j.asoc.2020.106796

Mohan, P., Sharma, L., & Narayan, P. (2021). Optimal path finding using iterative Sarsa. In *Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 811–817). IEEE. https://doi.org/10.1109/ICICCS51141.2021.9432345

Moosavi, S. K. R., Zafar, M. H., & Sanfilippo, F. (2024). Collaborative robots (cobots) for disaster risk resilience: A framework for swarm of snake robots in delivering first aid in emergency situations. *Frontiers in Robotics and AI, 11*, Article 1362294. https://doi.org/10.3389/frobt.2024.1362294

Peng, F., Liu, H., & Zheng, L. (2023). A Sarsa reinforcement learning hybrid ensemble method for robotic battery power forecasting. *Journal of Central South University, 30*(11), 3867–3880. https://doi.org/10.1007/s11771-023-5451-0

Phadke, A., & Medrano, F. A. (2024). Increasing operational resiliency of UAV swarms: An agent-focused search and rescue framework. *Aerospace Research Communications, 1*, Article 12420. https://doi.org/10.3389/arc.2023.12420

Puente-Castro, A., Rivero, D., Pedrosa, E., Pereira, A., Lau, N., & Fernandez-Blanco, E. (2024). Q-learning based system for path planning with unmanned aerial vehicles swarms in obstacle environments. *Expert Systems with Applications, 235*, Article 121240. https://doi.org/10.1016/j.eswa.2023.121240

Scharin, J., & Jansson, E. (2024). *Stigmergic interaction in robotic multiagent systems using virtual pheromones*. arXiv. https://doi.org/10.48550/arXiv.2401.12345

Singh, G., Lofaro, D. M., & Sofge, D. (2020). Pursuit-evasion with decentralized robotic swarm in continuous state space and action space via deep reinforcement learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART)* (Vol. 1, pp. 226–233). https://doi.org/10.5220/0008971502260233

Tan, J., Melkoumian, N., Harvey, D., & Akmeliawati, R. (2024). Evaluating swarm robotics for mining environments: Insights into model performance and application. *Applied Sciences, 14*(19), Article 8876. https://doi.org/10.3390/app14198876

Viseras, A., Losada, R. O., & Merino, L. (2016). Planning with ants: Efficient path planning with rapidly exploring random trees and ant colony optimization. *International Journal of Advanced Robotic Systems, 13*(5), Article 1729881416664078. https://doi.org/10.1177/1729881416664078

Wan, Y., Zhu, Z., Zhong, C., Liu, Y., Lin, T., & Zhang, L. (2025). Dynamic path planning for robotic arms based on an improved PPO algorithm. *Journal of System Simulation, 37*(6), 1462–1473. https://doi.org/10.16182/j.issn1004731x.joss.24-0122

Wei, D., Zhang, L., Liu, Q., Chen, H., & Huang, J. (2024). UAV swarm cooperative dynamic target search: A MAPPO-based discrete optimal control method. *Drones, 8*(6), Article 214. https://doi.org/10.3390/drones8060214

Yang, L., Li, P., Qian, S., Quan, H., Miao, J., Liu, M., Hu, Y., & Memetimin, E. (2023). Path planning technique for mobile robots: A review. *Machines, 11*(10), Article 980. https://doi.org/10.3390/machines11100980

Yildiz, B., Aslan, M. F., Durdu, A., & Kayabasi, A. (2024). Consensus-based virtual leader tracking swarm algorithm with GDRRT*-PSO for path-planning of multiple-UAVs. *Swarm and Evolutionary Computation, 88*, Article 101612. https://doi.org/10.1016/j.swevo.2024.101612

Zaghbani, I., Jarray, R., & Bouallegue, S. (2024). Comparative study of Q-learning and SARSA algorithms for UAV path planning in 3D environments. In *Proceedings of the 2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)* (pp. 245–250). IEEE. https://doi.org/10.1109/INES63318.2024.10629124

Zhou, Q., Lian, Y., Wu, J., Zhu, M., Wang, H., & Cao, J. (2024). An optimized Q-learning algorithm for mobile robot local path planning. *Knowledge-Based Systems, 286*, Article 111400. https://doi.org/10.1016/j.knosys.2024.111400