# Multi-Agent Reinforcement Learning in Collaborative Swarm Robotics: A Systematic Literature Review

Mahmut Osmanovic[1,2], Isac Paulsson[1,2]

***Abstract*—A systematic literature review (SLR) was conducted that investigates how multi-agent reinforcement learning (MARL)-based swarm robotic systems, and their extensions, contribute to improving collaboration and adaptability in search and rescue (SAR) missions. The utilization of swarm robotics within hazardous environments has the potential to reduce human exposure to danger. Recent research within the field has resulted in significant progress, however most research is conducted in simplified and static environments with small homogeneous swarms and single-stage objectives. For the review, twenty-four relevant articles from IEEE Xplore database were systematically gathered and analyzed. Identified key research gaps include the need for greater environmental fidelity, realistic communication constraints, and models that handle dynamic multi-objective tasks.**

## I. Introduction

Throughout nature, emergent swarm behavior allows relatively simple animals and insects to work collaboratively to solve complex tasks. The collaborative behaviors found in swarms do not rely on a centralized control structure. Instead, collaboration is decentralized and dynamic, governed by interactions at the individual level [1]. This forms the basis for an emerging field, swarm robotics, where agents with simple individual behavior collaborate to solve complex problems. The intersection between swarm robotics and multi-agent reinforcement learning (MARL) is a promising research direction, enabling agents to learn collaborative strategies similar to those observed in nature [2]. Hence, avoiding rigid control hierarchies and generating dynamic patterns of emergent behavior [3].

Search and rescue (SAR) operations provide a compelling application for swarm robotics [4]. In disaster environments, conditions are often uncertain, dynamic, and dangerous. Here, swarms of robots can be deployed to survey disaster sites, searching for victims and gathering real-time information [5]. The consequence is a reduction of the inherent risk found in SAR operations whilst simultaneously boosting search efficiency. To realize these capabilities, researchers have proposed a variety of multi-agent reinforcement learning models, designed to promote collaboration and adaptability in swarm robotics for search and rescue operations [6]. These approaches provide the foundation for this systematic literature review (SLR). This SLR is guided by the following research question (RQ):

**RQ1** How do multi-agent reinforcement learning swarm robotic systems, and their extensions, contribute to improving collaboration and adaptability in search and rescue missions?

[1]Jönköping School of Engineering, [2]Equal author contribution

This review identifies the need for greater environmental fidelity during simulation. Most research considers highly simplified static environments, without realistic constraints on communication. Additionally, further research is required on how agents can take uncertainty into account in more complex environments. In addition, future research should push for increased utilization of realistic environmental constraints and dynamics. With the purpose of driving development of highly adaptable agents, which are also deployable in representative real world experiments.

## II. Research Methodology

The research protocol followed in this systematic literature review is heavily inspired by the guidelines outlined by Barbara Kitchenham [7]. The protocol consists of six central pillars. The search process, inclusion/exclusion criteria, paper selection, quality assessment, data extraction and data analysis.

### A. Search Process

Given the specification of the technical research question and the quality assurance requirements of the review, research papers were sourced exclusively from IEEE Xplore. The reason being that the papers published in an IEEE journal or admitted into a conference are peer-reviewed. The final search query that was utilized to acquire the papers (which this SLR is based upon) is explicitly stated in table I. A total of 42 research papers were retrieved. The papers were subsequently downloaded in bulk and filtered through the application of the defined inclusion and exclusion criteria.

TABLE I
Search Query

```
    ("reinforcement learning" OR "RL")
AND ("multi-agent")
AND ("collaboration" OR "cooperation" OR "coordination")
AND ("swarm")
AND ("search" OR "rescue")
AND NOT ("survey" OR "review" OR "overview" OR "chapter")
```

There were four inclusion criteria and five exclusion criteria specified. The first inclusion criterion aids in ensuring a minimum standard for quality and novelty. The second criterion specify the languages in which the authors have an advanced user proficiency, in addition to the availability of research articles. The criterion is necessary as to ensure that a thorough understanding of the selected papers can be achieved. A thorough understanding enables in depth analysis of the material, which in turn strengthens the reliability of the drawn conclusions. The third criterion selects papers published within the past 18 years of research. It serves to provide comprehensive coverage of the relevant literature, resulting

in a holistic understanding of the domain. Lastly, the fourth criterion relates to the purpose of the paper. Its function is to aid in the alignment of purpose between the selected papers and ours.

Articles were included if they:

i. Peer reviewed
ii. Written in English
iii. Published between 09-27-2007 and 09-27-2025.
iv. The results focus on improving Multi-Agent Search Strategy, Collaboration, or Agent Adaptability

The first out of five exclusion criteria aims to exclude all papers that do not use MARL to address the research question. For instance, methods such as evolutionary computation or game theoretic based modeling. In criteria two, papers that base their models on imitation learning (a subfield of reinforcement learning) are excluded. Imitation learning does not rely on reward driven interactions with the environment. That is a defining trait of the MARL driven paradigm within swarm robotics. Exclusion criteria three and four filter out papers that do not adhere to our definition of swarm robotics. Criteria five excludes papers whose models or frameworks are deemed too elementary. These often include foundational techniques that many if not close to all of subsequent techniques rely upon.

Articles related to the following were excluded:

i. Model does not include RL (instead model is based on e.g., Evolutionary Computation, Game Theory, etc)
ii. Imitation Learning
iii. Centralized Task Allocation
iv. Static Centralized Communication Structure
v. Foundational or elementary models or frameworks

### B. Paper Selection

Using the previously defined inclusion and exclusion criteria, an inter-rater kappa analysis was performed on the collection of gathered 42 peer-reviewed research papers. The analysis was performed in two sweeps. The first sweep applied the inclusion and exclusion criteria exclusively to the titles and abstracts of all papers. The authors applied the defined criteria independently and sorted papers into one of two categories, "include" or "exclude". The end result of that procedure is detailed in table II-B. The authors managed to commonly include or exclude 40/42 papers, only disagreeing on two.

Next, "Cohen's kappa" was calculated to quantify the level of agreement between the two reviewers, beyond what would be expected by chance.

A high kappa value is indicative of clearly defined inclusion and exclusion criteria. That is since different reviewers interpreted and applied them similarly. Note the high kappa value, $\kappa_1 = 0.903$, after the first sweep. The derived confidence interval denotes $95\%$ confidence that the "true" $\kappa$ lies within the specified range ([0.754, 1.000]).

The second sweep was conducted on papers that had been mutually agreed upon for inclusion, as well as those on which disagreement remained. Contrary to the first sweep, the second sweep was performed on the full research articles. The second sweep resulted in complete agreement, as signified by $\kappa_2 = 1.000$ (II-B).

TABLE II
SUMMARY OF INTER-RATER RELIABILITY ACROSS SCREENING STAGES

| Screening Stage | #Screen Papers | Kappa ($\kappa$) | 95% CI |
|---|---|---|---|
| Title/Abstract Screening | 42 | $\kappa_1$: 0.903 | [0.754, 1.000] |
| Full-text Screening | 25 | $\kappa_2$: 1.000 | [1.000, 1.000] |

### C. Quality assessment

The quality of the 24 selected papers was assessed independently by both authors according to the eight specified criteria detailed in table III. These criteria were chosen with orthogonality in mind, aiming to capture different and complementary aspects of paper quality as broadly as possible. Each paper could obtain a final quality score between zero and eight points.

Figure II-C highlights the quality of assessments of utilized journals in virtue of their impact factor. Pearson's correlation coefficient yields a modest $r \approx 0.30$ between average quality score and journal impact factor (excluding IEEE Trans. Intell. Veh. journal). Nonetheless, the low quantity of journals does not enable us to make a generalized conclusion about it.

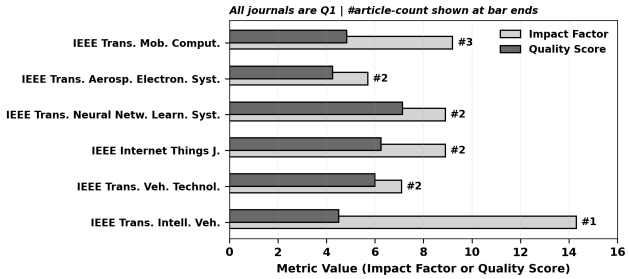


Fig. I. Exclusively includes the 13 journal papers. Grey bars represent the average quality score of the #articles papers from that journal. Impact factor for each journal is retrieved from IEEE Xplore [8]. All journals are Q1, signifying high qualitative value and methodological rigor. Journal rankings are retrieved from Scimago Journal & Country Rank [9].

Among the remaining papers, one was published in $A^*$ conference $ICRA$ [10]. Another conference paper was published in *IEEE International Conference on Automation Science and Engineering* which as of October 11, 2025 is yet to be ranked. The remaining conferences, although IEEE affiliated have not obtained an official ranking. The papers were nonetheless accepted by IEEE Conference Publishing Services (CPS) for inclusion in IEEE Xplore. This at least establishes that the conference adhered to IEEE's base criteria for publication. Specifically, a peer-review process, IEEE standard formatting requirements and scans for plagiarism [11].

Table IV quantifies the per quality criterion score agreement between the two authors. Specifically, the quantification measure is the Weighted Cohen's Kappa [12], using both linear and quadratic weighting. For ordinal categories (0, 0.5, and 1), linear weights penalize disagreement in direct proportion

to category distance, whereas quadratic weights penalize small disagreements less and large disagreements more strongly. Additionally, the $95\%$ confidence intervals quantify the error bound within two standard deviations.

TABLE III
QUALITY ASSESSMENT CRITERIA

| Category | Description |
|---|---|
| Problem formulation | Clear task definition, assumptions, agent roles, objectives (team-level and agent-level). |
| Environment fidelity | Is the environment a gross simplification of reality? [0p] Is the simulation a decent representation of reality? [0.5p] Was *sim-to-real* experimentation conducted? [1p] |
| Evaluation metrics | Collaboration/coordination metrics (e.g., throughput, time-to-rescue, coverage, collisions, bottleneck utilization) plus task success. |
| Baselines & Ablations | Comparison against strong baselines (non-MARL, centralized, heuristic) and ablations (e.g., communication disabled, reward variants, credit assignment). |
| Statistical robustness | Just run once? [0p] run and average over k iterations? [0.5p] performs statistical tests? [1p] |
| Generalization | Tested across maps, scenarios, team sizes, or heterogeneity; if simulation-only, evidence of robustness to noise, latency, failures. |
| Reproducibility | Code, data, configs, or sufficient detail to reproduce; compute budget and hyperparameters reported. |
| Threats to validity | Internal/external validity is explicitly discussed. |
| Scoring | [0p, 0.5p, 1p] assigned to each category. |

Two noteworthy insights can be drawn from table IV. First, $\kappa$ is at least approximately $0.90$ for all categories, this indicates almost perfect agreement in quality scoring between authors [13]. Secondly, the quadratically weighted $\kappa$ is generally greater than or equal to the linear $\kappa$. This implies that even when disagreements occurred, they were of a lesser magnitude. In addition, both observations are indicative of precisely formalized quality criteria, contributing to the reproducibility of this SLR.

Normalized scores for each author and paper were calculated. Subsequently, the by author normalized quality scores were themselves averaged to yield the final quality score.

### D. Data Extraction & Synthesis

The papers were systematically reviewed, and the following information was extracted: Paper (Shortened Title, Publication Year), Environment Fidelity (2D/3D, Continuous, Dynamics), Primary Objective, Communication Assumptions, Swarm Size, Swarm Diversity, RL Model (name, base). The extracted information is presented in table V together with the quality scores from the performed quality assessment.

## III. RESULTS & ANALYSIS

Agent behavior is influenced by communication, extending or altering communications can be beneficial for agent performance. Communication can be improved using GAT, convolution or attention based aggregation. Agents with these types of aggregation get access to filtered information. This information is either localized or of high importance to their current state. The results show improved performance for search and coordination [10, 17, 26, 27, 31, 35]. Another approach is to utilize pheromones, a bio-inspired communication mechanism. With this method pheromones can be distributed and sensed by the agents in the environment. The addition of pheromones aid coordination and lessens the reliance on other forms of communication [15, 16, 19, 32]. There are other considerations to make relating to communication. In simulations, it is easy to assume idealized communication. When transferring from simulation to reality this assumption cannot be justified. The addition of realistic signal modeling (Attenuation, Line of Sight, Path loss, Range) allows for testing in conditions that accurately reflect physical communication [16, 17, 23, 25, 26]. Consequently, adaptability and resilience can better be developed and evaluated [23, 25, 26].

TABLE IV
COMPARISON OF LINEAR VS. QUADRATIC KAPPA RESULTS WITH 95% CONFIDENCE INTERVALS ACROSS EVALUATION CATEGORIES

| Category | Linear $\kappa$ (95% CI) | | Quadratic $\kappa$ (95% CI) | |
|---|---|---|---|---|
| Baselines & ablations | 0.933 | (0.768, 1.000) | 0.944 | (0.813, 1.000) |
| Environment fidelity | 1.000 | (1.000, 1.000) | 1.000 | (1.000, 1.000) |
| Evaluation metrics | 0.892 | (0.625, 1.000) | 0.907 | (0.625, 1.000) |
| Generalization | 0.947 | (0.818, 1.000) | 0.960 | (0.865, 1.000) |
| Problem formulation | 1.000 | (1.000, 1.000) | 1.000 | (1.000, 1.000) |
| Reproducibility | 1.000 | (1.000, 1.000) | 1.000 | (1.000, 1.000) |
| Statistical robustness | 1.000 | (1.000, 1.000) | 1.000 | (1.000, 1.000) |
| Threats to validity | 0.933 | (0.779, 1.000) | 0.941 | (0.802, 1.000) |
| **Average** | **0.963** | **(0.874, 1.000)** | **0.969** | **(0.888, 1.000)** |

Environments vary greatly in fidelity across papers. With approximately $50\%$ of papers utilizing simplified 2D grid worlds. Furthermore, environments do not tend to dynamically change over time, with the exception of moving targets. In contrast, real applications are highly dynamic. Due to the architecture in MARL based swarm robotics, papers often conjecture real-time adaptation to dynamic environmental changes without justification. Nonetheless, these assumptions do still require validation [10, 23, 29, 30, 32, 33]. The current problem formulations mainly consider single stage tasks and short planning horizons. In reality, a search and rescue operation is not over when the target is found. This simplification overlooks critical complexities for success in real applications. Recent works consider the extended case where the target also has to be rescued by a separate entity after being located [19, 25]. Zhao et al. propose a two stage framework composed of a detection stage and an information aggregation stage, where agents have to alternate between stages in order to propagate information between each other [27].

| Paper | Environment | Objective(s) | Comm. Assumption | Swarm Size | Swarm Diversity | RL-Model | Quality |
|---|---|---|---|---|---|---|---|
| Decentralized Coop. Coverage Control (UAVs) (2023) [14] | Static 2D grid | Coverage | Idealized | 3 | Homogeneous | EPPO (PPO) | 2.5/8 |
| Deep-Q Learning Connectivity Aware UAVs (2024) [15] | Static 2D grid | Connectivity Maintenance | Limited (range) | 30-50 | Homogeneous | DQL (DQN) | 4.5/8 |
| Multi-AUV Coop. Search (Acoustic-Optical) (2024) [16] | Static 2D cont. | Target Search | Limited (acoustic-optical) (attenuation) | 3-7 | Homogeneous | AOSA (DDPG) | 4.0/8 |
| AAV Swarm Coop. Search (Digital Twin) (2025) [17] | Static 3D grid | Target Search | Limited (range) (complexity) | 1-40 | Homogeneous | SAMARL (PPO) | 5.0/8 |
| Age of Info-Aware Multi-Objective (UAV-USV-UUV) (2025) [18] | Static 3D cont. | Target Search | Idealized | 5 | Heterogeneous | AE-MVTD3 (DDPG) | 4.0/8 |
| Bio-Inspired Multi-Agent DQN (2024) [19] | Static 2D grid | Search & Rescue | Limited (pheromone) (diffusion) | 5-6 | Homogeneous | DQN | 2.5/8 |
| Collaborative Search & Tracking (Dynamic Map) (2024) [20] | Static 2D grid | Target Search | Limited (range) | 10 | Homogeneous | MAAC (AC) | 4.0/8 |
| Collaborative Search Planning (2023) [21] | Static 2D grid | Target Search | Idealized | 2-5 | Homogeneous | QMIX | 3.5/8 |
| Collaborative Target Search (Visual Drone) (2025) [22] | Static 3D cont. | Target Search | Limited (relative localization) | 1-3 | Homogeneous | PPO-AEC (PPO) | 7.5/8 |
| Deep RL for Decentralized Multi-Robot Exploration (2023) [23] | Static 2D grid and 3D cont. | Exploration | Limited (dropouts) | 2-6 | Homogeneous | MADE-Net (DDRQN) | 5.75/8 |
| DRL-Searcher (2024) [24] | Dynamic Graph | Target Search (moving target) | Idealized | 2-4 | Homogeneous | Distributional QL | 6.75/8 |
| Dynamic Task Allocation (PG-MAPPO) (2025) [25] | Dynamic 3D cont. | Search & Rescue | No communication | 45-100 | Homogeneous | PG-MAPPO (PPO) | 7.0/8 |
| Heterogeneous UAVs Traj Opt (2025) [26] | Static 3D cont. | Target Search | Limited (signal noise) (attenuation) | 6-32 | Heterogeneous | GATAC (AC) | 6.5/8 |
| Integrated RL Framework (Two-Stage) (2025) [27] | Dynamic 2D grid | Target Search and Information Aggregation | Limited (range) | 5-15 | Homogeneous | STDGNNet (TDGN) | 5.5/8 |
| Learning Collab. Multi-Target Search (Visual) (2023) [28] | Static 3D cont. | Target Search | No communication | 2-3 | Homogeneous | POCA-Mix (MA-POCA) | 4.0/8 |
| Learning to Routing in UAV Swarm Networks (2023) [29] | Static 3D cont. | Routing Optimization | Limited (bandwidth) (range) | 30 | Homogeneous | AC | 5.5/8 |
| MADRL Soft Actor-Critic (Self-Collab UAVs) (2024) [30] | None | Swarm Coordination | No communication | Undefined | Undefined | SAC | 1.0/8 |
| Multi-Agent Multi-Target Search (2025) [31] | Dynamic 2D grid | Target Search (moving target) | No communication | 8 | Homogeneous | MAAC (AC) | 4.25/8 |
| Multi-UAV Marine Search (2023) [32] | Dynamic 3D cont. | Target Search (moving targets) | Limited (pheromone) (diffusion) | 3 | Homogeneous | MADDPG (DDPG) | 3.0/8 |
| Proficiency Constrained UAV-UGV Teaming (2021) [33] | Dynamic 3D cont. | Target Search (moving target) | Idealized | 2-4 | Heterogeneous | Mix-RL (AC) | 3.25/8 |
| REPlanner (Economic RL) (2021) [34] | Static 2D grid and 3D cont. | Trajectory planning | Idealized | 3-9 | Homogeneous | Tabular QL | 4.0/8 |
| Spatial Intention Maps (2021) [10] | Static 3D cont. | Foraging and Search & Rescue | Idealized | 4 | Heterogeneous | DDQN | 6.5/8 |
| Two-Level Attention Trajectory Design (2025) [35] | Dynamic 3D cont. | Target Search | Limited (line of sight) (attenuation) | 2-11 | Homogeneous | MAPPO with Attention (PPO) | 5.5/8 |
| UAV Swarm Coop Target Search (2024) [36] | Dynamic 2D grid | Target Search | No communication | 3-10 | Homogeneous | MADDPG (DDPG) | 4.5/8 |

TABLE V

CONTRIBUTION SUMMARY OF REVIEWED SWARM REINFORCEMENT LEARNING PAPERS, CATEGORIZED BY ENVIRONMENT TYPE, OBJECTIVES, COMMUNICATION ASSUMPTIONS, SWARM SIZE, HETEROGENEITY, MODEL, AND QUALITY SCORE.

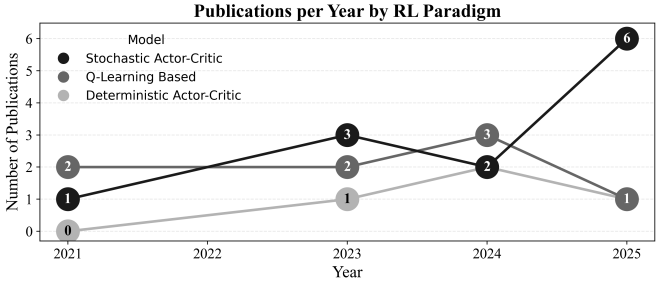


Fig. II. Popularity of RL paradigms over time: Stochastic Actor-Critic overtakes other approaches by 2025, with Q-Learning-based and Deterministic Actor-Critic showing flatter or declining trends.

Another finding from the review is the prevalence of small scale swarms. The majority of reviewed work utilizes less than 10 agents. This trend is likely caused by the computational complexity incurred when increasing the number of agents in traditional architectures. Notably a subset of the literature is advancing the frontier in terms of scalability. Wu et al. [25] demonstrate dynamic task allocation with up to 100 agents. Cao et al. [17] and Devaraju et al. [15] experiment with swarms of up to 40 and 50 agents, respectively. The methods employed utilize localized interactions (e.g., pheromones or neighbor-to-neighbor message passing) or learned information relevance (e.g., attention mechanisms) to reduce large scale information exchange while maintaining agent performance. Furthermore, Devaraju et al. [15] model agent failures that

dynamically alter swarm size. In the reviewed literature, few papers included agents with differing capabilities (heterogeneous swarms) [10, 18, 26, 33], the majority did not. A coalition with diverse agent capabilities may prove to be more advantageous than homogeneous swarms.

Despite only including January to September, the year 2025 stands out with the most publications (figure II). This is indicative of a growing interest in the field. Additionally, the popularity of Stochastic Actor-Critic based models increased in comparison to others.

## IV. Research Gaps

Even though reinforcement learning has made substantive contributions to swarm robotics, most of the reviewed studies were conducted in close to static environments. However, physical reality is permeated by uncertainties. This puts into question the adaptive capabilities of MARL systems in dynamic swarm robotic applications. A commonly found example of this is the focus on optimizing for a single objective. More realistic cases would likely require multi-stage and multi-objective MARL models to account for dynamics pertinent to the real world. As discussed previously, few studies incorporate multi-stage objectives or multi-objective scenarios [19, 25, 27], and none incorporate both.

Another topic relevant to adaptability is that of communication, studies scarcely impose realistic communication constraints in their environments [16, 17, 23, 25, 26]. Given further dynamic communication constraints, current models may need further capabilities to efficiently adapt. Regarding the uncertainty of realistic scenarios, the utilization of Bayesian techniques for quantifying uncertainty in internal representation is sparse (e.g., spatial intention by Wu et al. [10]). One such Bayesian framework is Friston's Free Energy Principle [37].

For the purpose of utility maximization, it is hence exceedingly important to reflect real world uncertainties and dynamics in future environment simulations. Yet, simulations (*in silico* experimentation) alone can not promise to fully capture such complexities. Thus, in an attempt to reinforce external validity, real world experimentation is equally indispensable.

## V. Proposed Study Outline

In order to bridge the gap between explored papers and identified research gaps (Section IV), we propose a study outline for future research. The aim would be to develop a multi-stage, uncertainty aware MARL model for heterogeneous swarm robotics that operates within a dynamic environment with the objective of search and rescue. There are several viable and interesting research questions. The one we propose is: *Does Bayesian uncertainty modeling improve adaptability under communication constraints?* Similarly, there are plenty of diverse environment and methodological configuration that could comprise compelling future studies. Our proposed environment should be dynamic, 3D, continuous, with varying hazards, and moving targets. Several such environments are to be constructed using a suitable simulation framework (e.g., *PyBullet*). The swarm should be heterogeneous and composed of 5-25 mixed UGV agents with varying capabilities (akin to [10]). The proposed learning framework is multi-agent PPO with Bayesian latent uncertainty modeling. Information sharing between agents should be limited with constraints on bandwidth and path loss. Drummond et al. [38] argues that more than a single performance measure probably is needed for scientific research to progress. Hence, there are several suggested evaluation metrics. Specifically, time-to-rescue, communication intensity, robustness to noise, and coverage efficiency. As a baseline the regular PPO model can be utilized.

A thorough ablation study is to be conducted, maximally varying one core component at a time. Additionally experiments should be conducted varying swarm size in steps of 5 agents and also varying team capabilities (including the homogeneous cases). Each experiment should be repeated 30 times, the mean and standard deviation for each metric should be recorded across runs. For validation a Bayesian t-test should be performed on the recorded data [39]. The Bayesian t-test compares the model with Bayesian uncertainty against the one without. The results of the Bayesian t-test yields the probability of one model being better than the other with respect to a set metric, hence providing a direct answer to the stated research question.

Finally, the environments should be physically constructed and the model results with and without Bayesian uncertainty awareness capabilities validated beyond simulation. A similar Bayesian t-test statistic should be performed for the sake of statistical robustness. Threats to internal validity may arise from bias in environment setup or in selection of RL model. External validity may be limited due to constrained real world experimentation, leaving out large scale constellations of swarms.

## VI. Study Limitations

Relying solely on IEEE Xplore for the SLR imposes limitations on the comprehensiveness of the search. Whilst IEEE Xplore is a prominent database in the field of computer science, there are several other databases that could have provided relevant articles not indexed in IEEE Xplore.

As noted by Brambilla et al. [40], the foraging problem can be seen as a generalization of the SAR problem, since both problems share many aspects. Hence, further insights into SAR may be gained from approaches utilized in the foraging problem. Additionally, our study reflects an intersection of accessible research. The coverage scope depends on the chosen search string, making it difficult to quantify what studies were left out, and is thus limiting external validity. A more inclusive search string could have improved coverage but would likewise require more processing time and would potentially retrieve additional irrelevant papers.

Lastly, internal validity may be threatened by the fact that only two researchers extracted the data, which could potentially lead to human errors and biased results.

## References


[1] J. Zhang, Q. Qu, and X. Chen. Understanding collective behavior in biological systems through potential field mechanisms. *Scientific Reports*, 15:3709, 2025.

[2] Jonas Kuckling. Recent trends in robot learning and evolution for swarm robotics. *Frontiers in Robotics and AI*, 10:1134841, 2023.

[3] Zhenyu Li, Yuzhe Zhang, Xiaotian Guo, and Jun Wang. Predator-prey survival pressure is sufficient to evolve swarming behaviors. *arXiv preprint arXiv:2308.12624*, 2023.

[4] David S. Drew. Multi-agent systems for search and rescue applications. *Robotics Reports*, 5(1):1–20, 2021.

[5] R. D. Arnold, H. Yamaguchi, and T. Tanaka. Search and rescue with autonomous flying robots through behavior-based cooperative intelligence. *International Journal of Humanitarian Action*, 3(1):18, 2018.

[6] M. Hüttenrauch, A. Šošić, and G. Neumann. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54):1–31, 2019.

[7] Barbara Kitchenham and Pearl Brereton. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12):2049–2075, December 2013.

[8] IEEE. Ieee xplore: Browse periodicals by title. https://ieeexplore.ieee.org/browse/periodicals/title, 2025. Accessed: 2025-10-11.

[9] SCImago Lab. Sjr — scimago journal & country rank [portal]. https://www.scimagojr.com/journalrank.php, n.d. Accessed: 2025-10-11.

[10] Jimmy Wu, Xingyuan Sun, Andy Zeng, Shuran Song, Szymon Rusinkiewicz, and Thomas Funkhouser. Spatial intention maps for multi-agent mobile manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8749–8756, 2021.

[11] IEEE Computer Society. Conference publishing services (cps). https://www.computer.org/conferences/cps, 2025. Accessed: 2025-10-11.

[12] Ming Li, Qiong Gao, and Tzu-Kuei Yu. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer*, 23:799, 2023.

[13] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

[14] Longbo Cheng, Guixian Qu, Jianshan Zhou, Dezong Zhao, Kaige Qu, Zhengguo Sheng, Junda Zhai, and Chenghao Ren. A decentralized cooperative coverage control for networked multiple uavs based on deep reinforcement learning. In *2023 IEEE International Conference on Unmanned Systems (ICUS)*, pages 205–210, 2023.

[15] Shreyas Devaraju, Alexander Ihler, and Sunil Kumar. A deep-$q$-learning-based base-station-connectivity-aware decentralized pheromone mobility model for autonomous uav networks. *IEEE Transactions on Aerospace and Electronic Systems*, 60(6):8682–8699, 2024.

[16] Xiang Li, Peijun Dong, Hang Tao, Pengyan Dong, Zhijie Feng, and Hanjiang Luo. A multi-auv cooperative search scheme based on acoustic-optical communication and deep reinforcement learning. In *2024 IEEE International Conference on High Performance Computing and Communications (HPCC)*, pages 785–790, 2024.

[17] Pan Cao, Lei Lei, Gaoqing Shen, Shengsuo Cai, Xiaojiao Liu, and Xiaochang Liu. Aav swarm cooperative search based on scalable multiagent deep reinforcement learning with digital twin-enabled sim-to-real transfer. *IEEE Transactions on Mobile Computing*, 24(6):5173–5188, 2025.

[18] Xiangwang Hou, Tianyu Xing, Jingjing Wang, Jun Du, Chunxiao Jiang, Yong Ren, and Dusit Niyato. Age of information-aware multi-objective optimization for heterogeneous uav-usv-uuv networks in underwater target hunting. *IEEE Transactions on Mobile Computing*, 24(11):11292–11304, 2025.

[19] Zhengkun Chen. Bio-inspired multi-agent dqn for maritime and underwater search and rescue. In *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 112–119, 2024.

[20] Yuechen Ge and Jianjun Ni. Collaborative search and tracking based on dynamic map and information fusion for uav swarm. In *2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 245–249, 2024.

[21] Liangji Zou and Yihua Tan. Collaborative search planning of uav swarms based on deep reinforcement learning. In *2023 5th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP)*, pages 1184–1187, 2023.

[22] Jiaping Xiao, Phumrapee Pisutsin, and Mir Feroskhan. Collaborative target search with a visual drone swarm: An adaptive curriculum embedded multistage reinforcement learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):313–327, 2025.

[23] Aaron Hao Tan, Federico Pizarro Bejarano, Yuhan Zhu, Richard Ren, and Goldie Nejat. Deep reinforcement learning for decentralized multi-robot exploration with macro actions. *IEEE Robotics and Automation Letters*, 8(1):272–279, 2023.

[24] Hongliang Guo, Qihang Peng, Zhiguang Cao, and Yaochu Jin. Drl-searcher: A unified approach to multirobot efficient search for a moving target. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3215–3228, 2024.

[25] Xiang Wu, Qingzhong Yan, Jiacun Wang, Yuhang Zhou, Qilong Huang, and Changhui Jiang. Dynamic task allocation for uav swarms in maritime rescue scenarios based on pg-mappo. *IEEE Internet of Things Journal*, 12(18):38073–38087, 2025.

[26] Tianyong Ao, Haoqiang Li, Kaixin Zhang, Huaguang Shi, Lei Shi, Fuqiang Liu, and Yi Zhou. Heterogeneous uavs trajectory optimization for post-disaster target

search based on marl with graph attention network. *IEEE Transactions on Vehicular Technology*, pages 1–14, 2025.

[27] Yijing Zhao, Sanqiang Lu, Chao Wang, Yumeng Liu, Yi Ding, and Hongan Wang. Integrated reinforcement learning framework for uav swarm two-stage cooperative multitarget detection tasks. *IEEE Internet of Things Journal*, 12(8):9435–9448, 2025.

[28] Jiaping Xiao, Yi Xuan Marcus Tan, Xinliang Zhou, and Mir Feroskhan. Learning collaborative multi-target search for a visual drone swarm. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 5–7, 2023.

[29] Zunliang Wang, Haipeng Yao, Tianle Mai, Zehui Xiong, Xiaohua Wu, Di Wu, and Song Guo. Learning to routing in uav swarm network: A multi-agent reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 72(5):6611–6624, 2023.

[30] Alexander Pascual and Soo Young Shin. Multi-agent deep reinforcement learning based on soft actor-critic for self-collaborating uavs in a swarm. In *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 346–348, 2024.

[31] Huiqin Pei and Zilong Luo. Multi-agent multi-target search with multi-head attention. In *2025 37th Chinese Control and Decision Conference (CCDC)*, pages 4035–4040, 2025.

[32] Kai Guo, Hanjiang Luo, Hang Tao, Rukhsana Ruby, Zhizun Qin, and Kui Liu. Multi-uavs collaborative search scheme in marine environments using deep reinforcement learning. In *2023 Eleventh International Conference on Advanced Cloud and Big Data (CBD)*, pages 39–44, 2023.

[33] Qifei Yu, Zhexin Shen, Yijiang Pang, and Rui Liu. Proficiency constrained multi-agent reinforcement learning for environment-adaptive multi uav-ugv teaming. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pages 2114–2118, 2021.

[34] Alvi Ataur Khalil, Alexander J Byrne, Mohammad Ashiqur Rahman, and Mohammad Hossein Manshaei. Replanner: Efficient uav trajectory-planning using economic reinforcement learning. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 153–160, 2021.

[35] Haowen Zhu, Junpeng Hui, and Zehua Guo. Two-level-attention-based continuous trajectory design and computation offloading for multi-uav cooperative target search. *IEEE Transactions on Mobile Computing*, pages 1–18, 2025.

[36] Yukai Hou, Jin Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. Uav swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Transactions on Intelligent Vehicles*, 9(1):568–578, 2024.

[37] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[38] Chris Drummond and Nathalie Japkowicz. Warning: Statistical benchmarking is addictive. kicking the habit in machine learning. *J. Exp. Theor. Artif. Intell.*, 22:67–80, 03 2010.

[39] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18:1–36, August 2017. Submitted 06/16; Revised 05/17; Published 08/17.

[40] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7(1):1–41, 2013.