ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προσαρμοστική πλοήγηση σμήνους με χρήση ενισχυτικής μάθησης

ΑΘΑΝΑΣΙΟΣ ΚΟΥΚΟΣΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία 2025- 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προσαρμοστική πλοήγηση σμήνους με χρήση ενισχυτικής μάθησης

ΑΘΑΝΑΣΙΟΣ ΚΟΥΚΟΣΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία 2025-2026

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

Adaptive swarm navigation with the use of reinforcement learning

ATHANASIOS KOUKOSIAS

FINAL THESIS

ADVISOR

Kolomvatsos Konstantinos
Associate Professor

Lamia 2025-2026

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ……/…../2026

Ο – Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Σε μια εποχή όπου η νοημοσύνη κατανέμεται σε πολλαπλούς συνεργαζόμενους πράκτορες, ο τρισδιάστατος χώρος γίνεται ένα δυναμικό περιβάλλον που απαιτεί προσαρμοστικότητα και συλλογική λήψη αποφάσεων. Παρόμοιες συμπεριφορές παρατηρούνται στη φύση, όπως σε σμήνη πτηνών και κοπάδια ψαριών, όπου η επιβίωση σε αβέβαιες και ταχέως μεταβαλλόμενες συνθήκες εξαρτάται από την αποτελεσματική τοπική αντίληψη και την προσαρμοστική απόκριση. Αυτές οι αρχές παρακινούν το σχεδιασμό τεχνητών συστημάτων ικανών για ισχυρό συντονισμό σε στοχαστικά περιβάλλοντα.

Αυτή η εργασία μεταφράζει αυτές τις έννοιες σε ένα αλγοριθμικό πλαίσιο για την εκπαίδευση τεχνητών πρακτόρων που αλληλοεπιδρούν τόσο με το περιβάλλον τους όσο και μεταξύ τους. Η ενισχυτική μάθηση παρέχει τη βάση για την προσαρμοστική επιλογή πολιτικής, επιτρέποντας στους πράκτορες να μάθουν όχι μόνο πώς να κινούνται αλλά και ποια στρατηγική πλοήγησης είναι η καταλληλότερη σε κάθε στιγμή.

Η προτεινόμενη προσέγγιση ενσωματώνει έναν βελτιωμένο αλγόριθμο σμήνους Boids με δυναμικά προφίλ σχηματισμού, επιτρέποντας στους πράκτορες να προσαρμόζουν τη συμπεριφορά τους μεταβαίνοντας μεταξύ επιθετικών, ουδέτερων και αμυντικών λειτουργιών με βάση την περιβαλλοντική ανατροφοδότηση. Αυτό επιτρέπει την αποτελεσματική εξερεύνηση διατηρώντας παράλληλα τη συνοχή και την ευρωστία.

Μετά την εξερεύνηση, ένας ισορροπημένος αλγόριθμος σχεδιασμού διαδρομής βασισμένος στον αλγόριθμο A* καθοδηγεί την πλοήγηση μέσω των περιοχών που έχουν ήδη ανακαλυφθεί, βελτιστοποιώντας από κοινού την αποδοτικότητα της διαδρομής και την περιβαλλοντική ασφάλεια. Τα πειραματικά αποτελέσματα καταδεικνύουν βελτιωμένη κάλυψη, αποτελεσματική επιλογή προσαρμοστικού σχηματισμού και αξιόπιστη απόδοση πλοήγησης, υπογραμμίζοντας την εφαρμοσιμότητα της προσέγγισης στην αυτόνομη ρομποτική, την νοημοσύνη σμήνους και τα κατανεμημένα δίκτυα αισθητήρων.

ABSTRACT

In an era where intelligence is distributed across multiple cooperating agents, three-dimensional space becomes a dynamic environment requiring adaptability, and collective decision-making. Similar behaviors are observed in nature, such as in flocks of birds and schools of fish, where survival in uncertain and rapidly changing conditions depends on effective local perception and adaptive response. These principles motivate the design of artificial systems capable of robust coordination in stochastic environments.

This thesis translates these concepts into an algorithmic framework for training artificial agents that interact with both their environment and each other. Reinforcement learning provides the foundation for adaptive policy selection, enabling agents to learn not only how to move but also which navigation strategy is most appropriate at each moment.

The proposed framework integrates an enhanced Boids Flocking Algorithm with adaptive formation profiles, allowing agents to dynamically adjust their behavior by transitioning between Offensive, Neutral, and Defensive modes based on environmental feedback. This enables efficient exploration while maintaining cohesion and robustness.

Following exploration, a balanced A*-based path planning algorithm guides navigation through discovered regions, jointly optimizing path efficiency and environmental safety. Experimental results demonstrate improved coverage, effective adaptive formation selection, and reliable navigation performance, highlighting the applicability of the approach in autonomous robotics, swarm intelligence, and distributed sensor networks.

Table of Contents

CHAPTER 1
INTRODUCTION

In recent years, the deployment of Edge Computing (EC) nodes has garnered significant attention, primarily due to advancements in robotics, artificial intelligence, and other fields (Huh & Seo, 2019). These systems are increasingly essential for applications requiring coordinated interactions among multiple autonomous nodes, particularly in dynamic environments. EC nodes are now integral to a broad spectrum of domains, including disaster response, surveillance and defense (Sakano et al., 2025). Effective coordination among these nodes is critical for mission success i.e., achieving the optimal solution to an optimization problem. The ability of nodes to maintain well organized formations as they dynamically adapt to their surroundings enhances efficient task execution (Bansla & others, 2024). This involves determining an optimal position for the entire swarm within the operational environment and adopting a formation that maximizes positional advantages while searching through the focused space.

This research examines a range of formation strategies employed by autonomous nodes in dynamic environments. Four distinct formation profiles are evaluated: the **Offensive Formation**, characterized by reduced cohesion and moderate separation, enabling efficient exploration and target engagement. The **Neutral Formation**, balancing cohesion and alignment for adaptability and flexibility, suitable for general purpose tasks such as scouting and communication. The **Defensive Formation**, featuring increased cohesion and minimized separation to protect critical nodes from high threat levels. Finally, the **Dynamic Formation**, which enables adaptive switching between formations in response to perceived threat intensity.

One significant research challenge that emerges during the optimal position exploration phase is identifying the most effective formation for the swarm and the corresponding movements in the search space. To address this, this study proposes a methodology for dynamically forming the swarm during search activities and guiding its movements based on the parameters of the underlying problem. In the initial section, the focus is on the **Optimal Position Scouting** (OPS) problem, which is fundamental to swarm based systems operating in stochastic environments.

The navigation capabilities of a dynamically adapting swarm in uncertain environments pose significant challenges. In such scenarios, traditional path planning algorithms may not be sufficient to ensure safe movement. Thus, the scope of this research later shifts towards the evaluation of the pathfinding capabilities within a comprehensive assessment methodology. A range of four path planning algorithms are investigated: **Dijkstra**, **A***, **Safety-First**, and the proposed **Balanced** approach. The performance of these methods is assessed through tens of thousands of independent experiments, which take into account varying environmental parameters, start-to-goal pairs, and random seeds to ensure statistical validity.

CHAPTER 2
FUNDAMENTALS OF SWARM BEHAVIOR

(Section 2.1) Biological Collective Behavior

Swarm behavior is one of the most observed collective phenomena in nature, emerging in biological systems where large numbers of individuals coordinate through local interactions without centralized control. Such collective motion appears in bird flocks, fish schools, and insect colonies. There, groups exhibit highly organized and adaptive behavior despite each individual possessing only limited perception of its surroundings. These systems demonstrate how decentralized decision making and local communication can produce global rules for adaptation to dynamic environments.

One of the most recognizable examples of collective behavior is observed in flocks of birds. During flight, birds continuously adjust their velocity and direction according to the motion of nearby flock members, allowing the flock to maintain cohesion while avoiding collisions. This collective coordination enables rapid directional changes, predator avoidance, and energy efficient migration. Similar behavior is observed in schools of fish, where local interactions between neighboring individuals produce synchronized movement patterns that improve protection against predators. In both cases, collective motion emerges not from centralized planning but from simple local interaction rules based on alignment, separation, and cohesion.

Collective intelligence is also visible in insect societies such as ants and bees. Ant colonies demonstrate path optimization through pheromone based communication. There, individuals coordinate by reinforcing efficient routes toward resources. Bee swarms exhibit distributed decision making when selecting new hive locations, with individuals sharing local observations until consensus emerges across the colony. These biological systems illustrate how large groups can collectively solve complex tasks while maintaining robustness against environmental uncertainty and individual failures.

The study of such natural systems has significantly influenced the development of computational swarm intelligence models. By abstracting the local interaction mechanisms observed in nature, researchers have developed algorithms capable of producing scalable, adaptive, and fault tolerant behavior in artificial multi agent systems. These principles form the foundation for modern swarm robotics and flocking algorithms, including the Boids model used throughout this work.

(Section 2.2) Computational Swarm Models

(Subsection 2.2.a) Boids Flocking Algorithm

The Boids Flocking Algorithm (BFA), introduced by Craig Reynolds (Reynolds, 1987), is one of the most influential computational models for simulating collective swarm movement. Originally developed to replicate the realistic motion of birds and fish in computer graphics, the model demonstrated that complex group behavior can emerge from simple local rules. Rather than relying on centralized coordination or predefined trajectories, each agent determines its movement based on the behavior of nearby neighbors within a range.

The Boids model is founded on three core principles: separation, alignment, and cohesion. **Separation** prevents collisions between nearby agents, **alignment** encourages agents to match the direction of neighboring agents, and **cohesion** drives agents toward

the local center of the group. Through the combination of these simple rules, large groups of agents exhibit coordinated flocking behavior that resembles natural swarm systems observed in nature.

A key characteristic of the Boids algorithm is its reliance on local perception. Each agent possesses only partial knowledge of the environment and interacts exclusively with nearby agents rather than the entire swarm. Despite this limited awareness, coherent global motion emerges collectively through repeated local interactions.

The concept of emergent motion is central to the success of the Boids framework. Global flocking patterns, obstacle avoidance, and adaptive group movement arise naturally from decentralized interactions rather than explicit high level planning. As a result, the Boids model has become a foundational approach in swarm robotics, and distributed multi agent systems(Bansla & others, 2024; Ren & Sorensen, 2008).

(Subsection 2.2.b) Extension of Boids in Robotics

Like mentioned in the above section, researchers have explored numerous extensions of the Boids Flocking Algorithm (BFA) to address practical challenges encountered in robotic and autonomous systems. These adaptations include parameter optimization for task-dependent behavior, integration of obstacle sensing, avoidance mechanisms. Such approaches demonstrate the effectiveness of local, behavior based control models in enabling decentralized swarm coordination while maintaining flexibility in dynamic environments (Lu, Zhao, & Niu, 2025).

One of the most significant areas of extension involves obstacle avoidance and environmental awareness. Classical Boids models mostly focus on inter-agent interactions and do not explicitly account for complex environmental constraints. To address this limitation, researchers have incorporated sensing mechanisms and repulsive steering behaviors that allow agents to dynamically avoid static and moving obstacles while preserving swarm cohesion(Guan & Li, 2020; Liu & Chen, 2015; Zhao & Wang, 2023). These approaches improve the applicability of Boids based systems in real world scenarios involving uncertain terrain, collision risks, and constrained operational spaces.

Another important research direction concerns formation control and coordinated movement in robotic swarms. While the original Boids model generates emergent flocking behavior, many robotic applications require more structured and mission oriented formations. As a result, several studies combine Boids like local interactions with higher level formation maintenance strategies to improve navigation efficiency, communication, and task execution(Lu et al., 2025; Ouyang, Wu, Cong, & Wang, 2021; Ren & Sorensen, 2008). These adaptations are quite relevant in Unmanned Vehicle (UV) swarms.

Recent work has also investigated the integration of learning mechanisms into Boids based systems. Approaches involving adaptive parameter tuning and reinforcement learning attempt to improve swarm responsiveness under changing environmental conditions and safety critical scenarios(Huang et al., 2023; Yang et al., 2024). Although these methods demonstrate progress toward adaptive swarm intelligence, they often focus primarily on low level safety prioritization or isolated parameter optimization rather than learning coordinated control policies capable of preserving Boids locality while enabling explicit formation adaptation and mission aware behavior (Hengstebeck, Jamieson, & Van Scoy, 2024).

Overall, these extensions highlight both the strengths and limitations of classical Boids based systems. While local interaction rules provide robust and scalable multi agent coordination, additional mechanisms are often required to support structured formation control, adaptive decision making, and reliable operation in realistic robotic environments.

(Subsection 2.1.c) Formation Control in Multi-Robotic Systems

Formation control in multi robotic systems is a mature research field that focuses on the coordinated movement of multiple agents while preserving a desired structure. Researchers have developed various techniques to generate, maintain, and reshape formations using approaches such as master-worker architectures and graph theoretic controllers (Ren & Sorensen, 2008). These methods enable groups of robots to move collectively while maintaining stability, communication, and collision free operation.

In many robotic applications, formation control is essential for improving task efficiency and environmental awareness. Structured formations allow robotic teams to distribute sensing coverage, maintain communication links, and coordinate exploration or transportation tasks more effectively than isolated agents. For example, Unmanned Aerial Vehicle (UAV) swarms may use wide formations for surveillance and area coverage, while compact formations may be preferred in environments containing obstacles or external threats. Similarly, ground robotic systems often rely on formation maintenance during search and rescue operations, environmental monitoring, and cooperative mapping tasks.

Several formation control strategies have been proposed in the literature. In master-worker approaches, one or more agents define the reference trajectory while the remaining robots adjust their movement relative to the leaders. These methods are computationally simple and easy to implement, but they usually suffer from reduced robustness if the leading agents fail(Ouyang et al., 2021; Ren & Sorensen, 2008). Virtual structure methods instead treat the entire swarm as a single rigid body, where each agent maintains a predefined position within the global structure. This allows precise formation maintenance but often requires stronger synchronization and communication between agents(Ouyang et al., 2021). Furthermore, this approach, lacks in adaptability in continuously changing environments where other environmental entities may act as obstacles in their way, thus disrupting their placement in the swarm(Guan & Li, 2020; Zhao & Wang, 2023).

Graph theoretic approaches represent the swarm as a dynamic interaction network where agents exchange information with neighboring robots. Such methods are particularly important in decentralized robotic swarms because they improve scalability and fault tolerance while reducing communication overhead(Z. Li et al., 2024; Ren & Sorensen, 2008). However, maintaining stable formations under noisy sensing conditions, communication delays, and dynamic obstacles remains a challenging problem in realistic environments(Ouyang et al., 2021; Zeng, Zhang, & Lim, 2020).

Modern formation control research increasingly combines classical control theory with adaptive and intelligent decision making mechanisms. Reinforcement learning, optimization methods, and hybrid swarm intelligence models have been explored to allow formations to dynamically adapt according to environmental conditions and mission objectives(Huang et al., 2023; Lu et al., 2025; Yang et al., 2024). Rather than maintaining a fixed geometric structure, adaptive formation systems attempt to balance cohesion, safety, coverage, and navigation efficiency. This transition from static formation maintenance toward adaptive swarm coordination is closely related to the objectives of the present work, where agents dynamically switch between formation profiles according to environmental feedback while preserving decentralized Boids based coordination(Hengstebeck et al., 2024; Lu et al., 2025).

(Section 2.3) Reinforcement Learning in Swarm Systems

Recent work has increasingly examined swarm systems operating within mobile and EC environments, where distributed intelligence is leveraged to support real time coordination, computation offloading, and adaptive task execution (Huh & Seo, 2019; Santos & Almeida, 2021). In such settings, EC infrastructures provide nearby computational resources that enable latency sensitive decision making and reduce reliance on centralized cloud services. This is particularly critical for highly mobile and bandwidth limited swarm nodes. Several studies investigate UAV and robotic swarms assisted by mobile EC, focusing on the joint optimization of communication latency or energy consumption through learning based or optimization driven scheduling mechanisms(B. Li, Fei, & Zhang, 2021; Zeng et al., 2020). These approaches demonstrate how edge assisted intelligence can enhance the scalability and responsiveness of swarms in dynamic environments. Beyond performance optimization, EC swarm systems have also been explored in mission critical and safety sensitive scenarios, including disaster response, emergency communications, and autonomous exploration in potentially hostile environments (Bansla & others, 2024; Sakano et al., 2025). In such applications, the combination of swarm coordination and EC enables resilient operation despite fragile connectivity and dynamic task requirements. However, much of this literature treats swarm coordination at an abstract or task allocation level, often assuming simplified motion models. This observation motivates the need for approaches that more tightly integrate low level swarm behaviors with formation aware control.

(Subsection 2.3.a) Model Based Reinforcement Learning

Model based reinforcement learning (MBRL) refers to reinforcement learning approaches in which agents learn or exploit an internal representation of the environment dynamics in order to improve decision making. Instead of relying exclusively on direct interaction with the environment, model based methods attempt to predict future states, rewards, or transitions, allowing agents to plan actions before executing them. This predictive capability often improves sample efficiency and enables faster adaptation in new environments compared to purely trial and error approaches.

In swarm robotic systems, model based reinforcement learning is particularly attractive because multi agent environments are often dynamic, partially observable, and resource constrained (Z. Li et al., 2024; Lowe et al., 2020). Agents operating in robotic swarms must continuously estimate the consequences of their movement decisions while maintaining coordination with neighboring agents. By incorporating predictive models of environmental dynamics, model based methods can support more stable collective behavior. In many cases, reinforcement learning is applied to optimize navigation, task allocation, or energy management in distributed environments assisted by EC infrastructures(Huh & Seo, 2019; Sakano et al., 2025). However, many existing methods replace classical swarm coordination mechanisms entirely with learned policies. As a result, the natural locality and decentralized characteristics of swarm intelligence are sometimes reduced in favor of centralized or heavily optimized control strategies. It should be noted, that such alternative introduces much greater computational costs that simply relying on a rule based system.

Another important limitation concerns formation awareness and spatial coordination. While reinforcement learning has shown promising results in autonomous navigation and multi agent cooperation, relatively few works fully integrate Boids style local flocking with structured formation control and mission aware behavior (Lu et al., 2025). Existing approaches often focus on isolated objectives such as obstacle avoidance, path optimization, or parameter tuning rather than adaptive coordination between different formation categories.

In particular, limited research has examined reinforcement learning systems capable of simultaneously supporting three dimensional exploration, decentralized flocking behavior, and dynamic formation selection under stochastic environmental conditions. This gap motivates the proposed approach of combining Boids based local interaction rules with a shared memory and reinforcement learning framework.

(Subsection 2.3.b) Model-Free Reinforcement Learning

Model free reinforcement learning (MFRL) refers to reinforcement learning methods in which agents learn optimal behaviors directly through interaction with the environment without relying on an explicit model. Instead of predicting future state transitions, agents improve their policies through repeated observation of rewards and penalties generated by their actions. Over time, the learning process allows agents to identify action strategies that maximize long term cumulative reward.

Model free approaches are widely used in robotics and autonomous systems (Huang et al., 2023; Yang et al., 2024) because they are capable of operating in highly complex and uncertain environments where accurate mathematical models may be difficult or impossible to construct. In swarm systems, agents often interact with dynamic obstacles, partially observable environments, and continuously changing neighboring behaviors. Under such conditions, model free learning provides flexibility by allowing agents to adapt directly from experience rather than depending on predefined environmental assumptions.

Several model free reinforcement learning algorithms have been applied in swarm intelligence and multi agent coordination problems. Value based methods such as Q-Learning and Deep Q Networks (DQN) learn policies by estimating the expected reward of actions within specific states(Z. Li et al., 2024; Lowe et al., 2020). Policy based methods instead directly optimize the action selection strategy, making them more suitable for continuous control problems frequently encountered in robotic navigation.

Despite these advantages, model free reinforcement learning also introduces several challenges in swarm systems. One major difficulty concerns the large state and action spaces generated by multi agent interactions. As the number of agents increases, the complexity of learning coordinated behaviors grows significantly, often resulting in unstable training or slow convergence. In addition, purely learned policies may produce unpredictable or unsafe behavior during exploration, particularly in safety critical environments involving physical robotic systems.

Another limitation is that many model free approaches focus primarily on optimizing reward signals while overlooking the structured coordination mechanisms commonly found in swarm intelligence models. Furthermore, relatively limited research investigates model free reinforcement learning in scenarios involving adaptive formation switching, formation aware exploration, and coordinated navigation in fully three dimensional environments.

For these reasons, hybrid approaches that combine classical swarm intelligence principles with reinforcement learning have gained increasing attention(Hengstebeck et al., 2024; Lu et al., 2025). By integrating Boids based local coordination with model free learning mechanisms, it becomes possible to preserve emergent swarm behavior while enabling adaptive high level decision making. In the proposed framework, reinforcement learning does not replace decentralized flocking rules but instead assists agents in selecting appropriate formation profiles according to environmental conditions, exploration objectives, and perceived threats. This combination aims to balance adaptability, robustness, and scalable swarm coordination in dynamic three dimensional environments.

CHAPTER 3
SYSTEM MODEL AND ENVIRONMENT

(Section 3.1 Environment Model)

(Subsection 3.1.a) 3D Space Definition

In the context of the proposed system, the environment is modeled as a bounded three dimensional space ($D = 3$). Each position inside the operational domain is represented through Cartesian coordinates x, y, z, allowing agents to navigate freely across horizontal and vertical axes. This representation, as seen in Figure 1 is particularly important for aerial and autonomous swarm systems operating in realistic environments where altitude variation, obstacle elevation, and volumetric exploration must be considered simultaneously. The bounded environment is defined by finite spatial limits along all three axes:

$$\Omega = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}] \times [z_{\min}, z_{\max}] \quad , \Omega \subset \mathbb{R}^D \tag{1}$$

where the limits define the accessible operational region of the swarm. Agents attempting to move beyond these boundaries are constrained within the valid exploration space. Such spatial constraints simulate realistic operational conditions where autonomous systems must navigate inside predefined mission areas or physically restricted environments.

The operational space in which the swarm based system operates is characterized as a bounded, multi-dimensional domain where autonomous nodes and environmental obstacles coexist and interact. The environment is mathematically represented as a set of vectors $(\mathcal{N}, \mathcal{O}) \in \mathbb{R}^D X \mathbb{R}^D$, with D representing the number of dimensions.

- $\mathcal{N}(t) \subset \Omega$ is dynamic subset of the environment representing the swarm nodes.
- $\mathcal{O} = \{\mathcal{O}_k\}_{k=1}^{N_O}, \mathcal{O}_k \subset \Omega$ is a static subset of the environment representing environmental obstacles.

The use of a three dimensional representation significantly increases the complexity of coordination and navigation compared to two dimensional swarm models. Agents must simultaneously maintain formation cohesion, avoid collisions, preserve safe inter-agent distances, and adapt movement trajectories while considering vertical displacement and obstacle distribution across all spatial dimensions.

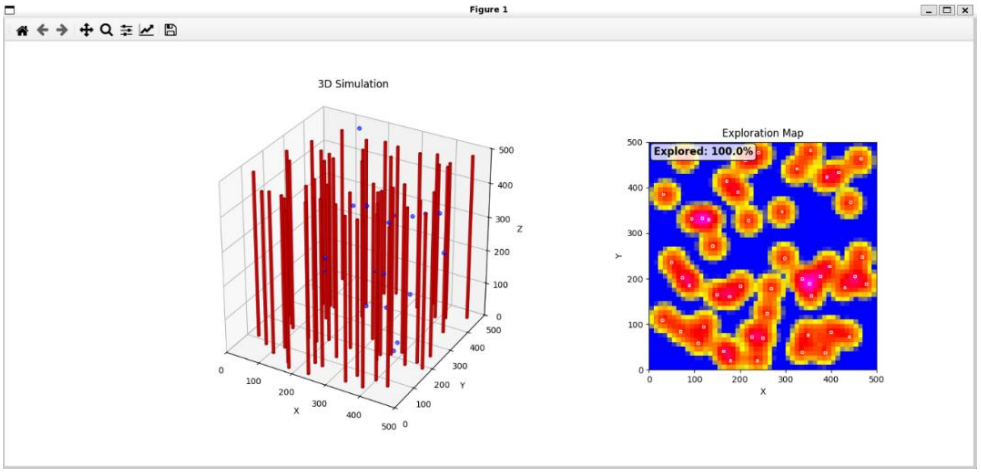

Figure 1: Initial Environment Simulation using Matplotlib Library in Python

(Subsection 3.1.b) Preliminaries

In this research, the scouting optimization problem is approached with a framework that combines biologically inspired swarm behavior with reinforcement learning based decision making. As mentioned earlier, it operates on three simple but powerful principles often referred, in computer science, by Boids Rules:

(a) Separation, ensuring nodes avoid collisions with neighbors.

(b) Alignment, encouraging nodes to align their velocities with nearby peers.

(c) Cohesion, driving nodes toward the center of their local neighborhood.

These rules allow nodes to exhibit emergent swarm behaviors, such as collective movement, obstacle avoidance, and spatial cohesion, without requiring any centralized control. This makes Boids particularly well suited for our context of EC and mobile computing in general, where nodes must act independently yet maintain group integrity in stochastic environments. Table 1 presents the notation adopted in the Thesis.

Notation	Description	Note
$\boldsymbol{\Omega}$	Operational Domain	$\boldsymbol{\Omega} \subset \mathbb{R}^{\mathbf{D}}$
$\mathbf{D}$	Spatial dimension- **D**=3 in this study	Positive integer
$\mathbb{R}^{\mathbf{D}}$	D-dimensional Euclidean space	Vector space
$\mathcal{N}(\mathrm{t})$	Active swarm agents at time t	set
N_A	Number of swarm agents	integer
$\mathcal{O} = \{\mathcal{O}_{\mathrm{k}}\}$	Obstacle regions	Set of regions
$\mathbf{o}_{\mathrm{k}}$	Center of reference point of obstacle k	Vector in $\mathbb{R}^3$
$\mathbf{p}_i(\mathrm{t})$	Position of agent i	Vector in $\mathbb{R}^3$
$\mathbf{p}_i^{cand}(t)$	Candidate next position	Vector in $\mathbb{R}^3$
$\mathbf{v_i}(t)$	Velocity of agent i	Vector in $\mathbb{R}^3$
Δt	Simulation time step	Positive scalar
$\mathbf{d}_{ij}(t)$	Euclidean Distance between agents $i-j$	Nonnegative scalar
$\boldsymbol{R}_{sense}$	Sensing Radius	Positive Scalar
$\mathcal{N}_i^{sense}(t)$	Agents sensed by agent i	Set
R_{sep}	Separation radius	Positive Scalar
R_{align}	Alignment radius	Positive Scalar
R_{coh}	Cohesion radius	Positive Scalar

$\boldsymbol{s}_i(t)$	Separation steering vector	Vector
$\boldsymbol{a}_i(t)$	Alignment steering vector	Vector
$\boldsymbol{c}_i(t)$	Cohesion steering vector	Vector
$\mathbf{p}_{com,i}(t)$	Local neighborhood center of mass for agent i	Vector
$\boldsymbol{u}_i(t)$	Combined steering command	Vector
$\boldsymbol{w}_s, \boldsymbol{w}_a, \boldsymbol{w}_c$	Boids rule weights	Nonnegative scalars
$\mathbf{F}(t)$	Active formation class	Categorical Variable
$\mathbf{J}_i(t)$	Per agent fitness or objective	Scalar
$\mathbf{A}_i(t)$	Adherence-to-dynamics score	Scalar
$\mathbf{d}_i(t)$	Distance to assigned target	Nonnegative scalar
$\mathbf{P}_i(t)$	Penalty term	Nonnegative scalar
$\mathbf{d}_{safe}$	Minimum obstacle safety distance	Positive scalar
$\boldsymbol{\tau}(\mathbf{p})$	Threat intensity at position p	Nonnegative scalar field
$\boldsymbol{\rho}(\mathbf{p})$	Opportunity value at position p	Scalar field
$\mathbf{q}(\mathbf{p})$	Environmental desirability	Scalar field
$\boldsymbol{\xi}_{\mathrm{k}}$	Threat coefficient of obstacle k	Nonnegative scalar
$\boldsymbol{\varepsilon}$	Regularization constant	Small positive scalar
$\mathcal{K}_i(t)$	Obstacles sensed by agent i	Index set
$\mathbf{C}(n)$	Traversal cost of grid cell n	Nonnegative scalar
$\mathbf{D}(n)$	Distance component of traversal cost	Normalized Scalar
$\mathbf{T}(n)$	Threat cost component	Normalized scalar
$\mathbf{M}_{F(n)}$	Formation dependent motion multiplier	Positive scalar
$\mathbf{g}(n)$	Accumulated path cost to n	Nonnegative scalar
$\mathbf{h}(n)$	A* heuristic	Nonnegative scalar
$\mathbf{E}(n)$	A* evaluation score	Nonnegative scalar
$\boldsymbol{\eta}$	Mean threat exposure along the path	Nonnegative scalar
$\boldsymbol{\mathcal{M}}$	Markov decision process	Tuple
$\mathcal{S}$	RL state space	Set
$\mathcal{A}$	RL action space	Set
$\mathbf{s}_i(t)$	State observed by agent i	State
$\mathbf{a}_i(t)$	Behavioral multiplier action	Action
$\mathcal{U}$	Available multiplier values	Finite set
$\mathbf{P}(\mathbf{s}'\|\mathbf{s},\mathbf{a})$	State transition Kernel	Probability
$\mathbf{r}_i(t)$	Immediate reward	Scalar
γ	Discount factor	[0,1]
a	Q-Learning rate	(0,1]
π	$\pi: \mathcal{S} \rightarrow \mathcal{A}$	Mapping

Table 1: Nomenclature

(Section 3.2 Swarm Agent Model)

(Subsection 3.2.α) Agent State and Parameters

Each swarm node is represented as a dynamic point within the three dimensional operational space $\mathbb{R}^3$, is described by the following equation:

$$\mathbf{p}_i(t)=\big(x_i(t), y_i(t), z_i(t)\big) \in \mathbb{R}^3 \tag{2}$$

The time dependence t emphasizes that positions change dynamically in response to environmental factors in real time. In any real world deployment, nodes must navigate around obstacles while maintaining formation. This is particularly important in

unpredictable stochastic environments, where new obstacles may appear in real time. To prevent collisions and inefficiencies, each node follows an avoidance rule:

$$\mathbf{p}_i^{\text{cand}}(t) \notin \mathcal{O}_k \; \forall \, k \in \{1, \ldots, N_O\} \tag{3}$$

This constraint ensures that no node ends up colliding with an obstacle, preserving both the safety of individual nodes and the overall swarm structure. When a node approaches a potential obstacle or enters an environment that could potentially be hazardous, maintaining a safety buffer. This Safety Distance acts as a protective perimeter, ensuring that nodes do not venture too close to areas of high risk. The constraint for safety aware movement is:

$$\| \, \mathbf{p}'_i(t) - o_k \, \|_2 > d_{safe} \forall \, k \tag{4}$$

This ensures that nodes do not approach threats beyond an acceptable risk threshold, minimizing unnecessary losses, and avoid losing time and resources to search an interval that does not seem to lead to the optimal solution. A key component in enhancing obstacle avoidance in a swarm system is the integration of a memory like system that keeps track of the locations and characteristics of previously encountered obstacles. When a node detects an obstacle, it stores the encounter not only as spatial coordinates but also with additional contextual information. Instead of relying on external infrastructure or continuous connectivity, the swarm maintains this memory in a fully distributed manner. Nodes periodically exchange lightweight summaries of their local obstacle maps with nearby agents. Through this decentralized sharing process, the swarm incrementally builds and maintains a coherent collective memory. If the swarm decides to revisit a previously traversed area, the nodes can refer to this memory to avoid previously encountered obstacles. In case the swarm decides to pass through or revisit an area with known obstacles, the safety distance(explained later on) prevents the nodes from coming too close to the regions where their trajectory may lead to collisions or other dangers.

Each agents motion is represented by a three dimensional velocity vector containing its Cartesian velocity components along the 3 axis. The following formula is used to update the agents position throughout the whole navigation process:

$$\mathbf{v}_i(t) = (v_{x,i}(t), v_{y,i}(t), \, v_{z,i}(t)) \; \in \; \mathbb{R}^3 \tag{5}$$

where:

- $\mathbf{v}_i(t)$ is the velocity vector of an agent
- $v_{x,i}(t), v_{y,i}(t), \, v_{z,i}(t)$ are the cartesian velocity components

In summary, the concept of obstacle avoidance within swarm systems involves more than just detecting and avoiding obstacles that may appear at a specific time instance. It requires a sophisticated, memory based strategy that enables nodes to learn from past encounters and adjust their paths dynamically.

(Subsection 3.2.b) Sensing

Nodes are naturally inclined to avoid areas with high threat levels, by leveraging their sensing range. The sensing range is defined as the *personal domain* of each node and describes all entities in the environment that are within a predefined range around its current position. This range is often described in real world adaptations as the *field of*

view or the *visibility radius* and can have different geometrical shapes like circular or spherical in case that $D = 3$, a conical or even elliptical shape as shown in Figure 2.

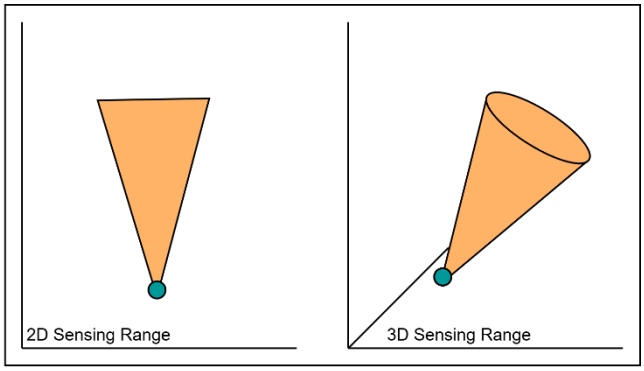

Figure 2:2D-3D Representation of an agents sensing range

The size of the range dictates the nodes' ability to detect environmental features, such as obstacles or other nearby nodes. The core formula to check if a node A_j with position p_j is in another node's A_i radius at time t is:

$$\mathcal{N}_i^{sense}(t) = \{j \neq i : d_{i,j}(t) \leq R_{sense}\}, d_{i,j} = \| \boldsymbol{p}'_i(t) - o_k \|_2 \quad (6)$$

The same function is used to detect non moving entities, i.e., obstacles in the environment, with the main distinction being that the entity with position $p_j(t)$ keeps it static at any time t, so this parameter is non mandatory.

(Subsection 3.2.c) Boids-Based Motion Model

The emergent result of the Boids local rules is depicted by a coherent global behavior, allowing the swarm to maintain formation, avoid collisions, and adapt to environmental changes without centralized control. As previously mentioned above the proposed model is governed by three fundamental rules:

- **Separation:** Each node steers to avoid collisions with nearby nodes (thus, it does not search the same areas for finding the solution to the optimization problem), ensuring local safety and preventing overcrowding within a protected range. A common formulation is to average the repulsive vectors from close vectors:

$$s_i(\mathrm{t}) = \sum_{\mathrm{n} \in \mathcal{N}_{i,sep}} f_{repel}(\mathbf{p}_i, \mathbf{p}_j) \quad (7)$$

Where $\mathcal{N}_{i,sep}$ is the set of neighbors in the separation radius while the repulsing force between two agents is defined as:

$$f_{repel}=\frac{\mathbf{p}_i-\mathbf{p}_j}{|\mathbf{p}_i-\mathbf{p}_j|2} \tag{8}$$

- **Alignment:** Each node adjusts its velocity to match the average heading of its neighbors, producing coordinated group motion. First, the average velocity of the neighbors $\mathbf{v}_{avg}=\frac{1}{n_i}\sum_{A_j\in N_i}\mathbf{v}_i$ is calculated (N_i is the set of neighbors for the ith node and $\mathcal{N}_j^{\mathbf{v}}$ is the velocity of the j-th node). Next, we calculate the steering vector required to reach this average velocity: $\mathbf{v}_{avg}-\mathbf{v}_i$. Then, the final alignment formula is formed:

$$a_i(\mathrm{t})=\frac{1}{n_i}\sum_{\mathcal{N}_j\in N_i}\mathbf{v}_j-\mathbf{v}_i \tag{9}$$

- **Cohesion:** Each node moves towards the perceived center of mass of the swarm. Another simple definition would be the steering force towards the average position of nearby nodes, maintaining group integrity and preventing fragmentation. A general formula of the cohesion rule is:

$$\mathbf{c}_i(\mathrm{t})=\mathbf{p}_{com,i}(t)-\mathbf{p}_i(t) \tag{10}$$

where p_{com} is the center of mass of the nodes in SR:

$$\mathbf{p}_{com}(\mathrm{t})=\frac{1}{\mathcal{N}_i}\sum_{A_j\in\mathcal{N}_i}\mathbf{p}_j(t)$$

Formally, the total steering force of node represented by $\mathcal{N}_i$ at time t can be expressed as:

$$\boldsymbol{u}_i=w_s s_i(t)+w_a a_i(t)+w_c c_i(t) \tag{11}$$

where w_s, w_a, w_c are smoothing weights controlling the relative influence of each rule. By tuning these weights, the swarm can exhibit different movement characteristics, from tightly cohesive formations to more dispersed exploratory behaviors. This flexibility makes our model well suited for stochastic environments, where nodes must continuously adapt to avoid threats and exploit opportunities while maintaining overall swarm integrity. The fitness function used to evaluate potential moves is as follows:

$$J_i(t)=w_1A_i(t)+w_2\rho(\mathbf{p}_i(t))-w_3d_i(t)-P_i(t) \tag{12}$$

where:

- $A_i(t)$: Adherence to swarm dynamics (alignment, cohesion, and separation), derived from the node's velocity behavior.
- $\rho(\mathbf{p_i}(t))$: Opportunity value at the node's position, inversely related to local threat intensity.
- $d_i(t)$: Euclidean distance between node i and its assigned task or target.
- w_1, w_2, w_3: Weighting coefficients controlling the relative importance of motion stability, opportunity maximization, and task proximity.

As for the penalty function:

$$P_i(t) = p_f I_{f,i}(t) + p_t I_{t,i}(t) + p_l I_{l,i}(t) \tag{13}$$

introduces negative rewards when an agent violates important constrains. It consists of binary occurrence indicators $I_{f,i}, I_{t,i}, I_{l,i} \in \{0,1\}$ and penalty coefficients $p_f, p_t, p_\tau > 0$. For example:

- $I_{f,i} = 1$ if the agent violates the currently active formation by exceeding the maximum allowable deviation from its neighboring agents otherwise $I_f = 0$.
- $I_{t,i} = 1$ if the agent enters a high threat region without adopting the Defensive profile otherwise $I_t = 0$.
- $I_{l,i} = 1$ if the agent exceeds the maximum allowed episode duration (or simulation step limit) otherwise $I_l = 0$.

In Boids, the initial velocity of an node is not strictly defined by a formula, instead it is typically defined by an initialization strategy chosen by the implementer (Reynolds, 1987). Regarding the update of the velocity for each node, we consider that it is calculated by adding the steering force f_i to the existing velocity v_i. Formulated, the new steered and unconstrained velocity is:

$$\mathbf{v}_i' = \mathbf{v}' + f_i \tag{14}$$

The speed constraint is a critical, supplementary mechanism and without it, the accumulation of acceleration from the Separation, Alignment, and Cohesion rules over time could cause nodes to achieve unrealistically high speeds, effectively *flying off* the simulation area, or conversely, cause them to stall entirely. The constraint defines a tunable operating range, i.e., a minimum and maximum speed that the magnitude of nodes' velocity vector is forced to remain within at the end of every simulation step. The following equation holds true:

$$\mathbf{v}_i(t + \Delta t) = \begin{cases} \mathbf{v}_i' & if\ v_{min} \leq \| \mathbf{v}'_i \|_2 \leq v_{max}, \\ \dfrac{\mathbf{v}_i'}{\| \mathbf{v}_i' \|} \cdot v_{max}, & if\ \| \mathbf{v}'_i \|_2 > v_{max}, \\ \dfrac{\mathbf{v}_i'}{\| \mathbf{v}_i' \|} \cdot v_{min}, & if\ 0 < \| \mathbf{v}'_i \|_2 < v_{min} \end{cases} \tag{15}$$

Where $\| \mathbf{v}_i(t) \|_2$ is the magnitude(speed) of the velocity vector:

$$\| \mathbf{v}_i(t) \|_2 = \sqrt{{v^2}_{x,i} + {v^2}_{y,i} + {v^2}_{z,i}} \tag{16}$$

(Section 3.3) Environmental Constraints and Obstacles

The spatial characteristics of the operational environment comprise a three dimensional domain $\mathbb{R}^3$ where computations are conducted, but for visualization purposes, it is often represented in two-dimensional form using vector-like representations. Obstacles within this space are modeled as elongated column shaped entities rather than discrete points, spanning continuous intervals along the vertical axis and projected onto two dimensional heatmaps as occupied grid regions. This visualization simplifies the monitoring of swarm movement, explored areas, and the intensity of surrounding threat regions throughout the environment.

Theoretically, an obstacle can be understood as any object or environmental feature that blocks the free movement of a node or poses a risk to the integrity of the swarm. In swarm intelligence systems, obstacles are not perceived solely as physical barriers but as environmental constraints that influence node decision making, trajectory selection, and collective movement behavior. Obstacles may vary in nature, including static structures, dynamic moving objects, or even neighboring hostile swarm agents. In the present study, obstacles are assumed to be static thus, the time index t is not mandatory, since their position does not variate over time. Therefore, their spatial position remains constant over time and the explicit temporal index is omitted from the obstacle representation. Collision conditions occur whenever an agent attempts to enter or intersect an occupied obstacle region or violates predefined safety distances relative to surrounding agents and environmental structures. Consequently, valid movement trajectories must preserve collision free navigation while maintaining swarm cohesion and formation stability.

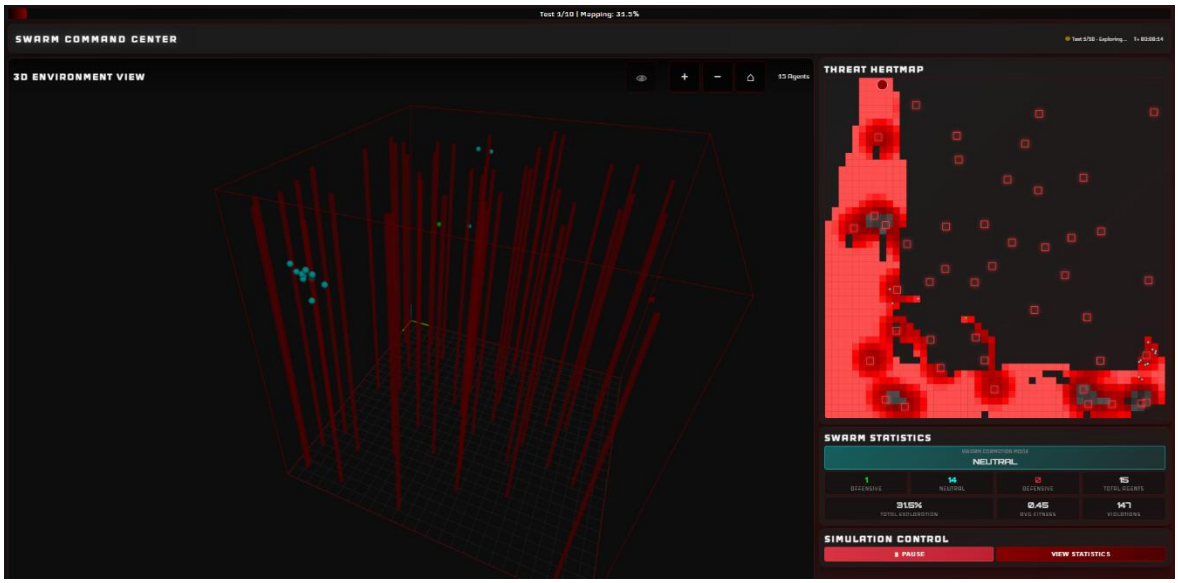

Figure 3:Finalized 3D web based interactive experimental environment

Collision conditions occur whenever an agent attempts to enter or intersect an occupied obstacle region or violates predefined safety distances relative to surrounding agents and environmental structures. Consequently, valid movement trajectories must preserve collision free navigation while maintaining swarm cohesion and formation stability. These constraints are particularly important in dense environments where navigation space is limited and the probability of inter agent or obstacle collisions increases significantly. Beyond direct physical obstacles, the environment may also contain threat regions associated with elevated navigation risk. Such regions do not necessarily block movement entirely but instead increase the potential cost or danger

associated with traversal. The mathematical modeling and detection of such threats are discussed in the following subsection.

(Subsection 3.3.a) Threat Field

By explicitly defining environmental obstacles and their surrounding influence regions, swarm nodes can plan movement trajectories that reduce collision probability and improve navigation safety. Although the operational environment is represented in three dimensions obstacle detection and threat visualization may also be projected onto two dimensional spatial maps for monitoring and computational simplification as shown below in Figure 4.

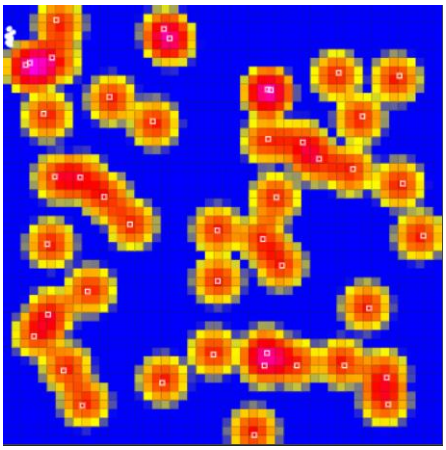

Figure 4:2D Heatmap Vectorization of the 3D simulation environment. Threat field is depicted as radiating heat around obstacles

In this representation, each obstacle occupies a spatial region that influences nearby agents through a localized threat field:

$$o_k=(x_k, y_k, z_k) \in \mathbb{R}^3, \qquad \mathcal{O}_k \subset \mathbb{R}^3, \qquad k = 1, \ldots, N_O \tag{17}$$

The threat field models the level of environmental risk associated with a specific position in space. Rather than treating obstacles solely as binary inaccessible regions, the proposed framework assigns continuous threat intensities around obstacles, allowing agents to adapt their behavior as they approach dangerous areas. Consequently, nodes experience a gradient like environmental pressure when navigating near obstacles or overlapping hazardous regions.

$$\tau(\mathbf{p}_i)= \sum_{k=1 \in \mathcal{K}_i(t)} \int_{\mathbf{q} \in \mathcal{O}_k} \frac{\xi_k}{\| \boldsymbol{p}_i - \boldsymbol{q} \|_2^2 + \varepsilon^2} d\mathbf{q}, \;\; \mathcal{K}_i(t) = \{k: dist(\mathbf{p}_i, \mathcal{O}_k) \leq R_{sense}\} \tag{18}$$

where:

- $\mathcal{O}_k$: The total number of obstacles in the sensing range of the node.
- ξ_k : The intrinsic threat intensity or influence coefficient of the k-th obstacle.
- $\mathbf{p}_i'$: A potential new position of the node within the obstacle region.
- ε : A small positive constant added to prevent singularities when $\mathbf{p_i} \approx \mathbf{p}_i'$.

The closer the node is to a threat, the higher the risk it experiences, thus, the stronger the negative contribution of the threat field. This formulation models threat intensity as an inverse distance dependent influence function. The closer a node moves toward a dangerous region, the larger the threat contribution becomes. As a result, nearby obstacles exert stronger repulsive influence on agent navigation, encouraging safer trajectory selection and collision avoidance.

An important characteristic of the proposed threat representation is the superposition of overlapping threat regions. When multiple obstacles are located near one another, their corresponding threat fields accumulate, naturally generating high risk navigation zones. Such regions become increasingly undesirable for traversal because the total environmental risk grows proportionally with obstacle density and proximity. This behavior is particularly important in cluttered or partially constrained environments where narrow passages and dense obstacle distributions significantly affect swarm movement and formation adaptation.

(Subsection 3.3.b) Opportunity Field

The opportunity field represents favorable regions within the operational environment that provide safe, efficient, or strategically advantageous navigation conditions. Unlike threat regions, which discourage traversal, opportunity regions positively influence swarm movement by attracting nodes toward areas associated with reduced environmental risk and improved navigational efficiency.

In the proposed framework, opportunity is inversely related to environmental threat. Regions containing low obstacle density, safe traversal conditions, or reduced collision probability naturally produce higher opportunity values. Consequently, agents are encouraged to move toward spatial regions that maximize safety while maintaining efficient exploration and coordinated swarm movement. Formally, the opportunity field is defined as:

$$\rho(\mathbf{p}) = \frac{1}{1+\tau(\mathbf{p})} \tag{19}$$

where:

- $\rho(\mathbf{p})$ The opportunity value associated with position $\mathbf{p}$
- $\tau(\mathbf{p})$ The threat intensity evaluated at the same spatial position

This inverse relationship ensures that regions with high threat intensity produce low opportunity values, while regions with minimal threat generate higher opportunity scores. Therefore, the opportunity field acts as a positive navigation incentive that guides agents toward safer and more favorable areas of the environment.

The opportunity field additionally supports cooperative exploration by indirectly reinforcing efficient traversal paths. In this research, areas associated with high opportunity may accumulate stronger pheromone deposition or increased traversal preference, encouraging neighboring agents to follow safer and more efficient routes. Over time, this collective behavior contributes to the emergence of adaptive exploration patterns and coordinated navigation pathways across the swarm.

(Subsection 3.3.c) Threat-Opportunity Function

To support balanced environmental evaluation, the proposed framework combines both threat and opportunity contributions into a unified environmental assessment function. This integrated representation allows swarm agents to simultaneously consider

environmental danger and navigational attractiveness when selecting movement trajectories and adapting formation behavior. The Threat-Opportunity Field function integrates these contributions into a single value:

$$f(\mathbf{p}_i)=\rho(\mathbf{p}_i)-\beta\tau(\mathbf{p}_i), \beta \geq 0 \quad (20)$$

The resulting function provides a continuous representation of environmental quality across the operational domain. Regions associated with low threat and high opportunity produce favorable navigation conditions, while areas containing concentrated obstacle influence generate increasingly unfavorable evaluations. Through this combined environmental assessment, swarm agents are capable of adapting their movement according to both safety constraints and exploration objectives.

This unified formulation is particularly important for the later reinforcement learning and path planning stages of the proposed framework. By continuously evaluating the surrounding environment through both positive and negative spatial influences, agents can make adaptive navigation decisions.

CHAPTER 4 SWARM FORMATION AND NAVIGATION STRATEGY

(Section 4.1 Formation Classification)

This section formalizes the problem of dynamic formation adoption by EC nodes. Main objective is to develop methodologies that enable swarms to achieve optimal formation configurations with high accuracy, even under uncertainty and dynamic conditions. This sets the stage for proposing solutions that balance individual adaptability with global structural integrity. The core challenge involves determining the most effective formation for the swarm within a dynamic environment $\mathbb{R}^3$, where nodes must adaptively reposition to enhance both individual and collective performance, responding to evolving environmental stimuli such as threats. Nodes dynamically adjust their behavior based on environmental context.

(Subsection 4.1.a) Formations

Formations are represented as parameter profiles, which modify node motion rules separation, alignment, and cohesion according to the situational feedback. The swarm evaluates its environment to classify threats and opportunities, then selects one of three formation modes by adapting hyperparameters. The following list presents the proposed formation schemes:

- **Offensive Formation (OF)**: It is activated in low threat and high opportunity scenarios. Nodes reduce cohesion to explore and cover more intervals in the search space while maintaining alignment for coordinated movement. Separation is moderate, allowing free movement without collisions. This configuration maximizes exploration speeds significantly and target engagement while keeping the core node minimally protected.

- **Neutral Formation (NF)**: It is used in moderate threat environments. Nodes maintain balanced cohesion and alignment, allowing the swarm to stay together but remain flexible to split if needed. Separation is balanced for safety and *maneuverability*. This mode supports general purpose tasks such as scouting, monitoring, or communication, providing adaptability without committing fully to offense or defense.

- **Defensive Formation (DF)**: Triggered in high threat environments. Cohesion is increased and separation minimized, forcing nodes to cluster tightly and protect the most critical nodes, such as the core data-holder node (a node dedicated to keep important information for the swarm - the study of the corresponding behavior lies beyond the scope of this thesis). Alignment ensures unified movement, allowing the swarm to collectively avoid hazards or respond to threats. This mode prioritizes survivability and preserves mission critical functionality.

By defining formations as adaptive parameter sets, the swarm retains the emergent, local rule based dynamics while exhibiting strategic behavior. The system continuously monitors the environment, selecting and updating formation parameters to maximize the exploration over time ratio.

(Subsection 4.1.b) Formation Adaptation via Environmental Feedback)

A key step is to enforce adherence to the current formation mode. Swarm based systems rely on well coordinated movement to maintain cohesion, avoid obstacles, and adapt to stochastic environments. To achieve this, nodes follow rules that guide their positioning, ensuring the swarm remains structured while effectively responding to unknown parameters. These rules include **formation integrity**, **obstacle avoidance**, **communication range maintenance**, and **deviation control**. Each rule plays a key role in keeping the swarm in a favorable configuration. In this approach, formation adherence is achieved by dynamically adjusting parameters according to environmental feedback.

At each time instant, the swarm's behavior follows the selected formation profile (Offensive, Neutral, or Defensive), maintaining cohesion and alignment according to the hyperparameter settings. A well defined formation profile allows nodes to work together efficiently while retaining flexibility. Obstacle avoidance remains fundamental: the separation weight in the algorithm is tuned dynamically to prevent collisions while maintaining cohesion. Communication connectivity is preserved as nodes maintain proximity to neighbors, ensuring decentralized information flow. Deviation control is enforced via parameter adjustments, preventing nodes from drifting too far from the swarm and maintaining effective coverage. By modulating algorithm's parameters, the swarm preserves emergent local rule behavior while achieving context sensitive formation control, maximizing survival, exploration, and operational efficiency.

(Section 4.2 Group Movement and Exploration)

Following the exploration phase of the previously unknown environment, the subsequent objective is to enable efficient, safe, and coordinated navigation within the newly acquired map. During exploration, swarm agents progressively construct a partial environmental representation by identifying traversable regions, obstacle locations, and surrounding threat distributions. Once sufficient environmental knowledge has been accumulated, the system transitions from exploration oriented behavior toward exploitation oriented navigation, where the swarm attempts to move efficiently between selected locations while preserving formation integrity and minimizing environmental risk.

Unlike classical pathfinding problems that assume complete environmental knowledge from the beginning, the proposed framework operates under the constraint that unexplored regions cannot initially be considered safe or traversable. Consequently, path planning is performed primarily over previously explored areas where environmental information has already been collected by the swarm. This restriction significantly increases the complexity of navigation, as the swarm must balance efficient movement with uncertainty avoidance and formation preservation.

Group movement in the proposed framework is therefore not treated as a simple shortest path problem. Instead, navigation decisions are influenced simultaneously by geometric distance, local threat intensity, obstacle density, and the currently active swarm formation profile. Different formation configurations naturally produce different mobility characteristics. Offensive formations prioritize rapid exploration and wider spatial coverage, while defensive formations prioritize safety and compact coordinated movement. As a result, the navigation strategy must continuously adapt according to both environmental conditions and the swarm's collective behavioral state.

To address these challenges, the proposed system introduces a balanced navigation algorithm that combines path efficiency and environmental awareness within a unified

cost formulation. The algorithm operates over a grid based abstraction of the explored environment, allowing the swarm to evaluate candidate movement trajectories according to both traversal distance and accumulated environmental risk.

(Subsection 4.2.a) Balanced Navigation Algorithm

The proposed balanced navigation algorithm is an A* based path planning method designed to explicitly balance geometric efficiency and environmental safety. Unlike traditional shortest path planners such as Dijkstra's algorithm, which prioritize only distance minimization, or purely defensive approaches that focus exclusively on safety, the proposed method treats both objectives as equally important. This balanced formulation allows the navigation strategy to operate as an intermediate behavior between aggressive exploration and highly conservative traversal.

The operational environment is discretized into a grid based search space, where each grid cell represents a traversable spatial region previously explored by the swarm. Let n denote a candidate grid cell and g the target destination. The total traversal cost associated with cell n is defined as:

$$C(n)=M_F(n)[w_d D(n) + w_t T(n)], \quad w_d, w_t \geq 0 \tag{21}$$

where w_d=0.5 and w_t=0.5 represent the relative importance assigned to distance minimization and threat avoidance, respectively.

By assigning equal weights to both terms, the algorithm achieves balanced navigation behavior that simultaneously considers traversal efficiency and environmental safety.

The geometric distance component is defined using the normalized Euclidean distance between the candidate node and the goal:

$$D(\mathbf{p})=\frac{\| \mathbf{p}_n - \mathbf{p}_g \|_2}{D_{max}} \tag{22}$$

where:

- $\mathbf{p}_n$ and $\mathbf{p}_g$ are the nodes position and goal position respectively
- D_{max} noting the maximum possible distance within the search space

The threat component $T(n)$ is defined as:

$$T(n)=\alpha_\tau \cdot \tau(n) \tag{23}$$

where:

- $\tau(n)$ is the normalized local threat level
- $\alpha = 10.0$ is a penalty scaling factor

This formulation ensures that regions associated with elevated environmental danger become significantly less desirable during path selection. As the local threat increases, the total traversal cost grows proportionally, discouraging the swarm from navigating through unsafe regions whenever safer alternatives exist.

An additional component of the proposed framework is the formation dependent motion penalty $M(n)$, which models the reduction in swarm mobility under increasing environmental threat conditions. Different formation profiles naturally exhibit different

movement characteristics and traversal efficiency. Compact defensive formations typically move more slowly due to increased cohesion requirements, while offensive formations prioritize rapid exploration and flexible movement. The motion penalty is therefore defined according to the currently active formation mode:

$$M_F(n) = \begin{cases} 1.0, & \tau(n) < 0.2\ (Offensive) \\ 1.33, & 0.3 \leq \tau(n) < 0.7\ (Neutral) \\ 2.0, & \tau(n) \geq 0.7\ (Defensive) \end{cases} \tag{24}$$

This adaptive penalty mechanism models the realistic reduction in swarm mobility observed under high threat conditions. Traversal through dangerous areas therefore incurs not only increased environmental risk but also additional movement cost associated with tighter and slower formation coordination.

The accumulated traversal cost from the starting location s to a candidate node n is represented through the A* g-score:

$$g(n_m)=\sum_{l=s}^{l} C(n_l), \qquad (n_0 = s, n_m = n) \tag{25}$$

while the evaluation function $E(n)$ that guides the search is defined as:

$$E(n)=g(n)+h(n) \tag{26}$$

where:

- $g(n)$ represents the accumulated traversal cost,
- $h(n)$ is a Euclidean heuristic estimating the remaining distance from the goal.

A priority queue is used to iteratively expand nodes with the minimum evaluation score first, enabling efficient path discovery while balancing both navigation efficiency and environmental safety. To further evaluate path quality, the proposed framework introduces the threat exposure ratio η, which quantifies the average environmental risk encountered along a traversal path:

$$\eta=\frac{1}{L}\sum_{l=1}^{L} \tau(\mathbf{p}_l) \tag{27}$$

where:

- L denotes the total number of path positions,
- $\tau\big(\mathbf{p}_i(t)\big)$ is the threat value experienced at position $\mathbf{p}_i(t)$

The threat exposure ratio provides an additional metric for evaluating navigation safety independently of total path length, enabling direct comparison between different path planning strategies.

For comparison purposes, the proposed balanced navigation approach is evaluated against a Safety First strategy that prioritizes threat avoidance over traversal efficiency. The Safety First algorithm assigns significantly larger weight to environmental risk, strongly discouraging traversal through dangerous regions even when longer paths are required. Its traversal cost function is defined as:

$$C_{Safety}(n) = 0.1D(n) + 50\tau(n) \tag{28}$$

where $D(n)$ is the distance component and the threat contribution dominates the overall navigation cost. Consequently, the Safety First strategy tends to generate highly conservative paths that minimize exposure to dangerous regions at the expense of traversal efficiency and exploration speed.

CHAPTER 5
REINFORCEMENT LEARNING FRAMEWORK

(Section 5.1 Problem Formulation as MDP)

The adaptive swarm coordination problem considered in this thesis is formulated as a Markov Decision Process (MDP), enabling autonomous agents to learn navigation and formation adaptation strategies through continuous interaction with the operational environment. The main objective of the learning framework is to allow swarm nodes to maximize exploration efficiency and navigation safety while minimizing threat exposure, unnecessary movement cost, and formation instability in stochastic three dimensional environments.

The use of an MDP formulation is particularly suitable for swarm intelligence systems because agent behavior evolves sequentially over time and each navigation decision directly affects future environmental states, swarm structure, and mission performance. At every time step, swarm agents observe local environmental conditions, select behavioral adaptations, and receive feedback through a reward mechanism that evaluates the quality of the selected actions. Through repeated interaction with the environment, the swarm gradually learns policies capable of balancing exploration, survivability, and coordinated movement.

Formally, the proposed reinforcement learning problem is represented by the tuple:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \tag{29}$$

where:

- $\mathcal{S}$ is the set of environmental states,
- $\mathcal{A}$ is the set of available actions,
- $P(s'|s,\alpha)$ represents the state transition probability,
- $R(s,\alpha)$ is the reward function,
- $\gamma \in [0,1]$ is the discount factor controlling the contribution of future rewards

The environment is considered partially observable because each node possesses only local sensing information rather than complete global awareness of the operational space. Consequently, agents must make decisions under uncertainty while simultaneously interacting with neighboring swarm members and dynamically changing environmental conditions. This decentralized perception model closely resembles natural swarm systems where global coordinated behavior emerges from localized interactions.

The state space combines positional, environmental, and swarm behavioral information. Let the state observed by an agent at time t be:

$$s_i(t) = (\mathbf{p}_i(t), \tau(\mathbf{p}_i(t)), F(t)) \in \mathcal{S} \tag{30}$$

where:

- $\mathbf{p}_i(t)$ is the node position
- $\tau(\mathbf{p}_i(t))$ denotes the local environment threat level,
- $F(t)$ represents the currently active swarm formation profile at time t.

The spatial coordinates allow the agent to reason about its relative location within the explored environment, while the threat value provides information regarding local environmental safety. The inclusion of the formation profile inside the state

representation allows reinforcement learning to adapt swarm behavior according to the currently active operational mode.

Unlike traditional navigation frameworks where reinforcement learning directly controls movement trajectories, the proposed approach applies learning at the behavioral level by adapting Boids hyperparameters. This design preserves the decentralized and emergent nature of swarm coordination while still enabling intelligent adaptation. The action space therefore consists of hyperparameter multiplier selections:

$$a_i(t) = \left(m_{s,i}(t), m_{a,i}(t), m_{c,i}(t)\right) \in \mathcal{A} \tag{31}$$

where:

- m_s controls seperation
- m_a controls alignment
- m_c controls cohesion

Specifically in this scenario the multipliers are selected from a discrete set that was chosen as initialization values throughout a trial-error procedure:

$$\mathcal{U} = \{0.0, 0.5, 1.0, 1.5, 2.0\}, \qquad \mathcal{A} \in \mathcal{U}^3 \tag{32}$$

allowing the swarm to dynamically strengthen or weaken local interaction rules according to environmental conditions. This mechanism enables the swarm to continuously reshape its collective movement behavior while maintaining local rule based coordination.

The transition dynamics $P(s'|s,\alpha)$ are influenced by several interacting factors, including obstacle proximity, neighboring swarm movement, active formation configuration, threat distribution, and environmental exploration state. Since swarm agents simultaneously influence one another through local interactions, the environment becomes highly dynamic and non stationary from the perspective of individual nodes. This significantly increases the complexity of the learning problem compared to single agent reinforcement learning scenarios. The reward function is designed to encourage safe, coordinated, and efficient swarm navigation. At each time step, the agent receives a scalar reward:

$$r_i(t) = -\lambda_\tau \tau(\mathbf{p}_i(\mathrm{t})) - \lambda_d \parallel \mathbf{p}_i(\mathrm{t}) - \mathbf{p}_g \parallel_2 - c_{step} + B \mathbb{1}\{\parallel \mathbf{p}_i(\mathrm{t}) - \mathbf{p}_g \parallel_2\} \leq R_{goal} \tag{33}$$

where:

- $\tau(\mathbf{p}_i(\mathrm{t}))$ is the threat intensity,
- λ_τ and λ_d are weight coefficients,
- c_{step} is a small movement penalty applied at every step,
- B is a positive reward associated with successful goal completion. The goal is considered reached when the goal position is within the sensing radius of the agent.

The threat penalty discourages traversal through dangerous regions, whereas the distance penalty promotes efficient navigation. The step penalty prevents excessively long trajectories, encouraging the swarm to discover more effective movement strategies.

An important characteristic of the proposed MDP formulation is that reinforcement learning does not replace swarm intelligence mechanisms. Instead, learning operates as a higher level adaptive component that continuously adjusts the intensity of local interaction behaviors. Consequently, the swarm preserves the scalability, robustness, and

emergent coordination properties of the Boids framework while simultaneously gaining adaptive decision making capabilities.

Overall, the proposed MDP formulation establishes the theoretical foundation for integrating reinforcement learning with decentralized swarm coordination. By combining environmental awareness, adaptive formation behavior, and local interaction dynamics within a unified learning framework, the proposed system aims to support robust and efficient swarm navigation in complex three dimensional operational environments.

The threat-exposure ratio η quantifies the accumulated environmental risk encountered along a path relative to the distance traveled. It is defined in discrete form as:

$$\eta=\frac{1}{L}\sum_{i=1}^{L}\tau(\mathbf{p}_i) \tag{34}$$

The Safety First algorithm finds paths that prioritize minimal threat exposure, heavily weighting threat avoidance over distance. Its cost function is:

$$C_{Safety}(\tau(\mathbf{p}_i(t)) = 0.1M_f(\tau(\mathbf{p}_i(t)) + 50\tau(\mathbf{p}_i(t)) \tag{35}$$

where $M_f(\tau(\mathbf{p}_i(t))$ is the distance component.

(Section 5.2 Policy Design)

The policy design of the proposed reinforcement learning framework focuses on enabling adaptive swarm behavior while preserving the decentralized and emergent characteristics of the Boids Flocking Algorithm. Instead of directly controlling the exact movement trajectory of each node, the reinforcement learning component operates at a higher behavioral level by dynamically modifying the intensity of local interaction rules. This approach allows the swarm to retain scalable and biologically inspired coordination properties while gaining adaptive decision making capabilities suitable for stochastic environments. Formally, the policy is defined as a mapping between environmental states and behavioral adaptation actions:

$$\pi: \mathcal{S} \rightarrow \mathcal{A} \tag{36}$$

where the policy π selects the most appropriate action for a given environmental state. In this framework, actions correspond to adaptive modifications of the Boids hyper parameters. These are the parameters defining cohesion, alignment and separation. By tuning these parameters, the formation profiles dynamically change as mentioned in the above section.

Unlike classical reinforcement learning navigation systems that directly generate low level movement commands, the proposed framework applies learning indirectly through behavioral adaptation. The resulting movement trajectories still emerge naturally from local interactions between neighboring agents rather than explicit centralized planning. The policy operates jointly with the previously defined formation profiles Offensive, Neutral, and Defensive. In practice, the formation profile determines the general swarm objective, whereas reinforcement learning adjusts the local movement behavior in order to improve navigation efficiency under current environmental conditions.

This adaptive mechanism enables context sensitive swarm coordination. In low threat environments, the policy may favor behaviors that increase exploration coverage and reduce unnecessary formation constraints. Conversely, in dangerous or highly constrained regions, the policy may strengthen formation cohesion and coordinated movement in order to improve swarm survivability and reduce fragmentation. Through continuous interaction with the environment, the swarm gradually learns how to balance movement flexibility, obstacle avoidance, and formation preservation.

An important characteristic of the proposed policy design is that learning remains decentralized. Each node evaluates its own local environmental observations and independently applies behavioral adaptations without requiring full global knowledge of the environment. Although agents operate independently, coherent global swarm behavior emerges collectively through neighboring interactions and shared environmental feedback. This decentralized structure improves scalability and fault tolerance while reducing communication overhead between swarm members.

The policy additionally supports smooth transitions between formation behaviors. This prevents excessive oscillations or sudden fragmentation that could negatively affect navigation efficiency and swarm cohesion. To balance exploration and exploitation during training, the framework employs an ϵ-greedy strategy. At each decision step, agents either select random actions to explore alternative behavioral adaptations or exploit the currently best known policy according to the learned Q-values:

$$\alpha_t = \begin{cases} uniformly\ sampled\ action\ from\ \mathcal{A}, & with\ probability\ \epsilon \\ arg\ max_{\alpha \in \mathcal{A}}\ Q(\boldsymbol{s}_t, \alpha), & with\ probability\ 1-\epsilon \end{cases} \tag{37}$$

This exploration mechanism prevents premature convergence to locally optimal behaviors and encourages broader policy discovery during training. Over time, the swarm gradually converges toward more stable behavioral adaptations that improve navigation safety and operational efficiency.

Overall, the proposed policy design establishes a hybrid swarm coordination framework where reinforcement learning and Boids based local interactions operate cooperatively rather than competitively. Reinforcement learning provides adaptive behavioral optimization, while decentralized Boids dynamics preserve emergent collective movement and scalable swarm coordination. This combination enables robust and adaptive navigation behavior within three dimensions.

(Section 5.3 Training Process)

The training process is designed to allow swarm agents to progressively learn adaptive behavioral strategies through repeated interactions. During training, agents continuously navigate, explore, and react to constant changing environmental conditions while reinforcement learning optimizes the behavioral adaptations applied to the swarm coordination model. The primary objective of training is to develop policies capable of balancing exploration efficiency, environmental safety, formation stability, and navigation performance. Each training episode begins with randomized initialization conditions in order to expose the swarm to diverse operational scenarios. The environment generation process varies obstacle placement, threat distribution, swarm initialization positions, and target locations between episodes. This variability is important because it prevents the learned policy from overfitting to a single environmental configuration and improves the generalization capabilities of the swarm under previously unseen conditions. At the beginning of each episode, swarm agents are initialized within the bounded operational space and assigned an initial formation profile according to the surrounding environmental threat conditions. The environment initially contains

partially unexplored regions, forcing the swarm to gradually construct a representation of the operational space through collective sensing and exploration. As agents navigate through the environment, explored regions, threat fields, and obstacle locations are progressively incorporated into the shared environmental representation. At every simulation time step t, each node performs the following sequence of operations:

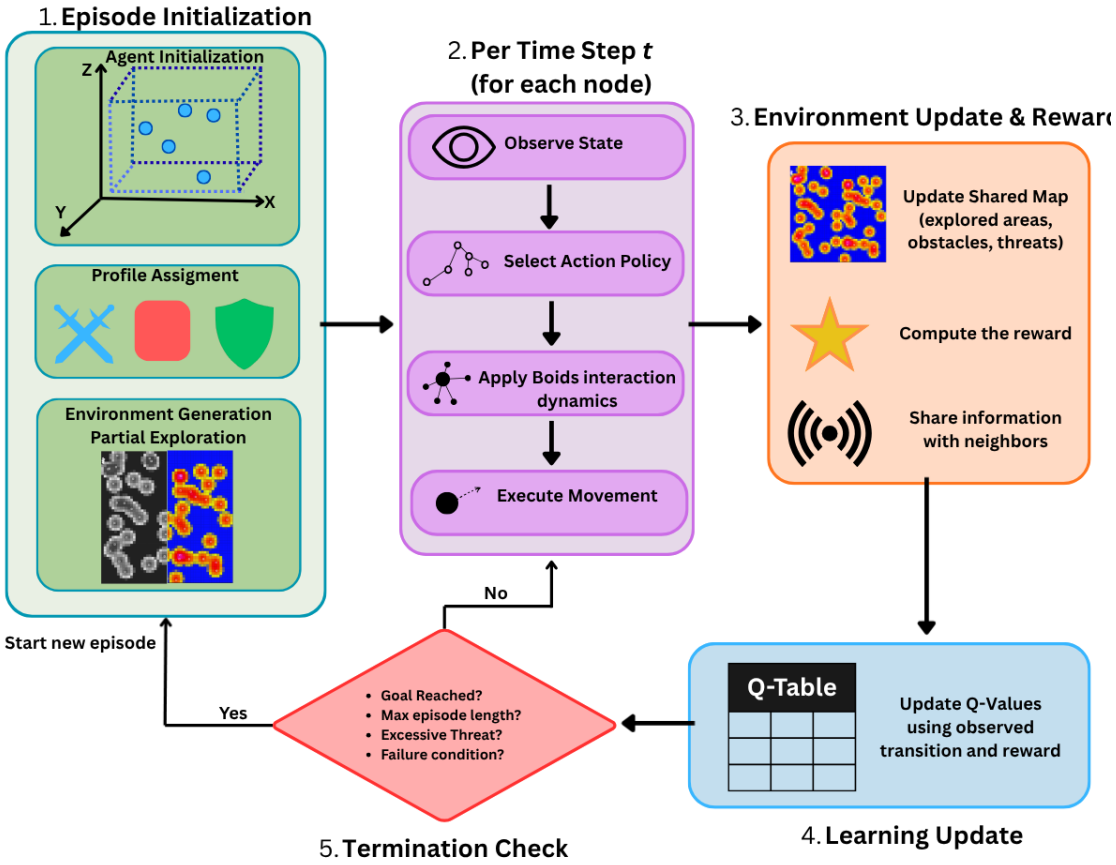

Figure 5: Workflow Diagram

This iterative interaction process allows the swarm to continuously refine its behavioral strategy through environmental feedback. The reinforcement learning framework employs tabular Q-learning in order to estimate the long term utility of behavioral adaptation actions under different environmental conditions. The Q-value update rule is defined as:

$$Q(\boldsymbol{s}_t, \alpha_t) \leftarrow Q(\boldsymbol{s}_t, \alpha_t) + \alpha[r_t + \gamma max_{\alpha' \in \mathcal{A}} Q(\boldsymbol{s}_t, \alpha') - Q(\boldsymbol{s}_t, \alpha_t)] \quad (38)$$

where:

- γ is the discount factor
- r is the immediate reward received after action execution

Through repeated updates, the Q-table gradually approximates the expected long term reward associated with selecting specific behavioral adaptations under different environmental conditions. The training process relies on continuous interaction between reinforcement learning and the Boids based swarm coordination model. After selecting an action, the resulting behavioral adaptations directly influence the swarm motion dynamics by modifying cohesion, alignment, and separation behavior. The updated movement behavior then affects future environmental states, exploration efficiency,

formation stability, and threat exposure. Consequently, the learning process becomes tightly coupled with the emergent swarm dynamics.

The swarm behavior during training is additionally influenced by the currently active formation profile. Offensive configurations generally encourage wider exploration and faster traversal, while Defensive configurations prioritize survivability and compact movement. The reinforcement learning framework therefore learns adaptive behavioral modifications within different strategic formation contexts rather than learning a single universal navigation behavior.

During early training stages, higher exploration probability encourages broader policy discovery and behavioral experimentation. As training progresses, the exploration rate gradually decreases, allowing the swarm to increasingly exploit the most successful behavioral adaptations discovered during previous episodes. Training episodes terminate when one of the following conditions is satisfied:

- The swarm reaches the designated target region
- The maximum allowed episode duration is exceeded
- Excessive threat exposure occurs
- Formation fragmentation or navigation failure is detected

Following episode termination, the environment is reinitialized and a new training episode begins. This episodic learning structure enables continuous policy refinement across a large number of environmental configurations and navigation scenarios. Several performance metrics are monitored throughout the training process in order to evaluate the quality of the learned policy. These metrics include:

- average episodic reward
- threat exposure ratio
- exploration coverage percentage
- swarm level fitness
- path traversal cost
- navigation success rate
- formation stability and cohesion

The objective of monitoring these metrics is not only to evaluate navigation efficiency but also to measure the swarm's ability to preserve coordinated behavior under environmental uncertainty. Policies associated with higher rewards, reduced threat exposure, improved exploration coverage, and stable formation maintenance are considered more successful.

Another important aspect of the training process concerns environmental uncertainty and partially unexplored space. Since agents cannot initially assume that unknown regions are traversable or safe, the swarm must continuously adapt its movement strategy according to newly discovered environmental information. This significantly increases the complexity of the learning problem because future navigation decisions depend heavily on the quality and completeness of the progressively constructed environmental map.

The integration of reinforcement learning with adaptive formation control highlights the importance of balancing individual flexibility with collective stability. Each node operates independently using local observations and decentralized decision making, yet the swarm as a whole remains capable of coordinated movement, environmental adaptation, and efficient navigation. This balance between local autonomy and global coordination is one of the defining characteristics of natural swarm systems and represents a key objective of the proposed framework.

Overall, the reinforcement learning framework presented in this chapter establishes the adaptive decision making foundation of the proposed system. The learned behaviors will subsequently be evaluated experimentally in the following chapter in terms of exploration performance, formation adaptation efficiency, threat avoidance capability, and navigation robustness.

CHAPTER 6
RESULTS AND DISCUSSION

(Section 6.2 Experimental Setup)

This chapter evaluates the effectiveness of the proposed framework through a series of controlled simulation experiments performed within stochastic three dimensional environments. The experimental evaluation focuses on the swarm's ability to adapt formation behavior, maintain coordinated navigation, avoid environmental threats, and efficiently explore previously unknown operational spaces. In addition, the experiments examine the impact of adaptive reinforcement learning driven coordination on swarm performance under different environmental conditions and navigation constraints.

The experimental process is divided into two major phases. During the **first phase**, the swarm performs exploration of the initially unknown environment while dynamically adapting formation behavior according to local environmental conditions. During the **second phase**, the generated environmental map is used for pathfinding and navigation evaluation using the proposed Balanced Navigation Algorithm and comparison methods including Dijkstra, A*, and Safety First navigation strategies. The swarm consists of multiple autonomous agents operating under decentralized control principles. Each node maintains only local environmental awareness through neighboring interactions and limited sensing capabilities. The environment contains vertically extended obstacle structures represented as column shaped spatial regions. Obstacles generate local threat fields that influence swarm navigation and formation adaptation behavior. During exploration, unexplored regions are treated as uncertain traversal areas, requiring agents to progressively construct an environmental representation before reliable path planning can occur. This constraint increases the realism of the navigation problem by preventing the swarm from assuming complete environmental knowledge from the beginning of the simulation. The reinforcement learning framework is trained using tabular Q-learning with discrete behavioral adaptation actions applied to Boids interaction parameters. Training is performed over multiple randomized episodes in order to expose the swarm to diverse environmental conditions and improve policy generalization. The learned policies are then evaluated according to exploration efficiency, threat exposure, formation stability, path traversal cost, and overall swarm fitness.

The primary simulation and swarm configuration parameters used throughout the experiments are summarized in the following table:

Parameter	Value	Description
Environment Dimensions	500x500x500	Size of the 3D operational space
Number of Swarm Agents	15	Total autonomous nodes
Number of Obstacles	45	Static environmental obstacles
Maximum Agent Speed	5.0	Maximum movement velocity
Maximum Steering Force	1.0	Upper steering constraint
Threat Detection Radius	35	Local threat sensing distance
Exploration Radius	7	Radius marked as explored
Map Resolution	50x50	Grid based exploration map

Obstacle Width	5	Obstacle geometric width
Obstacle Height Range	450-500	Vertical obstacle extension
High Threat Threshold	0.7	Switch to Defensive Profile
Low Threat Theshold	0.2	Switch to Offensive Profile

The Boids based formation profiles used during experimentation are presented in the following table:

Profile	Cohesion	Alignment	Separation	Behavioral Objective
Offensive	0.5	1.2	2.0	Exploration and spatial coverage
Neutral	1.0	1.0	1.2	Balanced coordination
Defensive	5.0	2.0	0.1	Survivability and compact movement

To evaluate navigation performance, four different pathfinding strategies were compared:

1. Dijkstra Algorithm
2. A* Algorithm
3. Safety First Navigation
4. Proposed Balanced Navigation Algorithm

Each algorithm was evaluated under Offensive, Neutral, Defensive, and Dynamic formation profile scenarios across multiple randomized navigation tasks.

The complete experimental evaluation consists of sixteen primary navigation scenarios generated from the combination of four pathfinding algorithms and four formation profile configurations. Each experiment was executed across multiple randomized start and goal locations within the explored operational environment in order to reduce bias associated with specific spatial configurations. Finally, 4 metrics were used on order to evaluate the performance of the experiments:

1. Path Cost
2. Path Length
3. Threat Exposure Ratio
4. Success Rate

Experiment Group	Pathfinding Algorithm	Formation Scenario
E1	Dijkstra	Offensive
E2	Dijkstra	Neutral
E3	Dijkstra	Defensive
E4	Dijkstra	Dynamic
E5	A*	Offensive

E6	A*	Neutral
E7	A*	Defensive
E8	A*	Dynamic
E9	Safety First	Offensive
E10	Safety First	Neutral
E11	Safety First	Defensive
E12	Safety First	Dynamic
E13	Balanced Navigation	Offensive
E14	Balanced Navigation	Neutral
E15	Balanced Navigation	Defensive
E16	Balanced Navigation	Dynamic

Assuming that N randomized navigation tasks are executed per scenario, the total number of pathfinding experiments performed during a complete evaluation cycle is:

$$E_{total} = 16 \times N_{tasks}$$

for the present study, $N = 200$, therefore $E_{total} = 16 \times 200 = 3200$. Thus, a total of **3200 pathfinding experiments** were conducted and **12800 metrics' measurements** collected for evaluation, enabling statistically meaningful comparison between navigation strategies and formation adaptation behaviors under stochastic environmental conditions.

(Section 6.2 Formation Profile Evaluation)

This section evaluates the impact of the proposed formation profiles on swarm navigation performance and environmental adaptability. The experiments compare Offensive, Neutral, Defensive, and Dynamic formation configurations under multiple pathfinding algorithms in order to examine how formation constraints influence traversal cost, path efficiency, and swarm adaptability within stochastic environments. The evaluation particularly focuses on the relationship between environmental threat conditions and swarm coordination behavior. Offensive formations prioritize exploration and movement flexibility, Defensive formations prioritize survivability and compact coordination, while the Dynamic profile adaptively switches between behaviors according to local environmental conditions. The obtained results demonstrate that formation behavior significantly influences overall navigation performance and traversal efficiency.

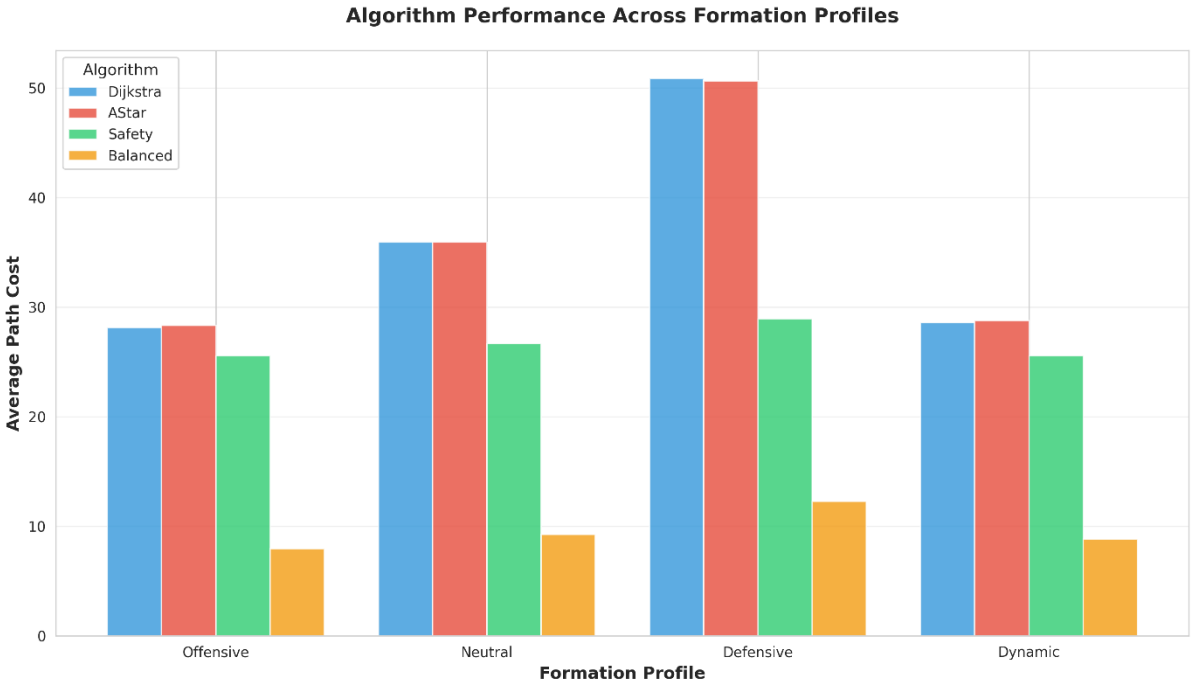

Figure 6: Algorithm Performance Across Formation Profiles

Figure 6 presents the average path traversal cost produced by each pathfinding algorithm under different formations. Several observations emerge from this comparison.

- First, all algorithms experience substantially increased traversal cost under the Defensive profile. This behavior is expected because defensive movement imposes stronger cohesion constraints, reduced movement flexibility, and tighter formation preservation requirements. Consequently, the swarm sacrifices traversal efficiency in favor of safety.
- In contrast, the Offensive profile consistently produces lower traversal costs because agents are allowed greater spatial flexibility and exploration freedom.
- The Neutral profile behaves as an intermediate configuration balancing movement efficiency and formation preservation.
- Most importantly, the Dynamic profile achieves lower average traversal cost compared to purely Defensive configurations while still maintaining adaptive threat aware behavior. This demonstrates the effectiveness of adaptive formation switching in reducing unnecessary movement penalties while preserving responsiveness to environmental conditions.
- Another important observation concerns the comparison between navigation algorithms. The proposed Balanced Navigation Algorithm consistently achieves the lowest traversal cost across all formation profiles. Its average cost remains significantly lower than Dijkstra and A* while also outperforming the highly conservative Safety First approach. This indicates that balancing environmental safety and traversal efficiency simultaneously produces more stable and efficient navigation behavior compared to strategies focusing exclusively on shortest path traversal or threat minimization.

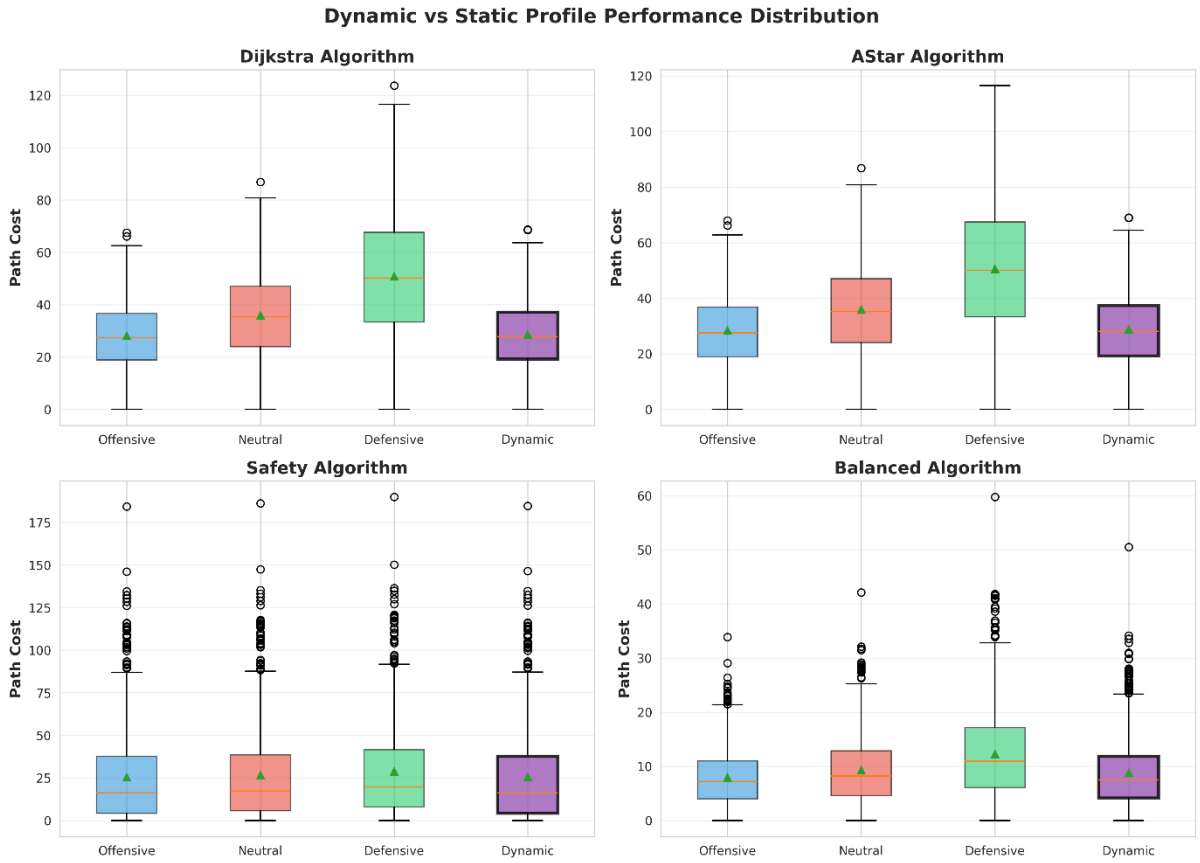

Figure 7: Dynamic vs Static Profile Performance Distribution

Figure 7 analyzes the statistical distribution of traversal costs across different formation profiles.

- The boxplot distributions reveal that Defensive profiles produce both larger median traversal costs and greater variance compared to Offensive and Dynamic profiles. This indicates that strict formation preservation increases not only average traversal cost but also navigation instability under varying environmental conditions.
- The Dynamic profile distributions are particularly important because they remain noticeably more compact and stable than the purely Defensive configuration. This suggests that adaptive formation switching, allows the swarm to avoid excessive movement penalties by selectively applying defensive behavior only when required by local environmental threats.
- The Balanced Navigation algorithm again demonstrates the narrowest and most stable distribution among all compared methods, indicating improved robustness and consistency across different environmental conditions.

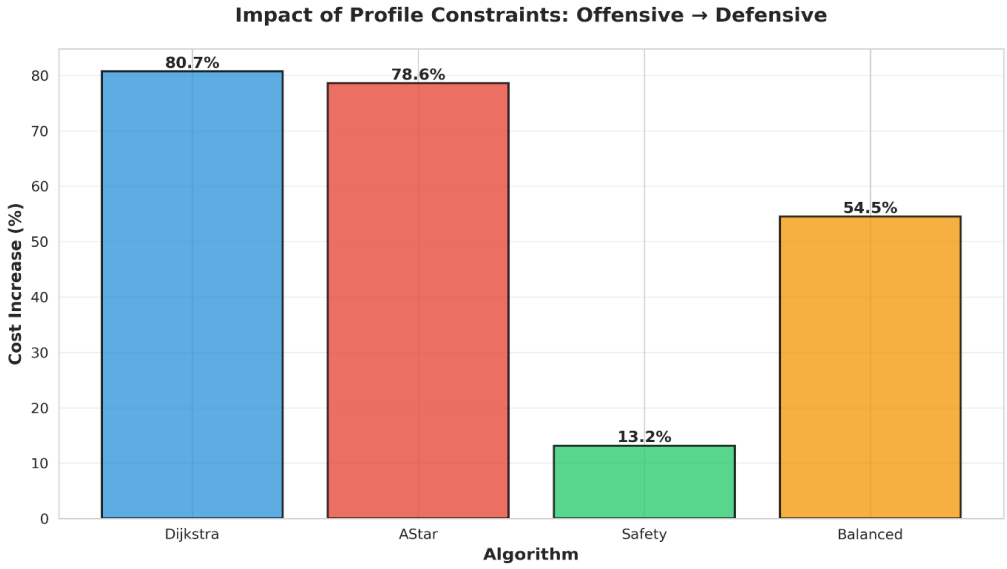

Figure 8: Impact of Profile Constraints: Offensive to Defensive

Figure 8 explicitly quantifies the increase in traversal cost when transitioning from Offensive to Defensive swarm behavior.

- The results clearly demonstrate that stronger formation constraints substantially increase navigation cost for all algorithms. Dijkstra and A* experience the largest cost increase, exceeding approximately 78 80%, indicating that shortest path oriented algorithms are strongly affected by reduced movement flexibility under compact defensive coordination.
- The Safety First algorithm exhibits a much smaller relative increase because its navigation strategy already prioritizes threat avoidance and conservative traversal behavior even under less restrictive profiles.
- The Balanced Navigation Algorithm shows intermediate behavior, with a moderate increase compared to Dijkstra and A*, demonstrating that the proposed method adapts more efficiently to restrictive formation conditions without excessively sacrificing navigation performance.

Overall:

- Static formation configurations showcase predictable movement behavior but may introduce unnecessary constraints under changing environmental conditions.
- In contrast, the Dynamic profile allows the swarm to continuously balance exploration flexibility and survivability according to local threat conditions, resulting in more efficient and context aware collective behavior.

(Section 6.3 Pathfinding Algorithms Evaluation)

This section evaluates the performance of the examined pathfinding algorithms under stochastic environmental conditions and adaptive swarm coordination constraints. The comparison includes Dijkstra, A*, Safety First, and the proposed Balanced Navigation Algorithm. The evaluation focuses on traversal cost efficiency, path length behavior, environmental threat exposure, navigation robustness, and adaptability under different formation profiles.

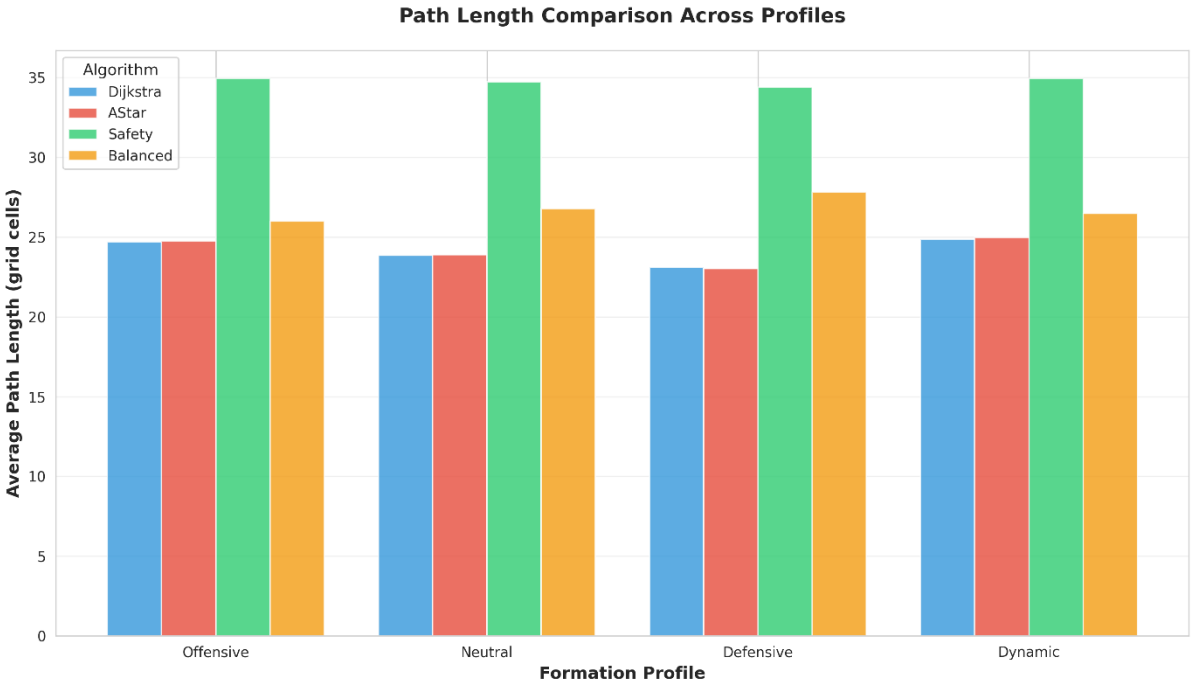

Figure 9: Path Length Comparison Across Profiles

Figure 9 compares the average path length produced by each navigation algorithm under Offensive, Neutral, Defensive, and Dynamic formation configurations.

- Dijkstra and A* consistently produce the shortest paths across all formation profiles because both algorithms prioritize geometric efficiency and direct traversal toward the goal location. However, these shorter paths frequently result in increased exposure to environmental threats and higher overall traversal costs under restrictive formation conditions.
- The Safety First algorithm consistently generates the longest paths because it strongly prioritizes threat avoidance over traversal efficiency. By aggressively avoiding dangerous regions, the algorithm often selects highly indirect routes that significantly increase traversal distance. Although this behavior improves environmental safety, it substantially reduces navigation efficiency and increases operational movement cost.
- The proposed Balanced Navigation Algorithm demonstrates intermediate path length behavior. While its generated paths are slightly longer than those of Dijkstra and A*, they remain significantly shorter than Safety First paths. This behavior reflects the core objective of the proposed approach: balancing traversal efficiency and environmental safety simultaneously rather than optimizing only one objective.

Figure 10: Algorithm Ranking by Cost Profile

Figure 10 directly compares the average traversal cost of each algorithm under the different formation profiles.

- Across all configurations, the proposed Balanced Navigation Algorithm consistently achieves the lowest average cost. This shows that combining threat awareness and distance minimization within a balanced cost function produces more efficient overall navigation than either purely shortest path or purely safety oriented approaches.
- Dijkstra and A* demonstrate very similar behavior across all profiles because both algorithms primarily optimize geometric path efficiency. However, their traversal costs increase significantly under Defensive profiles due to reduced swarm mobility and increased environmental constraints.
- The Safety First algorithm achieves lower cost than Dijkstra and A* under several profiles because aggressive threat avoidance reduces exposure penalties, although this occurs at the expense of substantially longer traversal paths as discussed previously.
- The Dynamic profile results are particularly important because they demonstrate that adaptive formation switching improves navigation efficiency across all algorithms. Under Dynamic behavior, traversal costs remain noticeably lower than purely Defensive configurations while still preserving adaptive environmental responsiveness. This observation confirms that

adaptive coordination provides significant advantages compared to static swarm behavior.

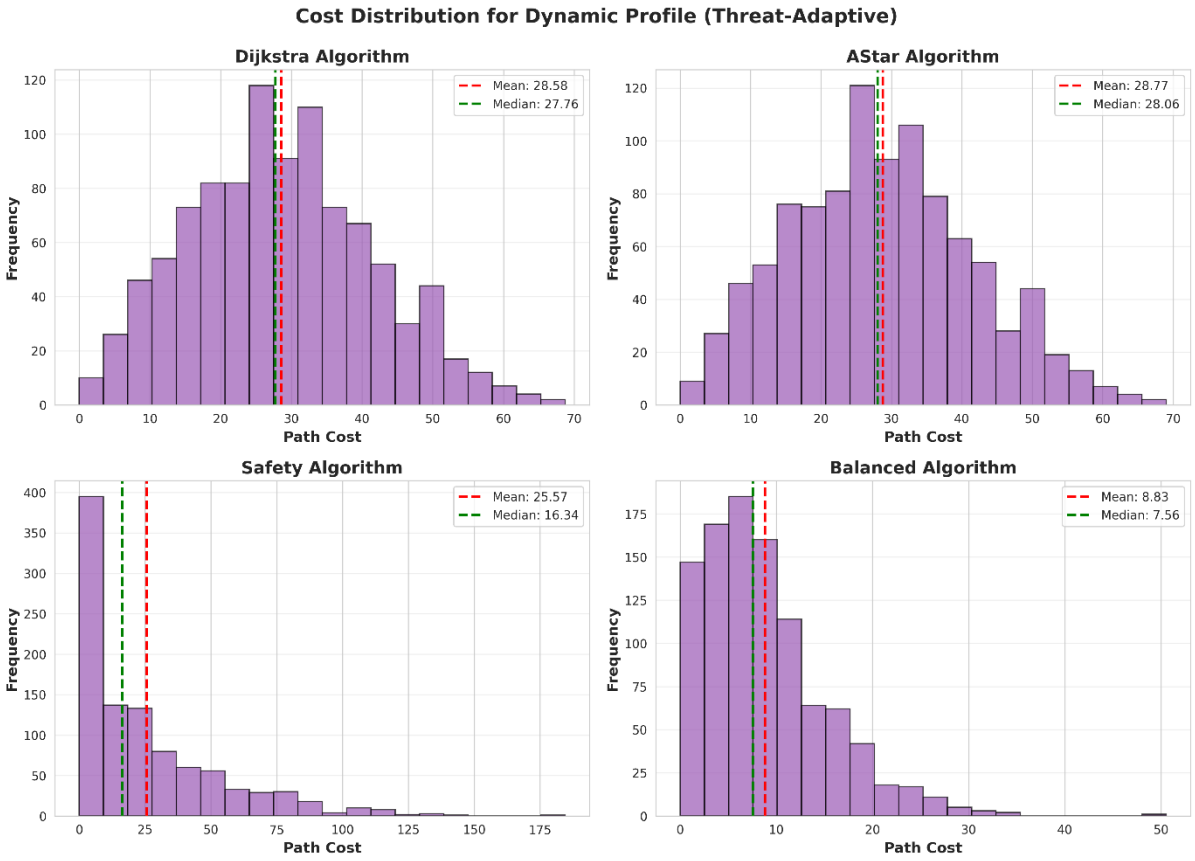

Figure 11: Cost Distribution for Dynamic Profile (Threat Adaptive)

Figure 11 examines the statistical distribution of traversal costs specifically under adaptive formation behavior.

- Dijkstra and A* produce relatively symmetric cost distributions centered around higher average traversal costs.
- The Safety First algorithm exhibits a heavily skewed distribution with large variance and numerous high cost outliers, reflecting the instability introduced by excessively conservative threat avoidance behavior.
- In contrast, the Balanced Navigation Algorithm produces the most compact and stable distribution with significantly lower average and median traversal costs. The close proximity between mean and median values additionally suggests improved consistency and reduced sensitivity to extreme environmental conditions. This result demonstrates that the proposed balanced strategy not only reduces average traversal cost but also improves navigation reliability across varying operational scenarios.

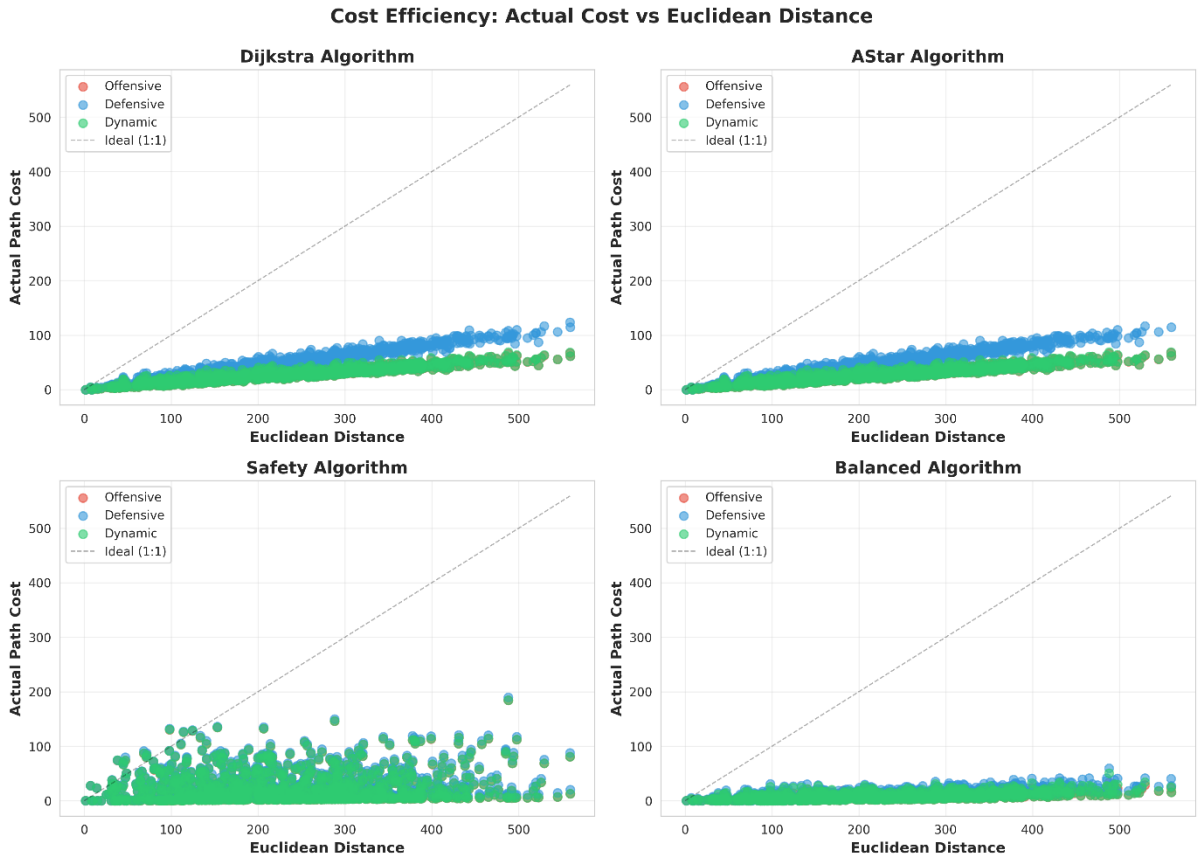

Figure 12: Cost Efficiency: Actual Cost vs Euclidian Distance

Figure 12 evaluates how efficiently each algorithm transforms geometric distance into traversal cost under different formation constraints.

- Dijkstra and A* demonstrate strong linear growth between Euclidean distance and actual path cost, particularly under Defensive profiles where traversal penalties become increasingly significant.
- The Safety First algorithm displays highly scattered behavior with large variability and weak proportionality between geometric distance and traversal cost. This reflects the algorithm's strong dependence on environmental threat avoidance rather than geometric efficiency.

In many cases, relatively short geometric distances still produce excessively high traversal cost because the algorithm avoids dangerous regions through highly indirect movement trajectories.

- The proposed Balanced Navigation Algorithm exhibits the most stable proportional relationship between geometric distance and traversal cost. The majority of traversal samples remain concentrated near the ideal efficiency region while still respecting environmental threat constraints. This behavior indicates that the proposed method achieves a more effective compromise between path optimality and environmental safety compared to the other examined strategies.
- Overall, the pathfinding evaluation demonstrates that the proposed Balanced Navigation Algorithm consistently provides the most stable and efficient navigation behavior across varying formation profiles and environmental conditions.

Unlike shortest path oriented methods that neglect environmental risk or purely defensive approaches that excessively sacrifice efficiency, the proposed framework

successfully balances traversal distance, threat exposure, and swarm movement constraints within a unified navigation strategy. The results additionally confirm the importance of integrating adaptive formation behavior directly into swarm navigation and path planning processes. Navigation performance is strongly influenced not only by the pathfinding algorithm itself, but also by the swarm's collective movement behavior and environmental adaptability. Consequently, the combination of reinforcement learning based formation adaptation and balanced threat aware navigation provides a robust framework for autonomous swarm coordination in complex three dimensional operational environments.

CHAPTER 7
CONCLUSION AND FUTURE IMPROVEMENTS

This thesis investigated adaptive formation aware navigation and path planning for swarm systems operating in stochastic three dimensional environments under heterogeneous environmental constraints. The proposed framework combined decentralized Boids based swarm coordination, adaptive formation control, reinforcement learning driven behavioral adaptation, and threat aware path planning within a unified navigation architecture. By integrating these components, the system was designed to support efficient exploration, coordinated movement, and robust navigation while maintaining scalability and decentralized behavior.

A major contribution of this work lies in the integration of adaptive formation behavior directly into the navigation process. Opposing to traditional swarm coordination approaches that rely on static formations or predefined movement strategies, this approach enables agents to dynamically transition between behavioral profiles according to environmental feedback and local threat conditions. This adaptive behavior allows the swarm to continuously balance exploration efficiency, survivability, and formation stability.

The experimental evaluation demonstrated that formation behavior significantly influences navigation performance and traversal efficiency. Offensive formations provided improved exploration flexibility and reduced traversal cost, while Defensive formations increased survivability and environmental robustness at the expense of movement efficiency. Most importantly, the Dynamic approach consistently achieved improved balance between these competing objectives by allowing the swarm to adapt its behavior according to environmental conditions. These findings highlight the importance of adaptive swarm coordination in realistic scenarios where environmental threats and navigation constraints continuously appear over time.

The pathfinding experiments demonstrated the effectiveness of the proposed Balanced Navigation Algorithm. Compared to traditional shortest path methods such as Dijkstra and A*, as well as highly conservative threat avoidance strategies such as Safety First navigation, the proposed approach consistently achieved lower average traversal costs with reduced variance across all evaluated formation profiles. The results further showed that the Balanced algorithm maintained a more favorable relationship between geometric distance and actual traversal cost, indicating superior optimization behavior under varying environmental conditions and formation constraints.

Another important outcome of this work is the demonstration that reinforcement learning can successfully augment decentralized swarm intelligence without replacing the emergent characteristics of local rule based coordination. Instead of directly controlling agent trajectories through centralized policies, reinforcement learning was used to adapt the intensity of local interaction dynamics, allowing coordinated global behavior to emerge naturally from neighboring interactions. This hybrid approach preserves the robustness, scalability, and flexibility commonly associated with natural swarm systems while simultaneously introducing adaptive high level decision making capabilities.

The proposed framework also demonstrated the importance of environmental awareness in swarm navigation. By incorporating threat fields, opportunity fields, obstacle influenced costs into the decision making process, the swarm was able to perform more realistic and context sensitive navigation. The experiments showed that considering environmental uncertainty directly during path planning significantly improves navigation robustness and operational adaptability compared to purely geometric planning approaches.

Despite the promising results, several limitations remain and motivate future research directions. The current framework assumes static obstacles and simplified environmental dynamics, whereas real world environments may contain moving threats, dynamic obstacles, uncertain operational conditions, and communication instability between swarm members. Extending the framework to support even more dynamic environmental modeling and real time environmental updates represents an important next step toward realistic deployment scenarios. Future work could additionally investigate more advanced reinforcement learning methodologies capable of supporting larger scale multi agent coordination and continuous action spaces. Deep reinforcement learning and multi agent reinforcement learning approaches may enable more sophisticated adaptive behaviors and improved generalization across more complex environments.

Finally, the transition from simulation to physical robotic deployment remains an important long term objective. Evaluating the proposed framework on real robotic platforms would allow investigation of practical challenges such as:

- Sensor noise
- Localization uncertainty
- Heterogenous swarm structure
- Energy constraints and
- Real world communication delays

Such validation would provide valuable insight into the operational feasibility of adaptive formation aware swarm coordination in realistic autonomous systems.

Overall, the proposed framework contributes toward the development of more adaptive, scalable, and context aware swarm systems capable of operating effectively under uncertainty and dynamic environmental conditions.

REFERENCES

- Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. In Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '87) (pp. 25 34). Association for Computing Machinery. https://doi.org/10.1145/37401.37406

- Ouyang, Q., Wu, Z., Cong, Y., & Wang, Z. (2021). Formation control of unmanned aerial vehicle swarms: A comprehensive review. IEEE Access, 9, 34572 34595. https://doi.org/10.1109/ACCESS.2021.3060864

- Sakano, T., Ojetunde, B., Suzuki, M., Temma, K., Nakamura, A., & Fukada, T. (2025). A portable and elastic edge computing network for disaster first responders. In 2025 27th International Conference on Advanced Communications Technology (ICACT) (pp. 357 363). IEEE. https://doi.org/10.23919/ICACT63707.2025.10908459

- Crespo, A. F., Silva, F. D., & Lima, P. S. (2017). A survey on swarm robotics: Algorithms, challenges, and applications. Future Generation Computer Systems, 68, 35 50. https://doi.org/10.1016/j.future.2017.08.060

- Santos, M. F., & Almeida, J. S. (2021). A swarm intelligence approach to a multi-node system for autonomous exploration in unknown environments. Future Generation Computer Systems, 115, 306 317. https://doi.org/10.1016/j.future.2021.05.001

- Hengstebeck, C., Jamieson, P., & Van Scoy, B. (2024). Extending Boids for safety-critical search and rescue. Franklin Open, 8, 100160. https://doi.org/10.1016/j.fraope.2024.100160

- Huh, J. H., & Seo, Y. S. (2019). Understanding edge computing: Engineering evolution with artificial intelligence. IEEE Access, 7, 164229 164245. https://doi.org/10.1109/ACCESS.2019.2951055

- Zeng, Y., Zhang, R., & Lim, T. J. (2020). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. IEEE Communications Magazine, 54(5), 36 42. https://doi.org/10.1109/MCOM.2016.7470933

- Li, B., Fei, Z., & Zhang, Y. (2021). UAV communications for 5G and beyond: Recent advances and future trends. IEEE Internet of Things Journal, 6(2), 2241 2263. https://doi.org/10.1109/JIOT.2018.2796248

- Lu, J., Zhao, J., & Niu, J. (2025). Boids-based integration algorithm for formation control and obstacle avoidance in unmanned aerial vehicles. Machines, 13(4), 255. https://doi.org/10.3390/machines13040255

- Zhao, Y., & Wang, L. (2023). Obstacle avoidance of a wheeled robotic swarm using virtual spring dynamics. Swarm Intelligence, 17(3), 181 200. https://doi.org/10.1007/s11721-023-00221-8

- Liu, X., & Chen, M. (2015). Bioinspired obstacle avoidance algorithms for robot swarms. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1234 1240). IEEE. https://doi.org/10.1109/IROS.2015.7353521

- Guan, Y., & Li, X. (2020). Obstacle avoidance and path planning methods for autonomous mobile robots: A review. Sensors, 20(11), 3573. https://doi.org/10.3390/s20113573

- Huang, Z., Yang, Z., Krupani, R., Şenbaşlar, B., Batra, S., & Sukhatme, G. S. (2023). Collision avoidance and navigation for a quadrotor swarm using end-to-end deep reinforcement learning (arXiv Preprint No. arXiv:2309.13285). arXiv. https://doi.org/10.48550/arXiv.2309.13285

- Yang, X., Hu, Y., Gao, H., Ding, K., Li, Z., Zhu, P., Sun, Y., & Liu, C. (2024). Risk-aware non-myopic motion planner for large-scale robotic swarm using CVaR constraints (arXiv Preprint No. arXiv:2402.16690). arXiv. https://doi.org/10.48550/arXiv.2402.16690

- Bansla, V., et al. (2024). Optimizing task execution in robotic swarms: Coordination strategies for enhanced efficiency. In Proceedings of the 7th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 1095 1100). IEEE. https://doi.org/10.1109/IC3I59117.2024.10550321

- Ren, W., & Sorensen, N. (2008). Distributed coordination architecture for multi-robot formation control. Robotics and Autonomous Systems, 56(4), 324 333. https://doi.org/10.1016/j.robot.2007.08.005

- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2020). Multi-agent actor-critic for mixed cooperative-competitive environments (arXiv Preprint No. arXiv:1706.02275). arXiv. https://doi.org/10.48550/arXiv.1706.02275

- Li, Z., Wu, L., Su, K., Wu, W., Jing, Y., Wu, T., Duan, W., Yue, X., Tong, X., & Han, Y. (2024). Coordination as inference in multi-agent reinforcement learning. Neural Networks, 172, 106101. https://doi.org/10.1016/j.neunet.2024.106101