drones

MDPI

Article

Flight Plan Optimisation of Unmanned Aerial Vehicles with Minimised Radar Observability Using Action Shaping Proximal Policy Optimisation

Ahmed Moazzam Ali, Adolfo Perrusquía *, Weisi Guo and Antonios Tsourdos

Faculty of Engineering and Applied Sciences, Cranfield University, College Road, Bedford MK43 0AL, UK; ahmed.ali.568@cranfield.ac.uk (A.M.A.); weisi.guo@cranfield.ac.uk (W.G.); a.tsourdos@cranfield.ac.uk (A.T.)
* Correspondence: adolfo.perrusquia-guzman@cranfield.ac.uk

Abstract: The increasing use of unmanned aerial vehicles (UAVs) is overwhelming air traffic controllers for the safe management of flights. There is a growing need for sophisticated path-planning techniques that can balance mission objectives with the imperative to minimise radar exposure and reduce the cognitive burden of air traffic controllers. This paper addresses this challenge by developing an innovative path-planning methodology based on an action-shaping Proximal Policy Optimisation (PPO) algorithm to enhance UAV navigation in radar-dense environments. The key idea is to equip UAVs, including future stealth variants, with the capability to navigate safely and effectively, ensuring their operational viability in congested radar environments. An action-shaping mechanism is proposed to optimise the path of the UAV and accelerate the convergence of the overall algorithm. Simulation studies are conducted in environments with different numbers of radars and detection capabilities. The results showcase the advantages of the proposed approach and key research directions in this field.

Keywords: UAVs; proximal policy optimisation (PPO); action-shaping; radar detection; Neyman–Pearson criterion; path planning

check for updates

Citation: Ali, A.M.; Perrusquía, A.; Guo, W; Tsourdos, A. Flight Plan Optimisation of Unmanned Aerial Vehicles with Minimised Radar Observability Using Action Shaping Proximal Policy Optimisation. *Drones* **2024**, *8*, 546. https://doi.org/10.3390/drones8100546

Academic Editor: Abdessattar Abdelkefi

Received: 31 August 2024
Revised: 26 September 2024
Accepted: 29 September 2024
Published: 1 October 2024

1. Introduction

Unmanned Aerial Vehicles (UAVs), commonly known as drones, have revolutionised modern transport and smart living applications [1]. Their ability to perform a variety of missions ranging from surveillance, package delivery, search and rescue, and reconnaissance, among others; have made them indispensable devices [2]. However, the increasing reliance on UAVs has also introduced significant challenges, particularly in the domain of safe and efficient navigation [3] through contested environments where radar systems are actively employed to detect and control the air traffic.

Radar systems are a primary method of detecting and tracking UAVs [4]. These systems emit electromagnetic waves that reflect off objects, allowing the detection of their presence, speed, and trajectory [5]. Here, due to the increasing use of drones across many applications, it has increased as well the cognitive load in the air traffic control system. This, in consequence, yields the following: (i) delays in flights, (ii) economic loss, (iii) accidents between UAVs or key assets, and (iv) issues in prioritising UAVs over others. There is currently a flourishing community that aims to reduce the cognitive burden of air traffic controllers by developing new path planning techniques that ensure the execution of the mission, whilst avoiding its detection.

UAV path planning [6] refers to the process of determining an optimal or feasible route for an Unmanned Aerial Vehicle (UAV) to follow from its starting point to a designated target or series of waypoints [7]. The objective is to navigate the UAV through its environment, while adhering to certain constraints, such as avoiding obstacles [8], minimising exposure to threats (like radar detection in contested environments) [9], optimising flight time [10], energy consumption, or maintaining safe flight conditions [11].

Path planning algorithms for UAVs typically focus on optimising flight paths based on factors such as distance [12], fuel efficiency, or specific mission objectives [13]. However, these methods often overlook the complexities of avoiding detection by advanced radar systems. While the UAVs under consideration in this research are not inherently stealthy, the techniques developed can also be applied to future stealth UAVs. As UAVs increasingly operate in congested environments where radar surveillance is pervasive, there is a critical need for path planning methods that integrate both mission efficiency and radar evasion [14]. In response to this challenge, researchers have explored a range of strategies, from heuristic-based algorithms to machine learning techniques. These approaches aim to enhance the ability of UAVs to navigate effectively in contested airspaces, minimising the risk of detection while achieving mission goals.

2. Related Work

Due to the advancement of radar detection systems, especially those capable of identifying stealth aircraft, there is an increasing need for sophisticated path-planning approaches that also minimise radar detection. The literature has evolved from basic heuristic algorithms to advanced machine-learning techniques to address this challenge.

2.1. Heuristic-Based Path Planning Approaches

Heuristic-based algorithms like A^* [15] and Dijkstra [16] algorithms are widely adopted for UAV path planning due to their simplicity and efficiency. These algorithms prioritize finding the shortest path to the target [17], often disregarding the risk of radar detection. A modified A^* algorithm has been used for fighter aircrafts [18]. The approach uses the values of the radar cross-section (RCS) to minimise the detection risk. By factoring in the RCS and terrain elevation, this approach demonstrates superior performance in planning paths that evade detection from advanced radar systems. However, while effective in certain scenarios, the A^* algorithm may not always adapt well to dynamic environments where radar positions or detection ranges can change unpredictably.

The sparse A^* algorithm [19] addresses the limitations of traditional A^* by incorporating sparse graphs to reduce computational complexity [20]. The algorithm introduces a heuristic that does not only consider the shortest path but also integrates a cost function that factors in radar detection probabilities [21]. This approach shows promise in balancing the trade-off between path length and detection risk, particularly in complex environments.

2.2. Reinforcement Learning-Based Path Planning

In recent years, reinforcement learning (RL) [22] has emerged as a powerful tool for autonomous systems, including UAVs [23]. By allowing UAVs to learn from interactions with their environment, RL-based models can develop strategies for avoiding radar detection while achieving mission objectives [24]. Proximal Policy Optimisation (PPO) [25], a popular RL algorithm, has been applied to UAV path planning with considerable success. PPO enables UAVs to learn optimal paths through trial and error, adjusting their strategies based on rewards and penalties associated with radar detection and mission completion [26]. However, traditional RL methods often require substantial computational resources and extensive training times [27], making them less practical for real-time or resource-constrained applications.

Q-learning has been applied to minimise the probability of detection by the sonar systems [28]. The success of this method in underwater environments suggests that similar techniques could be effectively applied to aerial environments, where radar detection is the primary concern. Other approaches use PPO for path planning [29]. The authors highlight the limitations of traditional PPO, particularly its susceptibility to poor convergence due to high variance in reward signals. To address this, they propose an enhanced version, FD-PPO, which decomposes rewards into frequency components, improving the stability and convergence of the learning process. The results indicate that FD-PPO outperforms standard PPO in navigating complex environments with minimal radar detection risk.

Recurrent neural networks such as the long-short term memory (LSTM) network have been used to address the limitations of single-step decision-making in existing deep reinforcement learning-based UAV path planning [30]. A Real-time Path Planning based on Long Short-Term Memory network [31] was proposed to leverage the memory capabilities of LSTM networks within the Deep Q-Network (DQN) framework. The RPP-LSTM algorithm represents a significant improvement in UAV path planning by incorporating the memory capabilities of LSTM networks into the deep reinforcement learning framework. This allows for more effective and adaptive decision-making in dynamic environments, offering improved performance over traditional methods.

The reviewed literature underscores the evolution of UAV path planning from simple heuristic methods to advanced RL-based approaches. While traditional algorithms like A^* have been adapted to consider radar detection, RL-based methods such as PPO and its variants offer more dynamic and adaptable solutions. The integration of RL with other deep learning techniques, such as LSTM, further enhances the ability of UAVs to navigate complex environments while evading detection. However, challenges remain in terms of computational demands and real-time applicability, which are critical areas for future research.

Despite the good solutions offered by RL agents there is an issue in terms of the smoothness of the final control policy as well as the time dedicated to train the RL agents. Reward shaping [32] has been widely used to improve the performance of the RL agents by adapting the reward function based on heuristic mechanisms [33]. Reward shaping has been exploited in multi-task problems to address the catastrophic forgetting of RL models [34]. Imitation learning [35] and inverse reinforcement learning [36] approaches have been also explored to improve the reward design based on expert trajectories. These models require a considerable set of expert demonstrations to infer the reward or policy from the data [37]. Primitive-based learning [38] has been applied to leverage preferences and learn a reward function model. Another technique that has also been applied to improve the performance of RL agents is action shaping [39]. The most simple approach is to discretize continuous actions (DC) into discrete actions, which is widely applied in tabular RL methods. Removing actions (RA) has been used to remove unnecessary actions [40] which notably accelerates the training convergence of RL agents.

2.3. Contributions

In view of the above, this paper aims to contribute to the field of UAV path planning approaches in radar-contested environments through the development and assessment of algorithms that prioritize radar detection avoidance. This is conducted by improving the decision-making capabilities of the standard PPO algorithm by integrating an action-shaping mechanism. This mechanism restricts the UAV action space to a subspace whose actions strategically bring UAV closer to the target goal. In contrast to previous path planning approaches using RL that learn the complete path to the target we apply only the RL agent in zones of radar detection, whilst non-detection zones are driven by the action shaping mechanism that aims to minimise the distance between the UAV and the target goal. This approach enhances the practicality of reinforcement learning in real-world scenarios where UAVs must operate under constrained conditions. Diverse simulation studies are conducted to show the benefits of the work.

The main contributions of this work are twofold

- A novel path-planning approach based on the PPO algorithm and an action-shaping mechanism that accelerates learning and avoids radar detection.
- A radar warning zone switching criteria is developed based on the Neyman–Pearson Criteria to improve the action selection in both warning and non-warning radar detection zones.

3. Problem Formulation

Figure 1 shows the high-level diagram of the proposed radar-evasion problem. The diagram consists of a 2D environment that models a contested environment with several radars with different detection range. The key goal of this paper is to design a path planning algorithm that is capable of driving a UAV to reach a target point without being detected by the radars located in the 2D environment, whilst minimising the travelled distance.

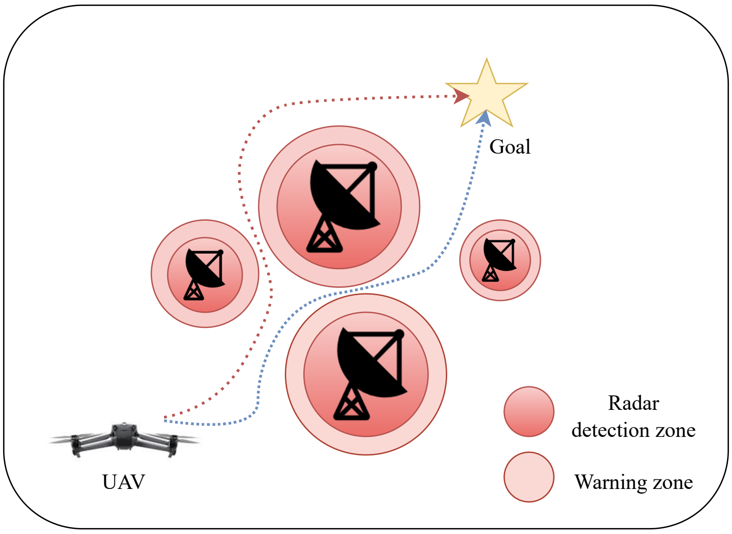


Figure 1. High-level diagram of the proposed radar evasion problem.

3.1. UAV Model

The UAV model is based on a simple kinematic model in the 2D plane, that is, we neglect the movement on the z-axis. The following kinematic model is considered

$$\begin{aligned} x_{k+1} &= x_k + T\cos(q)v \\ y_{k+1} &= y_k + T\sin(q)v, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_{UAV_k} = [x_k, y_k]^\top \in \mathbb{R}$ are the position coordinates of the UAV in time instance k, $q \subset \mathbb{R}$ is the orientation of the UAV, $T > 0$ is a sampling time, and $v \in \mathbb{R}$ is a nominal linear velocity. It is assumed that the UAV flies with a constant velocity. The orientation q of the UAV is used as a control input to drive the UAV to the desired target goal. In this paper, we set $T = 100$ s and $v = 10$ m/s to model that at each time step the UAV advances 1 km.

3.2. Technical and Theoretical Concepts

In order to design the proposed path planning algorithm, we need to consider key technical and theoretical concepts that allow a comprehensive understanding of the challenges involving the overall application. This section aims to outline these concepts that we consider in the proposed algorithm design.

3.2.1. Radar Detection

Radar systems are critical components of modern military defence, designed to detect, track, and identify airborne objects over long distances [41]. The effectiveness of radar is typically measured by its ability to detect an object, which is influenced by factors such as the Radar Cross Section (RCS) of the target, the distance between the radar and the target, and environmental conditions. The RCS is a measure of how detectable an object is by radar; it depends on the object's size, shape, material, and orientation relative to the radar. Stealth technology aims to reduce the RCS of an aircraft, making it less detectable by radar.

This is achieved through various design features such as angular shapes that deflect radar waves away from the source, radar-absorbent materials, and careful management of heat and other emissions. However, as radar technologies advance, even stealth aircraft face challenges in remaining undetected, particularly when operating in environments with multiple, overlapping radar systems [42]. In this research, while the UAV under consideration is not inherently designed as a stealth vehicle, the principles of radar detection avoidance are still applicable. The UAV's path planning must account for radar detection probabilities and strategically avoid areas with high radar coverage.

3.2.2. Neyman–Pearson Criterion

The Neyman–Pearson criterion [43] is a fundamental concept in statistical hypothesis testing, particularly in the field of signal detection theory, where it is often applied to radar detection scenarios. The criterion provides a rigorous method for making decisions based on observed data, specifically in determining whether to accept or reject a null hypothesis (typically that a signal is not present) in favour of an alternative hypothesis (that a signal is present), In the context of radar detection, it provides a framework for making decisions about the presence of a signal (e.g., whether a radar has detected the UAV) based on observed data while controlling for the probability of false alarms. In radar detection, the Neyman–Pearson criterion helps in setting thresholds for detecting targets by maximising the probability of detection (P_D) while keeping the probability of false alarm (P_{FA}) within acceptable limits. This criterion is used to calculate the detection probability of the UAV when it is within the range of a radar system, forming the basis for the cost function in the A^* algorithm and the reward function in the proposed PPO model.

The radar detection probability P_D is determined by the following components,

1. Signal-to-Noise Ratio (SNR): The SNR is a measure of the signal strength relative to the background noise. In the radar context, it helps to determine how easily a target can be detected by the radar. A higher SNR means better detectability of the target. SNR is calculated based on the distance between the UAV and the radar. The SNR decreases with the fourth power of the distance, i.e.,

$$\text{SNR} = \frac{\text{Radar Power}}{\text{Distance}^4}. \tag{2}$$

2. Eigenvalues of the Correlation matrix: The correlation matrix represents how similar or correlated the signals received by the radar are across different pulses or measurements. Suppose a radar transmits a series of pulses, and for each pulse, it receives a signal that may or may not contain the reflected signal from a target like a UAV. If the radar transmits N pulses, the signals it receives can be represented as a vector, where each entry corresponds to the received signal for one pulse. For a radar with N pulses and correlation ρ, the eigenvalues λ_i of the correlation matrix are

$$\lambda_i = \rho^i, \;\; i = 0, 1, \ldots, N-1. \tag{3}$$

3. Detection threshold: The detection threshold V_T is obtained by specifying the false alarm probability P_{FA} as [19]

$$P_{FA} = \exp(-V_T) \sum_{n=0}^{N-1} \frac{{V_T}^n}{n!}. \tag{4}$$

4. Detection probability: the detection probability is calculated by adding the contributions from each pulse while accounting for the SNR and the detection threshold, i.e.,

$$P_D = 1 - \sum_{i=1}^{N} \prod_{\substack{j=1 \\ j \neq i}}^{N} \left(1 - \frac{1 + \text{SNR} \cdot \lambda_j}{1 + \text{SNR} \cdot \lambda_i}\right)^{-1} \exp(-V_T/(1 + \text{SNR} \cdot \lambda_i)). \tag{5}$$

4. Methodology

The proposed approach is depicted in Figure 2. This approach consists of the design of an autonomous path-planning algorithm for UAVs capable of reducing the detection of radars and ensuring that the task is completed. The proposed approach consists of a modified proximal policy optimisation (PPO) algorithm based on an action-shaping mechanism that allows the PPO algorithm to focus only on the detection zones in order to avoid the detection of radars in real-time.

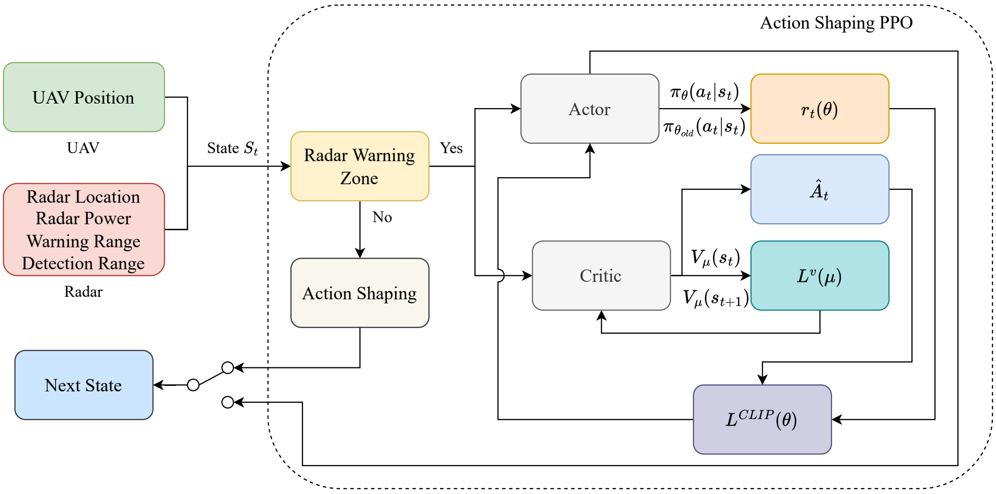


Figure 2. High-level diagram of the proposed Action-shaping PPO algorithm for detection avoidance of radars.

The main elements of the proposed architecture are as follows:

1. UAV and Radar State Information: these blocks provide the necessary information regarding both the locations of the UAV and radar, as well as the radar power, warning and detection ranges. This information plays a pivotal role in the decision-making process for action selection and manoeuvrability.
2. Radar Warning: in this block, the state is analysed to verify if the UAV is within a radar's warning zone. If the UAV is within a warning zone, then the PPO algorithm is applied for detection avoidance. Otherwise, it applies an action-shaping mechanism to select actions that drive the UAV towards the target location.
3. Action-shaping mechanism: this mechanism is used only when the UAV is not within a radar's warning range. Here, the action-shaping consists of reducing the action-space of the UAV to move only in the direction of the target goal and avoid using actions that may produce unnecessary movements.
4. Actor-Critic PPO Module: this module consists of two main elements,
    - Actor Network: which is responsible for the action selection based on a parametrised policy $\pi_\theta(a_t \mid s_t)$ with parameters θ.
    - Critic Network: which evaluates the value function $V_\mu(x_t)$ with parameter μ that enables to improvement of the actor policy.

    Both the actor and critic networks are updated using mini-batches of experiences collected from the environment–UAV interaction.
5. Reward function and policy update: the PPO algorithm uses a clipped objective function to prevent large and unstable updates. In the action-shaping implementation, the restricted action space complementary works with the clipped updates to stabilise the algorithm training and accelerate its convergence. The PPO updates the policy using a clipped objective $L^{CLIP}(\theta)$ based on the advantage estimate $\hat{A}_t$.
    The reward is designed to penalise actions leading to radar detection. This encourages the UAV to avoid those penalising paths, whilst moving towards to the location goal.

6. Action and policy iteration: based on the reward and value function updates, the actor selects the next action a_t that the UAV will apply to reach a new state s_{t+1} that is not within the detection range of the radars. This process is repeated until the goal is reached or if a terminal condition is met (e.g., the maximum number of steps is reached).

4.1. Proximal Policy Optimisation (PPO)

PPO is a well-known reinforcement learning (RL) architecture that has gained popularity for its effectiveness and simplicity in solving complex decision-making problems [44]. PPO is part of the policy gradient methods in RL, where the objective is to directly optimize the policy (the decision-making strategy) that maps states to actions to maximise cumulative rewards.

PPO operates by iteratively improving the policy while ensuring that updates to the policy are not too large. It uses a clipped objective function to prevent large policy updates, which could destabilise the learning process. This approach allows PPO to strike a balance between exploration and exploitation, making it suitable for dynamic environments where conditions change over time.

PPO uses an actor-critic framework, where the actor network represents the policy π_θ with parameter θ, and a critic network $V_\mu(s_t)$ with parameter μ. PPO introduces a surrogate objective function that includes a clipping mechanism to prevent large updates to the policy. The probability ratio $r_t(\theta)$ is defined as the ratio of the probability of taking an action under the new policy to the probability of taking the same action under the old policy

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}. \tag{6}$$

The clipped objective function is formulated as,

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)], \tag{7}$$

where $\hat{A}_t$ is an estimate of the advantage function, ϵ is a hyperparameter that determines the clipping range and the function, $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)$ limits $r_t(\theta)$ to a small range around 1, preventing the policy from changing too drastically.

The advantage function $\hat{A}(t)$ measures how good or worse an action is compared to the expected outcome. It helps to reduce the variance in the gradient estimates. The clipping mechanism in PPO acts as a form of regularisation to ensure that the policy does not change too much with each update. This makes the learning process more stable and prevents performance degradation.

In this paper, the state space is continuous within a user-defined bounded box. The action space consists of eight actions given by

$$\mathcal{A} = \{\rightarrow, \uparrow, \downarrow, \nearrow, \searrow, \leftarrow, \nwarrow, \swarrow\}. \tag{8}$$

where each action symbol is described in Table 1.

Table 1. Actions meaning.

Symbol	Meaning
$\rightarrow$	Forward
$\uparrow$	Upward
$\downarrow$	Downward
$\nearrow$	Forward-upward
$\searrow$	Forward-downward
$\leftarrow$	Backward
$\nwarrow$	Backward-upward
$\swarrow$	Backward-downward

The actions are transformed into their respective physical angles as

$$\mathcal{A} = \{\rightarrow, \uparrow, \downarrow, \nearrow, \searrow, \leftarrow, \nwarrow, \swarrow\} \mapsto q = \left\{0, \frac{\pi}{2}, -\frac{\pi}{2}, \frac{\pi}{4}, -\frac{\pi}{4}, -\pi, \frac{3\pi}{4}, -\frac{3\pi}{4}\right\}. \quad (9)$$

4.2. Action Shaping PPO

The objective of this paper is to train a path planning algorithm that drives the UAV to navigate from a starting point to a goal, whilst minimising the risk of detection by radar systems. This is achieved by taking strategic actions that avoid radar detection zones, whilst moving efficiently towards the goal. Here, the use of the standard PPO algorithm can lead to suboptimal learning and slow convergence, because the algorithm learns the complete path from scratch.

Warning radar zone switching criteria are developed based on the Neyman–Pearson criteria to determine the decision-making strategy that the UAV needs to follow when it is within a warning or non-warning radar detection zone areas.

The proposed action shaping PPO algorithm aims to improve the learning efficiency of standard PPO algorithm, whilst generating a higher-quality path that minimises radar detection. The key idea of the algorithm is to strategically limit the UAV's movements into directions that optimize its efficiency in reaching the target goal. Here, the PPO algorithm is applied only when the UAV enters the a warning area before reaching the detection zone, whilst the action shaping is used in non-detection zones and drives the UAV to reach the target goal. Algorithm 1 provides the pseudo-code of the action-shaping mechanism used in the proposed action-shaping PPO.

Algorithm 1 Action shaping mechanism

```
Given the UAV position and goal position
if $y_{UAV} = y_{Goal}$ then
    if $x_{UAV} > x_{Goal}$ then
        Allowed actions $\{\leftarrow\}$ (backward).
    else if $x_{UAV} < x_{Goal}$ then
        Allowed actions $\{\rightarrow\}$ (forward).
    end if
else if $y_{UAV} < y_{Goal}$ then
    if $x_{UAV} = x_{Goal}$ then
        Allowed actions $\{\uparrow\}$ (upward).
    else if $x_{UAV} > x_{Goal}$ then
        Allowed actions $\{\leftarrow, \nwarrow, \uparrow\}$ (backward, backward-upward and upward).
    else if $x_{UAV} < x_{Goal}$ then
        Allowed actions $\{\uparrow, \nearrow, \rightarrow\}$ (upward, forward-upward and forward).
    end if
else if $y_{UAV} > y_{Goal}$ then
    if $x_{UAV} = x_{Goal}$ then
        Allowed actions $\{\downarrow\}$ (downward).
    else if $x_{UAV} > x_{Goal}$ then
        Allowed actions $\{\leftarrow, \swarrow, \downarrow\}$ (backward, backward-downward and downward).
    else if $x_{UAV} < x_{Goal}$ then
        Allowed actions $\{\rightarrow, \searrow, \downarrow\}$ (forward, forward-downward and downward).
    end if
end if
```

The key advantages of the proposed action shaping PPO are fivefold,

- Restricted Movement Integration: The action space is reduced based on the UAV's position relative to the target goal and the radar warning zone. The UAV's movement options are dynamically adjusted during training. Depending on its environment, the action space can be narrowed or expanded, ensuring the UAV avoids unnecessary movements that would lead to the exploration of suboptimal paths.

- Actor-Critic Structure: The use of separate actor and critic networks allows for more stable learning in environments with complex dynamics.
- Adaptive Action Selection: The model adapts the action space depending on whether the UAV is in a radar warning area, improving efficiency and safety.
- Faster Convergence: By limiting the action space to strategic movements, the model converges faster, requiring fewer training steps.
- Enhanced Safety: The restricted movement prevents the UAV from making large, unpredictable moves, ensuring it stays within safe zones during both training and deployment.

The reward function r is designed as a sparse reward with the following components

- If the UAV is in a non-detection zone, that is, the action-shaping mechanism is applied then the reward is $r = -\|\boldsymbol{x}_{UAV} - \boldsymbol{x}_{Goal}\|$, where $\boldsymbol{x}_{UAV} = [x_{UAV}, y_{UAV}]^\top$ is the position of the UAV and $\boldsymbol{x}_{Goal} = [x_{Goal}, y_{Goal}]^\top$ is the position of the target goal.
- If the UAV moves to a previous visited state, then it is penalised with a reward of $r = -1$.
- If the UAV moves to a position that reduces the distance to the target, then a positive reward of $r = 20$ is given.
- If the UAV enters into the radar detection range and exceeds a threshold of 0.2, then it is penalised by a reward of $r = -1000P_D$.
- If the UAV is within the radar range, then it is penalised by a reward of $r = -20$.
- If the UAV reaches the target then a positive reward of $r = 1000$ is given; otherwise, it is penalised with a reward of $r = -0.1$.

4.3. Modified Sparse A^* Algorithm

For fair comparisons of the proposed action-shaping PPO, we design a modified version of the sparse A^* algorithm [45] that is capable of optimising the trajectory path in environments with multiple radars, enabling the aircraft to avoid regions of high radar detection probability, whilst maintaining efficient flight paths. This is achieved by designing a relative complex cost function based on the Neyman–Pearson criteria. This makes the algorithm more suitable for scenarios like radar detection avoidance, where stealth and safety are as important as reaching the goal.

The following key features are considered in the optimisation cost,

- Total distance: this term accounts for the cumulative distance travelled. This term accounts for the cumulative distance travelled along the current path. The algorithm aims to minimise the overall distance, which aligns with operational efficiency and fuel conservation.
- Distance to Goal: This heuristic guides the path toward the goal. It is the Euclidean distance between the current node and the goal. This term encourages the UAV to move closer to the target in a straight line when possible.
- Cumulative Radar Detection Probability: It considers the cumulative radar detection probability encountered along the path. Lower detection probabilities are preferred, so paths that keep this metric low are favoured. The radar detection probability is computed using a Neyman–Pearson criterion-based model that accounts for factors like signal-to-noise ratio (SNR), and the characteristics of the radar (e.g., pulse number, correlation).
- Height Difference: This term measures the altitude change between consecutive nodes. Large altitude changes are undesirable due to aircraft performance limits, energy consumption, or increased radar visibility. For 2D environments, it can be set to zero.
- Immediate Radar Detection Probability: This term considers the radar detection probability at the current node. It ensures that paths moving into high-risk areas (high detection probability) are heavily penalised.

The node with the lowest overall cost is selected for further expansion. This process repeats iteratively until the goal is reached. The radar model also considers the detection range, warning range, and radar field of view. Nodes outside the radar's detection range are considered safe, while those within the range are evaluated based on the radar model.

Therefore, the proposed optimisation cost is designed as a linear combination of the aforementioned features as follows

$$\begin{aligned} J = \;& k_1 \cdot \text{Total distance} + k_2 \cdot \text{Distance to Goal} + k_3 \cdot \text{Cumulative detection probability} \\ & k_4 \cdot \text{Height difference} + k_5 \cdot \text{Immediate radar detection}, \end{aligned} \tag{10}$$

where $k_i, i = 1, \ldots, 5$ are user-defined scalars that weight each feature. By carefully balancing the cost function's components, the algorithm can generate paths that are both stealthy and efficient.

Remark 1. *Both the cost of the sparse A^* and the reward function of the PPO algorithm aim to minimise the travel distance with minimum detection probability. Despite sharing the same task, they are designed differently due to the design nature of A^* and RL algorithms.*

5. Results

We test the proposed approach in different environments under different target goal locations and radars with different capabilities. These configurations were consistent across all models to ensure fair comparisons.

We consider a square area of 400×400 km^2. This area is divided into a 1200×1200 grid to train the proposed approach. Here, the positions of the kinematic model are discretised to match the closest position location of the 1200×1200 grid. This allows to model a time-varying velocity instead of the constant velocity of the kinematic model. It is assumed that the UAV is equipped with a radar detection sensor, capable of identifying radar systems and issuing warnings prior to the UAV entering their detection range. Multiple radars were placed within the environment, each with defined detection and warning ranges, power, number of pulses, correlation, and false alarm rate. These parameters are used to calculate the radar detection probability using the Neyman–Pearson criterion. The radar detection probability for each position of the UAV is computed based on its distance from the radar and the radar's characteristics. Table 2 summarizes the characteristics of the radars used in the proposed implementation.

Table 2. Radar parameters.

No	Radar Detection	Warning Range	Radar Power	Pulse	Correlation	False Alarm
1	30	40	5×10^{-5}	5	0.5	1×10^{-6}
2	40	55	1×10^{-4}	10	0.5	1×10^{-6}
3	50	70	3.9063×10^{-3}	15	0.5	1×10^{-6}

The neural network of the PPO algorithm consists of a multi-layer perceptron (MLP) with two hidden layers. Each hidden layer consists of 64 neurons and a ReLU activation function. The output layer is determined by the dimension of the action space of the environment. The hyperparameters of the simulation are set to a learning rate of 0.0001 and a clip range of 0.2. The number of steps that the UAV requires to reach the target goal is denoted as an episode. In each episode we modify the number of radars and their locations to ensure good generalisation of the PPO algorithm. We consider 100,000 time steps to train the PPO algorithm. For the modified sparse A^* algorithm, we use the weights of Table 3 for the cost design.

Notice that the scalar weights k_i depend on the radar detection range. These values are proposed to ensure the UAV behaves similarly for all radars.

Table 3. Proposed scalar weights of the modified sparse A^* cost.

No	Radar Detection	Weights $(k_1, k_2, k_3, k_4, k_5)$
1	30	$(0, 1 \times 10^{-4}, 0.8, 0, 0)$
2	40	$(1 \times 10^{-6}, 1 \times 10^{-4}, 8, 0, 1 \times 10^{2})$
3	50	$(1 \times 10^{-9}, 1 \times 10^{-4}, 8, 0, 9.5)$

5.1. Comparison against Traditional PPO

We first motivate the proposed approach by first training a simple PPO algorithm without an action-shaping mechanism. The results of the PPO algorithm in episode 5 are observed in Figure 3. Here, the strong orange circles denote the radar detection range, whilst the soft orange contour models the radar warning zones.

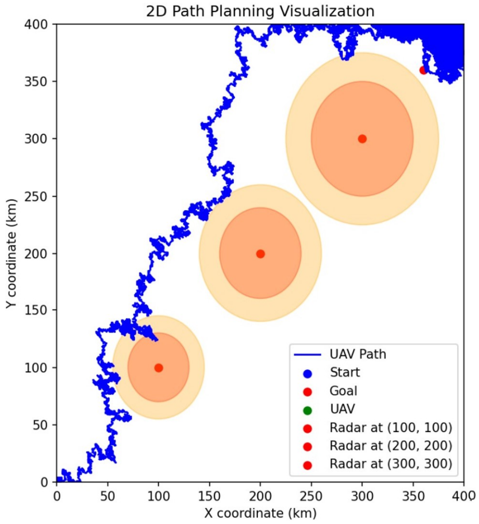


Figure 3. Path planning of UAV using PPO algorithm: Trajectory obtained in the episode 5.

The result of Figure 3 shows that the PPO algorithm takes unnecessary actions in zones without radar detectability. This translates into more steps to reach the target goal, specifically, it takes 74,879 time steps to reach the target with a cumulative detection probability of 49.99 and a total travelled distance of 86,048.72 km. Figure 4 shows the results using the proposed action-shaping PPO algorithm in episode 5.

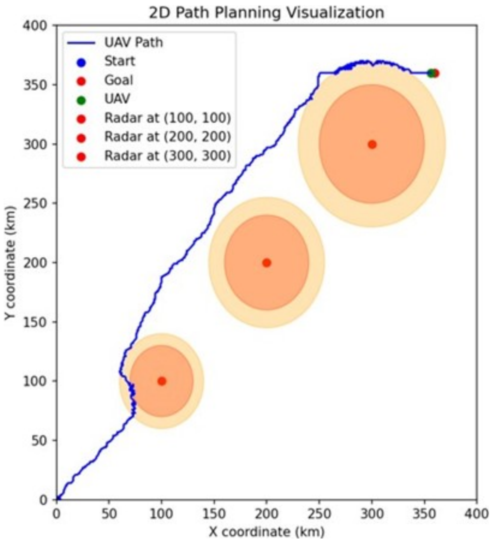


Figure 4. Path planning of UAV using action-shaping PPO: Trajectory obtained in the episode 5.

Notice that incorporating the action-shaping mechanism allows having a smooth trajectory that accelerates convergence to the goal in fewer steps. Here, the action-shaping PPO takes 887 time steps to reach the goal with a cumulative detection probability of 15.99 and a total travel distance of 1504.12 km. This clearly shows the benefits of adding this simple mechanism into the PPO algorithm to obtain better and more reliable results. Table 4 summarises the comparison results between PPO and action-shaping PPO.

Table 4. Comparison of UAV path planning outcomes using standard PPO and action-shaping PPO under the Scenario 1.

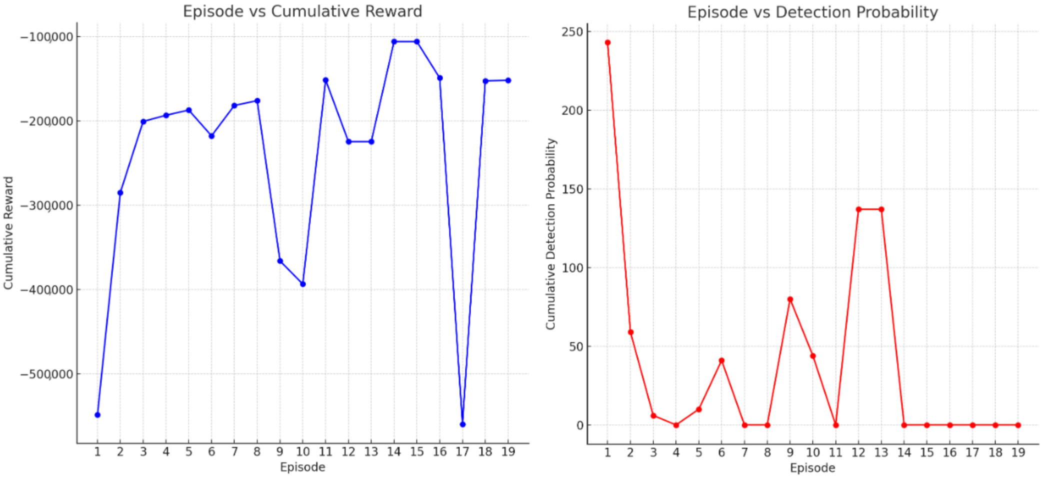
Methods	Time Steps	Cum. Probability	Dist. Travelled
PPO	74,879	49.99	86,048.72 km
Action-shaping PPO	887	15.99	1504.12 km

Remark 2. *The distance travelled by each agent is variable in each time step due to the discretisation of the position of the drone in the proposed large state space grid.*

The comparisons allow us to conclude the following key benefits

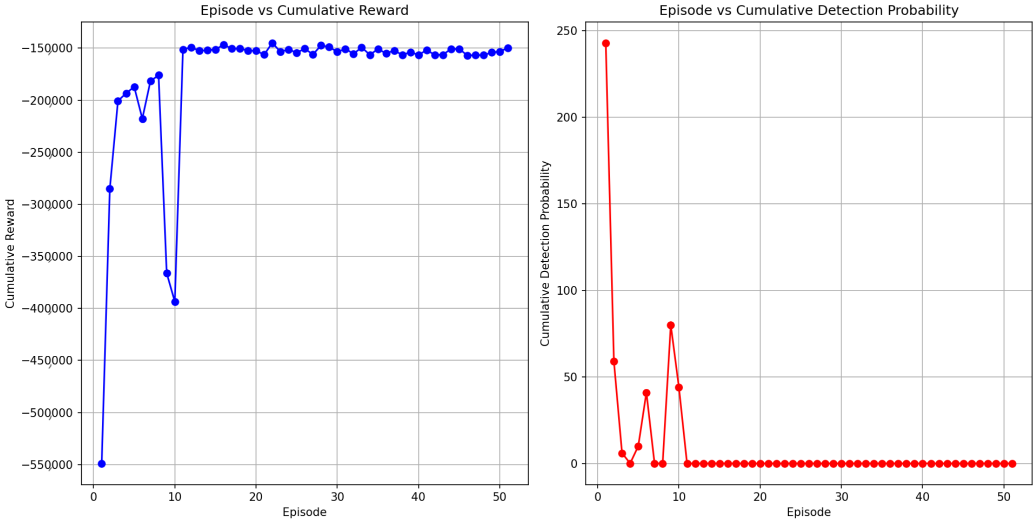
- Efficiency: the UAV using the action-shaping PPO reaches the goal much faster compared with the UAV using the standard PPO. This means that the action-shaping mechanism provides an effective tool to guide the UAV towards the goal without applying unnecessary actions.
- Detection probability: the cumulative detection probability is notably reduced which is critical for this particular implementation to ensure the survivability of the UAV.
- Path smoothness: as previously discussed, the path of the standard PPO is not smooth due to the random selection of actions in zones without radar detectability. In contrast, the proposed approach overcomes this issue such that the PPO is only applied in the coverage area of the radar.
- Distance travelled: this is a key benefit of the proposed approach since the UAV is capable of reaching the goal in less number of steps which directly affects the travel distance that is a critical element when the UAV resources are limited, e.g., battery time or fuel.

In view of this, we can conclude that action-shaping brings value to the standard PPO in this specific application. Figure 5 shows the cumulative reward and detection probability in each episode of the action-shaping PPO. Here, the cumulative reward fluctuates due to the different travel distances obtained from the change in the number of radars and their locations. However, we can observe that the detection probability is minimised across diverse environment configurations. If the number of radars and their locations are maintained fixed we obtain the results of Figure 6. In this case, the cumulative reward and detection provability converge in approximately 11 episodes.



(**a**) Cumulative reward (**b**) Total detection probability

Figure 5. Training results of the action-shaping PPO under random environment configurations.



(**a**) Cumulative reward

(**b**) Total detection probability

Figure 6. Training results of the action-shaping PPO under a fixed environment configuration.

We further compare the proposed approach using the modified sparse A^* algorithm designed in this paper across different implementations. Here, we used the policy generated by the action-shaping PPO trained under individual environments after 50 episodes.

5.2. Scenario 1

Consider first the results of the modified sparse A^* algorithm in a scenario of three consecutive radars located in front of the UAV and the target goal. The result is shown in Figure 7.

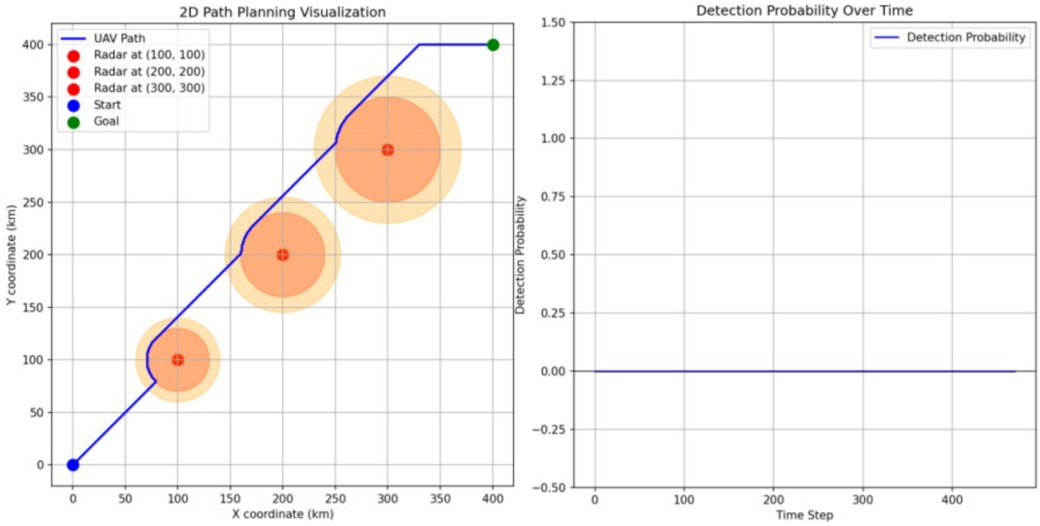


Figure 7. Scenario 1: Penetration Path of the modified sparse A^* algorithm.

The results demonstrate that the modified sparse A^* is effective in avoiding the detection of the radars across all the paths. Here, the UAV takes 4700 time-steps to reach the goal with a cumulative detection probability of 0 and a total travelled distance of 613.32 km. Figure 8 shows the results of the action-shaping PPO.

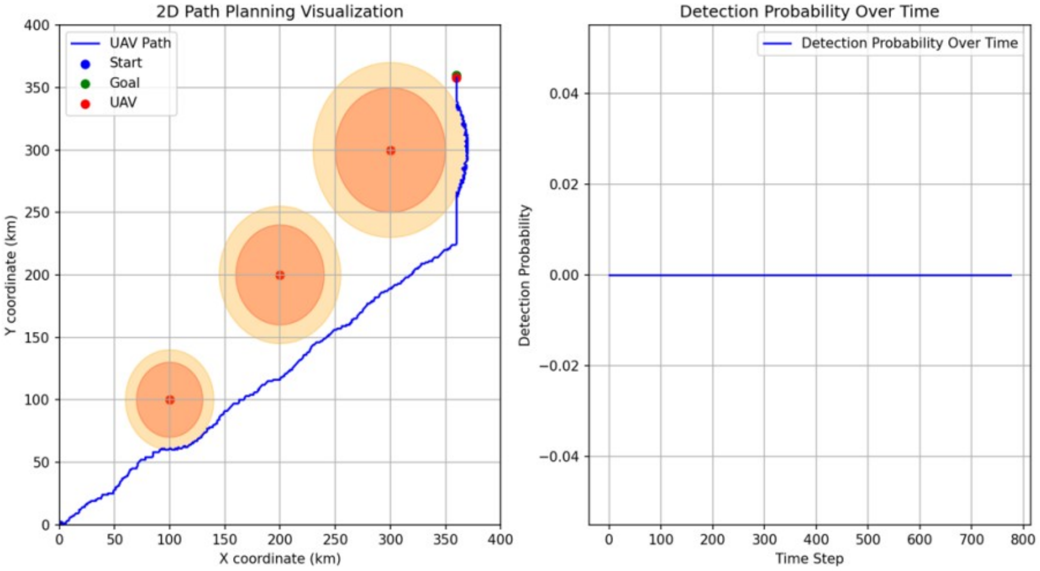


Figure 8. Scenario 1: Penetration Path of the action-shaping PPO algorithm.

Similar to the results observed in Figure 4, the action-shaping PPO is effective in reaching the goal, whilst avoiding the radar detection. In this case, the action-shaping PPO takes more steps (and therefore, travel distance) compared with the modified sparse A^* algorithm. This is evident since the modified sparse A^* assumes knowledge of the map which facilitates the optimisation problem. Here, the action-shaping PPO takes 778 time steps to reach the goal, with 0 cumulative detection probability and a travel distance of 1394.76 km. Notice that the behaviour of the action-shaping PPO can be further improved by considering other action-shaping mechanisms to consider the radar detection range as an indicator for action selection based on the location of the UAV. The obtained results are summarised in Table 5 for visualisation purposes.

Table 5. Comparison of UAV path planning outcomes using modified A^* and action-shaping PPO under Scenario 1.

Methods	Time Steps	Cum. Probability	Dist. Travelled
Modified sparse A^*	470	0.0	613.32 km
Action-shaping PPO	778	0.0	1394.76 km

5.3. Scenario 2

Consider a more complex scenario which consists of five radars located in different positions on the map with different detection ranges. The result of the modified sparse A^* algorithm is shown in Figure 9.

A similar performance is observed in this specific scenario, where the path followed by the modified sparse A^* algorithm stays in the borderline of the radar detection range. This performance can be risky since a wrong manoeuvre can lead to the UAV entering the radar range and increasing the likelihood of detection. Nevertheless, the path followed by the UAV effectively reach the target goal with 0 detection probability. Here, the modified sparse A^* algorithm reaches to the goal in 636 time steps which is equivalent to 767.72 km of travel distance.

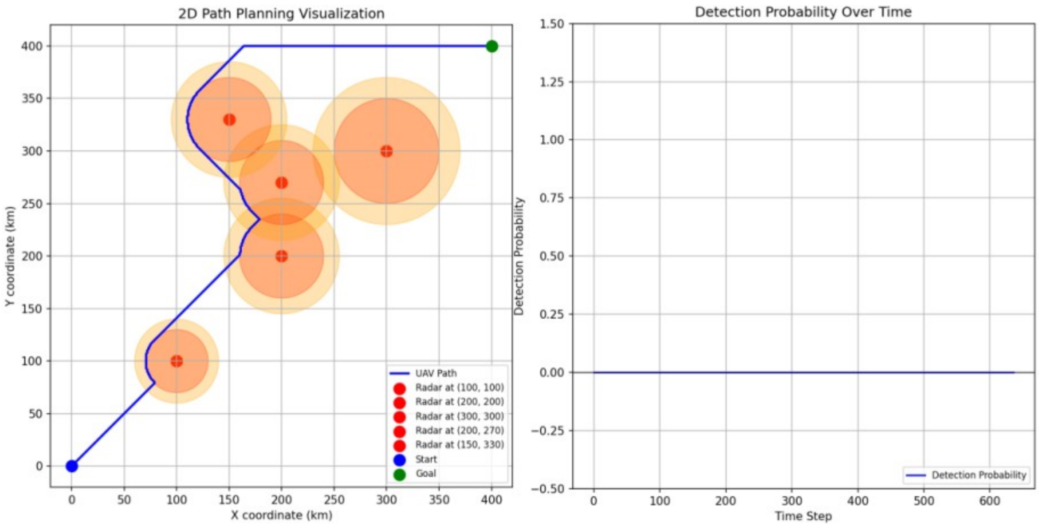


Figure 9. Scenario 2: Penetration Path of the modified sparse A^* algorithm.

Figure 10 shows the result using the action-shaping PPO algorithm. In this case, the path followed tries to move in the zones where the radar detection is null such that the detection probability is zero. This implies moving at a relatively larger distance, i.e., the algorithm takes 681 time steps equivalent to 1323.03 km of travel time. Table 6 summarises the results obtained from this particular scenario.

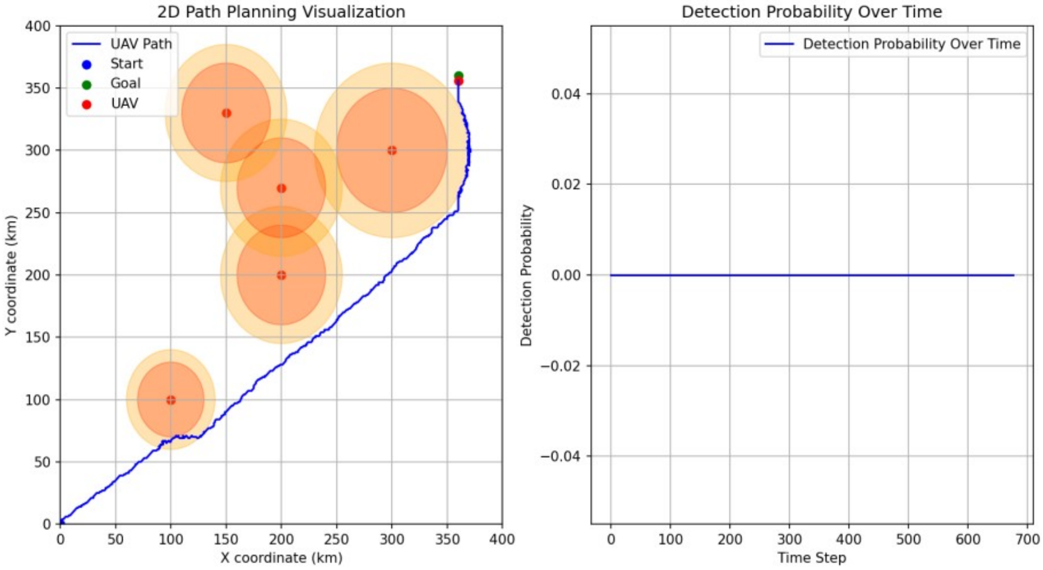


Figure 10. Scenario 2: Penetration Path of the action-shaping PPO algorithm.

Table 6. Comparison of UAV path planning outcomes using modified sparse A^* and action-shaping PPO under the Scenario 2.

Methods	Time Steps	Cum. Probability	Dist. Travelled
Modified sparse A^*	636	0.0	767.72 km
Action-shaping PPO	681	0.0	1323.03 km

5.4. Scenario 3

Consider other complex scenarios with five radars with only three potential paths to reach the target goal. Figure 11 shows the results of the modified sparse A^* under this new scenario.

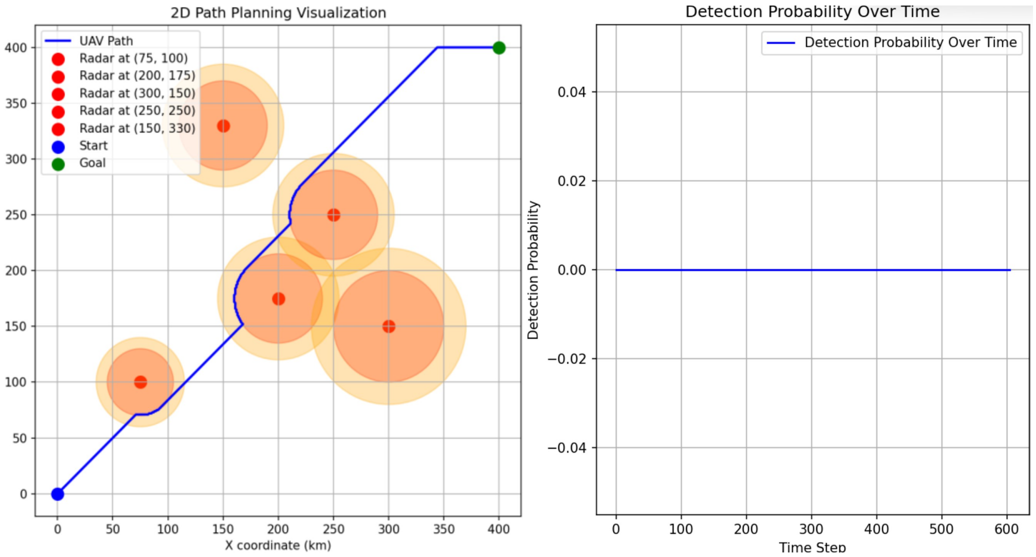


Figure 11. Scenario 3: Penetration Path of the modified sparse A^*.

Similar results to Scenario 2 are observed, that is, the modified sparse A^* algorithm follows a path in the borderline of the radars detection range. As we previously discussed, this performance is risky such that the UAV can be detected under windy conditions or wrong manoeuvres. Here, the modified sparse A^* algorithm reaches the target goal in just 472 steps equivalent to a travelled distance of 615.31 km with zero detection probability. Figure 12 shows the results using the action-shaping PPO, where the UAV reaches the goal in 609 time steps with zero detection probability and a travelled distance of 1197.18 km.

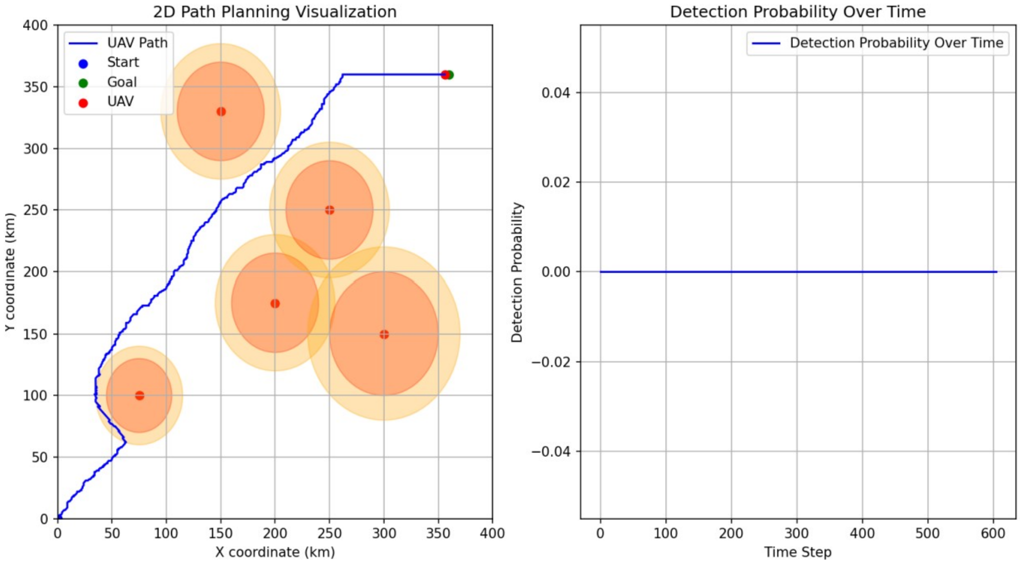


Figure 12. Scenario 3: Penetration Path of the action-shaping PPO.

Table 7 summarizes the obtained results using both the modified sparse A^* and action-shaping PPO algorithms.

Table 7. Comparison of UAV path planning outcomes using modified sparse A^* and action-shaping PPO under the Scenario 3.

Methods	Time Steps	Cum. Probability	Dist. Travelled
Modified sparse A^*	472	0.0	615.31 km
Action-shaping PPO	609	0.0	1197.18 km

5.5. Improvement of the Action-Shaping PPO

To enhance the performance of the action-shaping PPO, we modify the action-shaping mechanism and use the Euclidean distance to select the state that is closer to the target goal. Figure 13 shows the results obtained by this slight improvement.

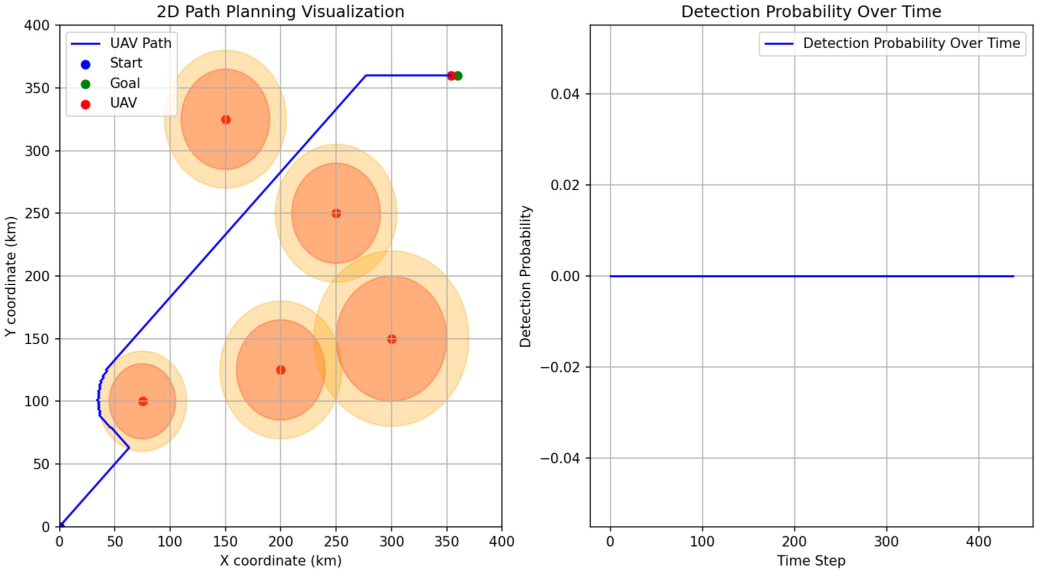


Figure 13. Scenario 3: Penetration Path of the improved action-shaping PPO.

The results show that the incorporation of the Euclidean distance in the action-shaping mechanism smooths the trajectory path in the zones without radar detection. Here, the algorithm reaches the target goal in just 444 time steps and reduces the travel distance of the action-shaping PPO to 1100.53 km. Table 8 compares the results of the modified sparse A^*, the action-shaping PPO, and the improved action-shaping PPO algorithms.

Table 8. Comparison of UAV path planning outcomes using modified sparse A^*, action-shaping PPO and improved action-shaping PPO under the Scenario 3.

Methods	Time Steps	Cum. Probability	Dist. Travelled
Modified sparse A^*	472	0.0	615.31 km
Action-shaping PPO	609	0.0	1197.18 km
Improved Action-shaping PPO	444	0.0	1100.53 km

5.6. Limitations and Future Work

New developments in the field of path planning for UAVs have started to include measurements of the Radar Cross Section (RCS) to improve the effectiveness of the path planning algorithm. This is intended to minimise the radar visibility of UAVs, especially in stealth operations. However, the proposed approach did not incorporate RCS as an additional feature for the design of the path planning models. The RCS of a stealth aircraft is a critical metric in assessing the radar visibility of an object, which involves understanding the following key factors,

- Geometric Data of the Aircraft: this includes the dimensions of the aircraft such as its length, wingspan, height, and overall shape. In addition, the surface materials for the absorption of radar waves, and the panel configuration are critical aspects that affect the RCS.
- Data on Stealth Features: this covers the design features (e.g., edge alignment, RAM coating, cooling techniques) and operational profiles (e.g., typical flights and angles that the aircraft operates) that play important roles in the performance of the RCS.
- Radar Characteristics:
  - Radar frequencies that interact differently with the aircraft's surface. Higher frequencies (shorter wavelengths) are more sensitive to smaller details on the aircraft's surface, while lower frequencies (longer wavelengths) interact more with the overall shape.
  - Polarisation of the radar signal (vertical, horizontal, or circular) can affect how the radar waves interact with the aircraft. The RCS can vary depending on the polarisation of the incoming radar signal.
  - Incident Angle where the radar waves hit the aircraft is critical. RCS is highly dependent on the aspect angle, and the orientation of the aircraft relative to the radar source.
- Environmental Conditions: include both atmospheric conditions (e.g., humidity, temperature and pressure) and background noise which can reduce the radar signal strength and/or the radar readings.
- Radar Cross-section Data: such as monostatic and bistatic RCS. Here, the monostatic RCS is a measurement taken when the radar transmitter and receiver are at the same location. It is the most common method and provides a direct measure of how much energy is reflected back to the radar source. On the other hand, the bistatic RCS is the measurement taken when the radar transmitter and receiver are at different locations. These data help in understanding how radar waves are scattered in different directions, not just back towards the radar source.
- Computational Simulations:
  - Electromagnetic Modelling to predict the RCS, electromagnetic modelling techniques like the Method of Moments (MoM), Finite Element Method (FEM), or Finite Difference Time Domain (FDTD) are used. These simulations require detailed geometric and material data of the aircraft.
  - Simulation Parameters such as frequency range, incident angles, and polarisation settings, which need to align with the actual measurement conditions.
- Historical RCS Data: which consists of data from previous tests or from similar aircraft models used for comparison. This helps to understand the effectiveness of the stealth features and identify areas for improvement.

Incorporating RCS data into the proposed path planning algorithms is a research direction that we are keen to address as future work. In addition, the development of hybrid approaches that combine the strengths of heuristic-based methods like Sparse A^* with RL techniques is becoming an attractive field for future research.

6. Conclusions

This paper investigates the development of path-planning systems for military UAVs, emphasising radar detection avoidance to ensure safe navigation in contested environments. An action-shaping PPO is proposed to address this problem. The approach uses the Neyman–Pearson criterion to measure the probability of radar detection, ensuring consistent evaluation of UAV exposure to radar. The algorithm incorporates an action-shaping mechanism to impose additional constraints on UAV movements and refine the path-planning process. This method is designed to enhance the training speed and efficiency of traditional PPO algorithms by limiting the UAV's action space to only the most strategic movements. This action-shaping mechanism not only accelerates the learning process, but also enhances the UAV's ability to avoid radar detection more efficiently. For comparison purposes, we design a modified version of the sparse A^* algorithm that evaluates the cost of reaching the goal from different nodes while considering radar detection probabilities. The results demonstrate that the proposed models inject into the UAV the ability to learn effective strategies for avoiding radar while navigating towards its target.

Author Contributions: Conceptualisation, A.M.A. and A.P.; methodology, A.M.A.; software, A.M.A.; validation, A.M.A., A.P. and W.G.; formal analysis, W.G.; investigation, A.P.; resources, A.T.; writing original draft preparation, A.P. and A.T.; visualisation, A.M.A.; supervision, A.P. and W.G.; project administration, A.T.; funding acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article material, further inquiries can be directed to the corresponding author.

DURC Statement: Current research is limited to the research in autonomous path planning algorithms using drones, which is beneficial for the development of new smart living applications and does not pose a threat to public health or national security. Authors acknowledge the dual-use potential of the research involving drones and confirm that all necessary precautions have been taken to prevent potential misuse. As an ethical responsibility, authors strictly adhere to relevant national and international laws about DURC. Authors advocate for responsible deployment, ethical considerations, regulatory compliance, and transparent reporting to mitigate misuse risks and foster beneficial outcomes.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DC	Discretize continuous actions
DQN	Deep Q-Networks
FDTD	Finite Difference Time Domain
FEM	Finite Element Method
LSTM	Long Short-Term Memory
MoM	Method of Moments
PD	Probability Distribution
PPO	Proximal Policy Optimisation
RCS	Radar Cross section
RA	Removing actions
RL	Reinforcement Learning
RPP	Real Path Planning
UAV	Unmanned Air Vehicle

References

1. Wang, N. "A Success Story that Can Be Sold"?: A Case Study of Humanitarian Use of Drones. In Proceedings of the 2019 IEEE International Symposium on Technology and Society (ISTAS), Medford, MA, USA, 15–16 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
2. Cui, Y.; Osaki, S.; Matsubara, T. Autonomous boat driving system using sample-efficient model predictive control-based reinforcement learning approach. *J. Field Robot.* **2021**, *38*, 331–354. [CrossRef]
3. Amendola, J.; Miura, L.S.; Costa, A.H.R.; Cozman, F.G.; Tannuri, E.A. Navigation in restricted channels under environmental conditions: Fast-time simulation by asynchronous deep reinforcement learning. *IEEE Access* **2020**, *8*, 149199–149213. [CrossRef]
4. Thombre, S.; Zhao, Z.; Ramm-Schmidt, H.; García, J.M.V.; Malkamäki, T.; Nikolskiy, S.; Hammarberg, T.; Nuortie, H.; Bhuiyan, M.Z.H.; Särkkä, S.; et al. Sensors and AI techniques for situational awareness in autonomous ships: A review. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 64–83. [CrossRef]

5. Fraser, B.; Perrusquía, A.; Panagiotakopoulos, D.; Guo, W. A Deep Mixture of Experts Network for Drone Trajectory Intent Classification and Prediction Using Non-Cooperative Radar Data. In Proceedings of the 2023 IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 5–8 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
6. Gasparetto, A.; Boscariol, P.; Lanzutti, A.; Vidoni, R. Path planning and trajectory planning algorithms: A general overview. In *Motion and Operation Planning of Robotic Systems: Background and Practical Approaches*; Springer International Publishing: Cham, Switzerland, 2015; pp. 3–27.
7. Gruffeille, C.; Perrusquía, A.; Tsourdos, A.; Guo, W. Disaster Area Coverage Optimisation Using Reinforcement Learning. In Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS), Chania, Crete, Greece, 4–7 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 61–67.
8. Vagale, A.; Bye, R.T.; Oucheikh, R.; Osen, O.L.; Fossen, T.I. Path planning and collision avoidance for autonomous surface vehicles II: A comparative study of algorithms. *J. Mar. Sci. Technol.* **2021**, *26*, 1307–1323. [CrossRef]
9. Bildik, E.; Tsourdos, A.; Perrusquía, A.; Inalhan, G. Swarm decoys deployment for missile deceive using multi-agent reinforcement learning. In Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS), Chania, Crete, Greece, 4–7 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 256–263.
10. Li, G.; Hildre, H.P.; Zhang, H. Toward time-optimal trajectory planning for autonomous ship maneuvering in close-range encounters. *IEEE J. Ocean. Eng.* **2019**, *45*, 1219–1234. [CrossRef]
11. Shaobo, W.; Yingjun, Z.; Lianbo, L. A collision avoidance decision-making system for autonomous ship based on modified velocity obstacle method. *Ocean Eng.* **2020**, *215*, 107910. [CrossRef]
12. El Debeiki, M.; Al-Rubaye, S.; Perrusquía, A.; Conrad, C.; Flores-Campos, J.A. An Advanced Path Planning and UAV Relay System: Enhancing Connectivity in Rural Environments. *Future Internet* **2024**, *16*, 89. [CrossRef]
13. Lyu, D.; Chen, Z.; Cai, Z.; Piao, S. Robot path planning by leveraging the graph-encoded Floyd algorithm. *Future Gener. Comput. Syst.* **2021**, *122*, 204–208. [CrossRef]
14. Hameed, R.; Maqsood, A.; Hashmi, A.; Saeed, M.; Riaz, R. Reinforcement learning-based radar-evasive path planning: A comparative analysis. *Aeronaut. J.* **2022**, *126*, 547–564. [CrossRef]
15. Tang, G.; Tang, C.; Claramunt, C.; Hu, X.; Zhou, P. Geometric A-star algorithm: An improved A-star algorithm for AGV path planning in a port environment. *IEEE Access* **2021**, *9*, 59196–59210. [CrossRef]
16. Kang, H.I.; Lee, B.; Kim, K. Path planning algorithm using the particle swarm optimization and the improved Dijkstra algorithm. In Proceedings of the 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, China, 19–20 December 2008; IEEE: Piscataway, NJ, USA, 2008; Volume 2, pp. 1002–1004.
17. Luo, M.; Hou, X.; Yang, J. Surface optimal path planning using an extended Dijkstra algorithm. *IEEE Access* **2020**, *8*, 147827–147838. [CrossRef]
18. Yang, R.; Ma, Y.; Tao, Z.; Yang, R. A stealthy route planning algorithm for the fourth generation fighters. In Proceedings of the 2017 International Conference on Mechanical, System and Control Engineering (ICMSC), St. Petersburg, Russia, 19–21 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 323–327.
19. Guan, J.; Huang, J.; Song, L.; Lu, X. Stealth Aircraft Penetration Trajectory Planning in 3D Complex Dynamic Environment Based on Sparse A* Algorithm. *Aerospace* **2024**, *11*, 87. [CrossRef]
20. Meng, B.b. UAV path planning based on bidirectional sparse A* search algorithm. In Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation, Changsha, China, 11–12 May 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 3, pp. 1106–1109.
21. Zhaoying, L.; Ruoling, S.; Zhao, Z. A new path planning method based on sparse A* algorithm with map segmentation. *Trans. Inst. Meas. Control* **2022**, *44*, 916–925. [CrossRef]
22. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
23. Panov, A.I.; Yakovlev, K.S.; Suvorov, R. Grid path planning with deep reinforcement learning: Preliminary results. *Procedia Comput. Sci.* **2018**, *123*, 347–353. [CrossRef]
24. Yang, Y.; Xiong, X.; Yan, Y. UAV Formation Trajectory Planning Algorithms: A Review. *Drones* **2023**, *7*, 62. [CrossRef]
25. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
26. Lei, X.; Zhang, Z.; Dong, P. Dynamic path planning of unknown environment based on deep reinforcement learning. *J. Robot.* **2018**, *2018*, 5781591. [CrossRef]
27. Bae, H.; Kim, G.; Kim, J.; Qian, D.; Lee, S. Multi-robot path planning method using reinforcement learning. *Appl. Sci.* **2019**, *9*, 3057. [CrossRef]
28. Tascioglu, E.; Gunes, A. Path-planning with minimum probability of detection for auvs using reinforcement learning. In Proceedings of the 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), Antalya, Turkey, 7–9 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
29. Qi, C.; Wu, C.; Lei, L.; Li, X.; Cong, P. UAV path planning based on the improved PPO algorithm. In Proceedings of the 2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE), Qingdao, China, 26–28 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 193–199.
30. Wang, H.; Lu, B.; Li, J.; Liu, T.; Xing, Y.; Lv, C.; Cao, D.; Li, J.; Zhang, J.; Hashemi, E. Risk assessment and mitigation in local path planning for autonomous vehicles with LSTM based predictive model. *IEEE Trans. Autom. Sci. Eng.* **2021**, *19*, 2738–2749. [CrossRef]

31. Zhang, J.; Guo, Y.; Zheng, L.; Yang, Q.; Shi, G.; Wu, Y. Real-time UAV path planning based on LSTM network. *J. Syst. Eng. Electron.* **2024**, *35*, 374–385. [CrossRef]
32. Ma, H.; Luo, Z.; Vo, T.V.; Sima, K.; Leong, T.Y. Highly efficient self-adaptive reward shaping for reinforcement learning. *arXiv* **2024**, arXiv:2408.03029.
33. Chu, K.; Zhu, X.; Zhu, W. Accelerating Lifelong Reinforcement Learning via Reshaping Rewards. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 619–624.
34. Kliem, J.; Dasgupta, P. Reward Shaping for Improved Learning in Real-Time Strategy Game Play. *arXiv* **2023**, arXiv:2311.16339.
35. Zare, M.; Kebria, P.M.; Khosravi, A.; Nahavandi, S. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Trans. Cybern.* **2024**, 1–14. [CrossRef] [PubMed]
36. Wu, F.; Ke, J.; Wu, A. Inverse reinforcement learning with the average reward criterion. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
37. Perrusquía, A.; Guo, W.; Fraser, B.; Wei, Z. Uncovering drone intentions using control physics informed machine learning. *Commun. Eng.* **2024**, *3*, 36. [CrossRef]
38. Singh, U.; Suttle, W.A.; Sadler, B.M.; Namboodiri, V.P.; Bedi, A.S. PIPER: Primitive-Informed Preference-Based Hierarchical Reinforcement Learning via Hindsight Relabeling. *arXiv* **2024**, arXiv:2404.13423.
39. Kanervisto, A.; Scheller, C.; Hautamäki, V. Action space shaping in deep reinforcement learning. In Proceedings of the 2020 IEEE Conference on Games (CoG), Osaka, Japan, 24–27 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 479–486.
40. Zahavy, T.; Haroush, M.; Merlis, N.; Mankowitz, D.J.; Mannor, S. Learn what not to learn: Action elimination with deep reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 3566–3577.
41. Zhuang, J.Y.; Zhang, L.; Zhao, S.Q.; Cao, J.; Wang, B.; Sun, H.B. Radar-based collision avoidance for unmanned surface vehicles. *China Ocean Eng.* **2016**, *30*, 867–883. [CrossRef]
42. Safa, A.; Verbelen, T.; Keuninckx, L.; Ocket, I.; Hartmann, M.; Bourdoux, A.; Catthoor, F.; Gielen, G.G. A low-complexity radar detector outperforming OS-CFAR for indoor drone obstacle avoidance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9162–9175. [CrossRef]
43. Scott, C.; Nowak, R. A Neyman–Pearson approach to statistical learning. *IEEE Trans. Inf. Theory* **2005**, *51*, 3806–3819. [CrossRef]
44. Li, S.E. Deep reinforcement learning. In *Reinforcement Learning for Sequential Decision and Optimal Control*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 365–402.
45. Zhou, L.; Ye, X.; Yang, X.; Shao, Y.; Liu, X.; Xie, P.; Tong, Y. A 3D-Sparse A* autonomous recovery path planning algorithm for Unmanned Surface Vehicle. *Ocean Eng.* **2024**, *301*, 117565. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.