Comparative Review and Analysis of Artificial Intelligence Based Learning Systems Used in Air Combat Simulation Environments

Emre Şafak[1,*], Tolga Erol[1], Hüseyin Furkan Ceran[1], Emre Ayvaz[2]

Abstract

The advancement of Artificial Intelligence (AI) technology is also manifesting itself in the defence sector. This study comprehensively reviews and comparatively analyses AI-based learning systems used in air combat simulation environments. The studies were categorised according to within-visual-range and beyond-visual-range scenarios and compared based on the methods used and the results obtained. These studies demonstrate that AI agents can develop realistic tactics, share situational awareness, and compete effectively with human pilots. In this study, methods such as Proximal Policy Optimisation (PPO) and Hierarchical Multi-Agent Reinforcement Learning (HMARL) were observed to be particularly prominent in terms of performance. Significant problems such as long training durations, high computational costs, partial observability, and sensitivity to sensor noise are the major challenges encountered. As a result of the study, it is concluded that intelligent agents developed in air combat simulation environments can enhance training efficiency by serving as teammates and opponents in pilot training.

Keywords: Air combat simulation, reinforcement learning, military simulation training, deep reinforcement learning, multi-agent systems, AI-driven tactical decision making, human-in the-loop.

Received: 10.11.2025 Accepted: 24.04.2026 Review Article

[1] HAVELSAN Inc., Ankara, Türkiye
[2] Turkish Air Force Command, Ankara, Türkiye
* Corresponding author.
✉ esafak@havelsan.com.tr

Cite this article as:
https://doi.org/10.65834/jdsi.12.30
Şafak, E., Erol, T., Ceran, H.F. & Ayvaz, E. (2026). Comparative Review and Analysis of Artificial Intelligence Based Learning Systems Used in Air Combat Simulation Environments. *Journal of Defence and Security Industries: Strategy and Technology 1(2)*, 88-109.

1. Introduction

Air combat environments are evolving towards sixth-generation platforms, collaborative Manned-Unmanned Teaming (MUM-T) concepts, and increasingly complex rules of engagement. Therefore, the simulation infrastructures used for pilot training and tactical development must also become "learning" and adaptable at the same pace. This necessity represents the evolution from rule-based Computer Generated Forces (CGF) approaches to today's Deep Reinforcement Learning (DRL) and Multi-Agent Reinforcement Learning (MARL)-based autonomous agents. Current studies integrating DRL into air combat aim to generate more realistic, adaptable, and less labour-intensive tactics. The ability to generate more engagements in a virtual environment than a human pilot could experience in real life will usher in a new era in pilot training. Unlike High Level Architecture and Distributed Interactive Simulation (HLA and DIS) simulators that run predefined scenarios, the Live-Virtual-Constructive (LVC) paradigm demands intelligent simulators. These systems must generate tactics, share situational awareness, and interact with humans online (Neville et al., 2015).

AI agents in high-fidelity air combat simulators are no longer static entities executing predefined scenarios. They have evolved into learning systems capable of reacting to opponents' manoeuvres, explaining discovered tactics, and sharing situational awareness across platforms. Mei et al. (2024) automated basic air-to-air manoeuvre decisions by framing single-aircraft manoeuvre generation with DRL. Kong et al. (2023) focused on coordinated behaviour by designing a Hierarchical Multi-Agent Reinforcement Learning (HMARL) framework for short-range, multi-aircraft scenarios. Sun et al. (2021) developed a self-play-based multi-agent policy gradient that enabled the emergence of tactics. Dantas et al. (2025) contributed to the DRL pipeline regarding engagement timing by using supervised learning to determine the optimal conditions and timing for missile launches in the Beyond Visual Range (BVR) environment. Toubman and Roessingh (2015) strengthened the practical aspects of guiding agent behaviour through novel reward-method development in training simulations, bringing it closer to application. Thus, these diverse yet complementary studies indicate that DRL/MARL-based agents for both Within Visual Range (WVR) and BVR air combat are approaching the maturity required to meet operational needs (Mei et al., 2024; Kong et al., 2023; Sun et al., 2021). Recent work has also extended this picture by addressing explainability, partial observability, and human-in-the-loop integration in increasingly realistic training settings. Despite this rapid progress, a significant gap remains in the literature: because these AI-based learning approaches are reported with different assumptions, fidelity levels, and human-in-the-loop integrations, it is still unclear which type of learning system a force structure should prefer. The objective of this article is to systematically explain AI-based learning systems used in air combat simulation environments according to the criteria of learning paradigm (single-agent DRL, MARL, hierarchical/curriculum learning), human-machine interaction (human-in-the-loop, explainability, super-agent integration), and operational realism (partial observability, obscured vision, multi-aircraft coordination, missile-launch timing). This study will enable command, control, and training authorities to distinguish

the differences between high-fidelity but costly-to-train systems and simpler yet rapidly adaptable systems. Thus, it will allow them to select the most suitable agent type for their specific network-centric warfare architecture.

This article reviews and provides a comparative analysis of studies utilizing artificial intelligence (AI) technology in air combat simulation environments. The second section covers the AI algorithms used in air combat simulation environments and the examined studies. The third section presents a detailed comparative analysis of the reviewed studies. The fourth section contains the conclusions reached from the examined studies and suggests potential future research directions.

2. Methodology

In the study, AI algorithms used in air combat simulation environments were first explained. Studies were categorized into WVR, BVR, and hybrid scenarios. Literature review was conducted using the IEEE Xplore, ScienceDirect, SpringerLink, Scopus, Web of Science, and ACM Digital Library databases. The search included English and Turkish journal and conference publications from the period 2012–2024, with selected early-access 2025 records included only where publication details could be verified. Specific Inclusion Criteria (IC) and Exclusion Criteria (EC) were applied to the selected studies to enhance the effectiveness of the research. These criteria are presented as:

- **IC1.** Studies must have been published in a conference, journal, press, or similar venue.
- **IC2.** Sufficient information must be provided in the abstract and the full text of the study.
- **IC3.** Studies using AI in air combat simulations are included in the research.
- **EC1.** Books, letters, notes and patents are not included in the review.
- **EC2.** Studies published in languages other than English and Turkish between 2012 and 2025 were excluded.
- **EC3.** Studies must be accessible; otherwise, they are not included in the review.
- **EC4.** Purely, theoretical articles lacking simulations are not included in the research.

2.1. AI Algorithms Used in Air Combat Simulation Environments

2.1.1. Reinforcement Learning

Reinforcement Learning (RL) is a branch of machine learning that focuses on how agents can learn to make decisions through trial and error to maximize the rewards received. RL enables machines to learn by interacting with an environment and receiving feedback based on their actions. This feedback occurs in the form of rewards or penalties (Sivamayil et al., 2023). In reinforcement learning, the algorithm works with an unlabelled dataset focused on a specific outcome. Each action the algorithm takes to explore the dataset generates positive, negative, or neutral feedback. This feedback constitutes the reinforcement part of the learning process. In the final stage, the model will be able to determine the best way to achieve the outcome. The general outline of how RL algorithm works is shown in Figure 1.

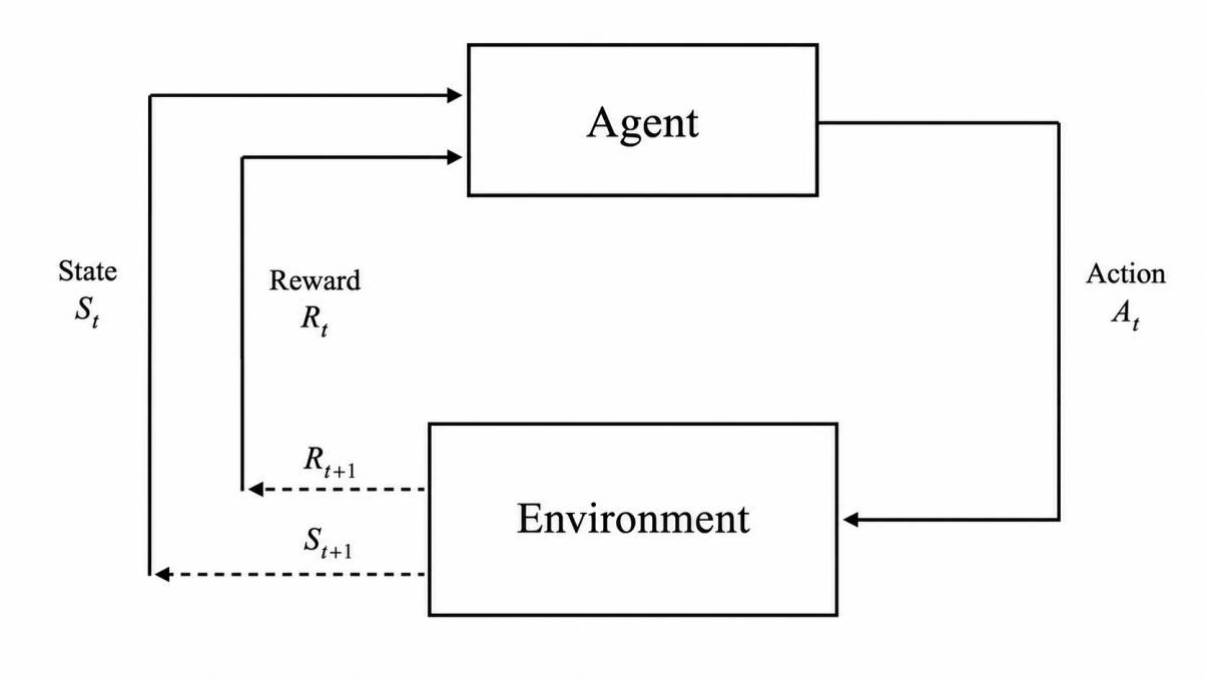

Figure 1. RL architecture (Sivamayil et al., 2023).

As shown in Figure 1, the RL process consists of four fundamental components: the agent, environment, action, and reward. The agent serves as the learner and decision-maker within the system. The external world with which the agent interacts is referred to as the environment. At each step, the agent makes choices through an action. The feedback the agent receives after each action, which indicates whether the outcome meets the expected level, is called a reward (Sivamayil et al., 2023). RL is divided into two types: model-based and model-free. In model-based reinforcement learning, the agent attempts to build a model of the environment. This model enables the agent to predict the consequences of its actions before executing them, allowing for a more planned and strategic approach. In model-free reinforcement learning, an explicit model of the environment is not constructed. Instead, the agent directly learns the optimal policy by associating actions with rewards based on the feedback received. RL algorithms are commonly used in dynamic environments due to their ability to adapt quickly to changing conditions and develop new strategies (Shakya et al., 2023).

2.1.1.1. DRL

DRL is a method that combines deep learning and RL algorithms. Thanks to deep learning, agents can learn rules directly from unstructured data. DRL is fundamentally based on Q-Learning, policy gradients, and actor-critic systems (Terven, 2025). Q-learning is a model-free RL algorithm that aims to enable an agent to make optimal decisions by interacting with the environment. The agent does not require a model of the environment; instead, it learns from its experiences by exploring different actions through trial and error. In Q-Learning, the goal is for the agent to build a Q-table that stores Q-values. Each Q-value estimates how correct it is to perform a specific action in a given state in terms of future rewards. The agent updates this table based on the feedback it receives (Jang et al., 2019). Policy gradient is a method that

directly models and optimizes the policy for the agent to obtain optimal models. Since the environment is generally unknown, it aims to address the challenge of the effect a policy update has on the state distribution (Francois-Lavet et al., 2018). The actor-critic algorithm provides a parallel training method. Each agent interacts with its own environment, using multiple actors that learn in parallel. The agents add updates to a shared global model, enabling faster and more stable training (Grondman et al., 2012). Curriculum learning is a method where training data is ordered from simple to complex examples and then used in the training process. By sequencing the data, the model used in Deep RL can learn better and achieve improved generalisation (Soviany et al., 2022).

2.1.1.2. MARL

While single-agent RL has made significant progress, real-world problems often involve multiple agents. This has created a need for multi-agent learning systems. MARL is an approach in which multiple agents can communicate and simultaneously influence the environment. The MARL method enables a multi-agent structure by allowing agents to learn, cooperate, and compete. A key challenge faced by MARL is the non-stationarity of the environment, caused by agents learning and adapting to one another (Ning & Xie, 2024). In MARL, agents can exhibit cooperative, competitive, or mixed behaviours depending on their interactions and goals. In cooperative MARL, agents work in coordination to complete a task, share a reward, and find the optimal policy for the overall system. Success depends on effective coordination and communication. Competitive MARL is a method in which agents have conflicting goals: they try to maximise their own rewards while minimising those of other agents. Mixed MARL combines cooperation and competition, where agents may cooperate but also face competition (Huh & Mohapatra, 2023). In multi-agent learning, the iterative process in which an agent improves its own performance, thereby leading to environmental variability that affects both itself and other agents, is called self-curriculum. The self-curriculum cycle results in different learning stages, where each stage is built upon the previous one. Self-curriculum is particularly evident in competitive environments where groups of opposing agents continuously change their strategies in response to their opponents' actions (Baker et al., 2019).

2.1.2. Long Short-Term Memory

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) designed to prevent the decay of feedback loop signals between network outputs based on the input. Thanks to these feedback loops, recurrent networks demonstrate better performance in pattern recognition. LSTM mitigates the vanishing gradient problem, a situation where the activation function becomes ineffective, thereby achieving better performance compared to other RNN architectures (Hochreiter & Schmidhuber, 1997). The LSTM architecture, which consists of three gates input, forget, and output is shown in Figure 2.

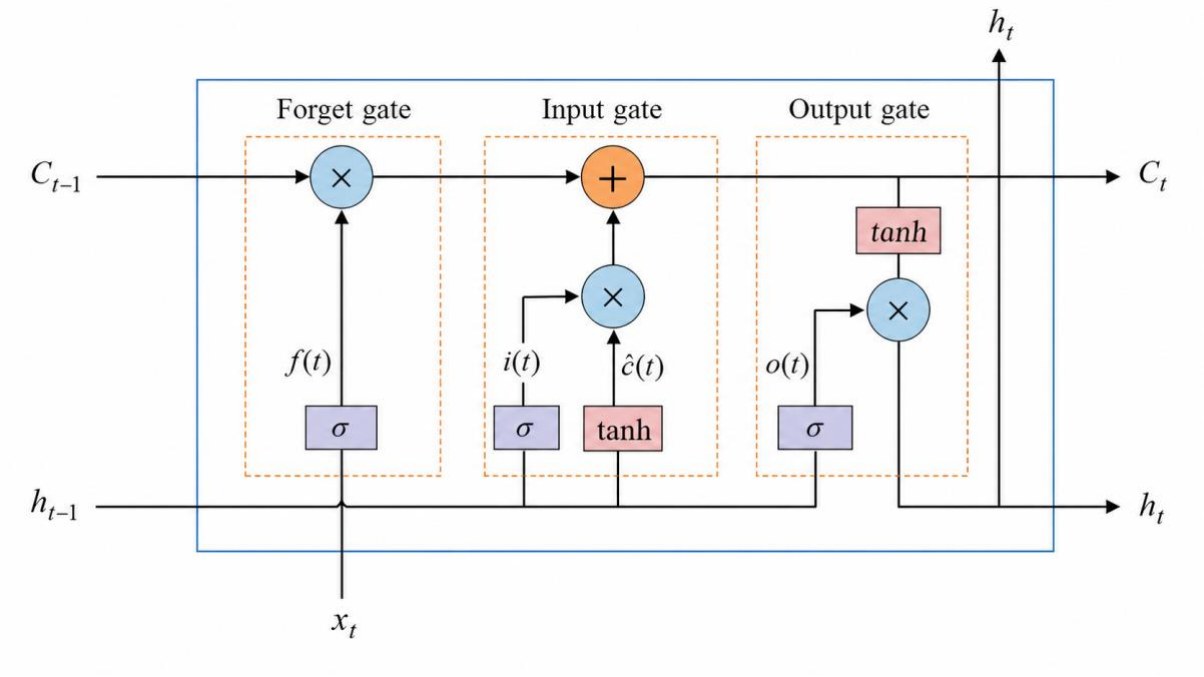

Figure 2. LSTM architecture (Mohsen et al., 2021).

As shown in Figure 2, the input gate determines which information will be added to the memory cell. The forget gate determines which information will be removed from the memory cells. The output gate identifies which information will be provided as output (Mohsen et al., 2021). Bidirectional LSTM (BiLSTM) is a type of LSTM that enables the bidirectional learning of time series and sequential data. BiLSTM networks pass the input data through LSTM layers twice, which allows for additional training and enhances performance. BiLSTM consists of two components: a forward LSTM and a backward LSTM. The forward LSTM facilitates learning from the first-time step to the last time step. The backward LSTM facilitates learning from the last time step to the first-time step. After the data passes through the two LSTM components, the outputs are merged (Chang et al., 2020).

2.2. AI Algorithms Used in Air Combat Simulation Environments

2.2.1. WVR

WVR is an air combat scenario that takes place within the range where the pilot can visually see the enemy aircraft. Due to the close distance, short-range weapons are used. The pilot's physical skill, situational awareness, and the aircraft's manoeuvrability are the most decisive factors. AI-based studies conducted for the WVR scenario are explained in this section.

Teng et al. (2012) compared the effectiveness of an Adaptive Computer-Generated Force (CGF), capable of learning and adapting through real-time interactions, is compared against traditional, doctrine-based CGFs used in aviation simulations for pilot training. The proposed method developed an Adaptive CGF model based on RL, capable of learning counter-tactics from human pilots in real-time. The foundation of the model is a self-learning neural network called FALCON, based on Adaptive Resonance Theory (ART). The value functions (Q-values)

of state-action pairs are learned using the FALCON algorithm integrated with the Temporal Difference method (TD-Falcon). An Adaptive ε-Greedy Policy was used for action selection. The experiments utilized the STRIVE CGF Studio simulation platform. The TD-FALCON model was connected to the simulator using a client-server architecture. In the initial trials, the Adaptive CGF generally achieved losses and draws against novice pilots. Over time, it learned effective strategies and achieved more victories by the 3rd trial. However, it could not maintain this advantage as the novice pilots also adapted quickly. Against veteran pilots, the Adaptive CGF could not establish a clear superiority and mostly suffered defeats and draws. Yet, over time, it learned defensive strategies and managed to increase the number of draws. Compared to traditional, non-adaptive CGFs, the Adaptive CGF demonstrated surprising manoeuvres and unexpected behaviours, thereby providing a more realistic and challenging training environment for pilots. However, its learning speed was found to be insufficient, especially against experienced veteran pilots (Teng et al., 2012).

Bae et al. (2023) developed an AI model for WVR air combat manoeuvres. The model operates in a realistic setting with incomplete enemy data, sensor limits, and measurement errors. The Soft Actor-Critic (SAC) (Haarnoja et al., 2018) algorithm was used for the training process due to its effectiveness in continuous action spaces and robustness against noise. The JSBSim aerodynamic model was used as the simulation environment, and OpenAI Gym with the Python programming language was utilized for developing the reinforcement model. SAC-LSTM demonstrated better performance in all experiments compared to the model using only fully connected layers. When the visual sensor range was reduced from 10 km to 1 km, SAC-LSTM models completed the training and achieved a win rate of over 75% against the rule-based model. In contrast, SAC-FC, which does not use LSTM, had a low success rate of 27.3% at the 1 km range. When noise was added to sensor measurements, the SAC-LSTM model achieved a win rate of over 80%, while the SAC-FC model remained at 49.2%. The model trained without the proposed curriculum learning learned more slowly and achieved lower final performance compared to the model trained with the curriculum. The study focused solely on one on-one air combat scenarios and was not extended to multi-agent or swarm scenarios (Bae et al., 2023).

Selmonaj et al. (2024) focused on the explainability of MARL models used in air combat scenarios. The HMARL model developed consists of a high-level commander policy and low-level control policies such as attack, engagement, and defence. The agents were trained using the Proximal Policy Optimisation (Schulman et al., 2017) algorithm. The study utilised a custom 2D simulation platform developed in Python, based on Dassault Rafale aircraft dynamics. Explainability was pursued through methods such as policy simplification, reward decomposition, feature-contribution analysis, hierarchical modelling, and causal modelling. Experiments showed that integrating the hierarchical commander policy improved air combat performance. In explainability experiments, it was observed that as the commander's detection range increased, it selected more cautious (defensive) manoeuvres against numerically superior enemies. At lower detection ranges, situational awareness decreased, leading to persistence in aggressive tactics. In an analysis conducted for a specific aircraft, the bearing angle was

identified as the most critical factor in the commander's decision-making. The commander selected attack mode when the agent was in an advantageous position and defence mode when the enemy agent was directly confronting it. However, the realism of the study is limited by the fact that the simulations were conducted only in a 2D environment (Selmonaj et al., 2024).

Selmonaj et al. (2023) proposed a HMARL-based air combat manoeuvre framework for teams consisting of multiple and heterogeneous agents. The study addresses challenges such as high-dimensional state/action space, partial observability, and nonlinear flight dynamics by dividing the decision-making process into two abstraction layers: commander and unit level. In the proposed HMARL method, lower-level policies control individual aircraft, while a higher-level commander policy issues macro commands for the entire mission. The model was trained using PPO on two basic policies: engagement and disengagement. The policies utilised a sophisticated Actor-Critic neural network incorporating self-attention and Gated Recurrent Unit modules. Parameter sharing exists between agent types. A custom, lightweight 2D Python simulation platform was developed for fast simulation, based on Dassault Rafale dynamics, for use in training and experiments. Despite being trained in 2v2 scenarios through curriculum learning and the self-attention mechanism, the lower-level policy agents were able to operate in 5v5 scenarios as well. The commander policy improved combat performance in small team sizes (3v3). However, in larger team sizes, performance decreased due to partial observability, often resulting in draws or near-equal win/loss ratios. The simulations are limited to a 2D environment, which restricts realism (Selmonaj et al., 2023).

Kong et al. (2023) developed a hierarchical manoeuvre control architecture that was developed for multi-aircraft close air combat. In contrast to traditional one-on one air combat studies, the research focused on variable-sized formations, which are common in modern air combat. The study employed a HMARL structure. Three subtasks were defined: attack, defence, and firing. The Recurrent Soft Actor-Critic (RSAC) algorithm and self-play (SP) method were used to learn the sub-strategies. The open-source JSBSim flight dynamics model was used for aircraft dynamics, with the F-16 Fighting Falcon selected as the aircraft model. The simulation frequency was set to 50 Hz, and the command input frequency was set to 10 Hz. Experimental results achieved an average success rate of 75% in defence and a 50%-win rate in attack. In aggressive firing, the average damage per episode was 0.8. Since the action and state spaces were simplified, complex scenarios are not yet supported (Kong et al., 2023).

Zhu et al. (2024) developed an agent that was developed for close air combat using DRL. To address the problem of neural network plasticity loss in traditional curriculum learning and to develop an effective air combat strategy in a realistic digital twin environment, a Motivational Curriculum Learning based Distributed Proximal Policy Optimization (MCLDPPO) algorithm was proposed. This algorithm observes the agent's performance deficiencies, adds appropriate rewards, and ensures the agent's continuous improvement. A mechanism was designed that allows the agent to interrupt its current action and make a new decision when its threat perception changes (such as an enemy missile warning or entering/exiting the enemy's visual field). A Deep LSTM-based Actor Critic network was chosen to process local and global

observations. Using Unity3D, a high-fidelity digital twin air combat simulation environment was created, incorporating aerodynamics, radar, infrared, and missile systems. The simulation was modelled with the F-22 fighter jet and AIM-120 missiles. In tests against an expert rule-based Finite State Machine (FSM), the MCLDPPO agent achieved a 66%-win rate. The addition of the interruption mechanism improved the agent's performance by approximately 10%. The study focused solely on scenarios involving a single friendly aircraft against a single enemy aircraft (Zhu et al., 2024).

Saldiran et al. (2024) achieved global explainability was achieved for RL-trained AI agents in close air combat. In the safety-critical domain of air combat, understanding AI decision processes is of great importance alongside performance improvement. A Deep Q-Network-based RL approach, combined with a reward decomposition method that extends local explainability to a global level, was employed. The agent's decisions were explained by decomposing them into different reward types. The state space was divided into tactical regions such as attack, defence, neutral, and head-on, allowing for regional interpretation of the agent's behaviour. The relative contributions of different reward types were visualized to enhance comprehensibility. A 3DOF model was adopted for aircraft dynamics, and Python with the CleanRL library was used for training. The reward decomposition method was found effective in understanding the agent's tactical preferences. However, the method was not directly applied to policy-based RL methods, and the visualization approach was limited to only three reward types (Saldiran et al., 2024).

2.2.1.1. Dogfight

Israelsen et al. (2018) addressed the optimisation of behavioural parameters for AI-based sparring partners in aviation combat training. The primary focus is on optimising and adapting AI pilots, particularly in air-to-air dogfight scenarios, within a variable and noisy environment. The study employed the Gaussian Process Surrogate-based Bayesian Optimisation (GPBO) method. To address GPBO's limitations in handling highly variable and noisy objective functions, a new sampling technique called Hybrid Repeat/Multi-point Sampling (HRMS) was developed. HRMS combines Repeat Sampling (RS) and Multi-point Sampling (MS) strategies to both enhance the reliability of optimisation and enable the construction of a global model of the objective function. Acquisition functions used included Expected Improvement (EI), Upper Confidence Bound (UCB), and Thompson Sampling (TS). The GPBO method supported by HRMS produced more consistent and repeatable optimisation results compared to traditional Single-point Sampling (SS) methods. However, in high-dimensional spaces ($d \geq 6$), the performance of methods like Thompson Sampling decreases (Israelsen et al., 2018).

Pope et al. (2023) examined AlphaDogfight, developed as part of DARPA's Air Combat Evolution (ACE) program, is examined. A competitive experimental framework was established to measure the capability of AI agents to conduct air combat against human pilots in a simulation environment. Within the scope of the AlphaDogfight Trials, autonomous agents based on DRL were developed. The study proposes a novel Hierarchical Deep Reinforcement Learning (HDRL) approach called PHANG-MAN (Policy Hierarchy for Adaptive Novel

Generation of MANeuvers). In the proposed method, agents are trained separately at lower and upper levels. In lower-level policies, agents are trained separately for different combat situations (e.g., attack, defence, control), such as Control Zone, Aggressive Shooter, and Conservative Shooter. The high-level policy selector chooses the most suitable sub-policy based on the current context. The JSBSim open-source software was used for basic flight dynamics. The environment was extended based on OpenAI Gym, and multi-aircraft scenarios were supported. A realistic damage model was implemented by defining the Weapon Engagement Zone (WEZ). The adversarial self-play method, which involves agents competing against themselves and their past versions, was employed. In a match between the trained agent and an F-16 pilot, the AI agent defeated the human pilot with a score of 5-0. The most successful agents demonstrated high-precision weapon usage, rapid decision-making, and low response times. While the training only covered 1vs1 (single combat) scenarios, multi-agent coordination was not tested (Pope et al., 2023).

2.2.2. BVR

BVR is an air combat scenario where the target is not visually seen but is detected and engaged solely through radar, data links, or sensors. The decision-making process is based on sensors, radar warning systems, electronic warfare, and tactical data sharing. The missiles used are radar-guided medium/long-range air-to-air missiles. The pilot typically engages the target without ever seeing it. AI-based studies conducted for the BVR scenario are explained in this section.

Toubman et al. (2015) developed RL-based behaviour generation was developed for CGF used in air combat training simulations. Firstly, to address the problem of the chance factor in missile hits leading to incorrect rewards during the learning process, a new reward function named Probability-of-kill Rewards is proposed. The Probability-of-kill Rewards function rewards agents based on the expected outcome (the missile's probability of kill) rather than the actual outcome (whether the missile hits or not). The Dynamic Scripting RL agent method was used for training. The study utilized a customized air combat simulation environment modelling F-16 fighter jets. The scenario is based on a 2v1 air combat engagement. The proposed Pk reward function demonstrated statistically significant higher performance compared to traditional binary (win/loss) reward functions and expert knowledge-based reward functions. The proposed method consistently showed performance improvement of approximately 10% to 19.4% across all scenarios. The highest performance increase, 19.4%, was achieved against an enemy using close-range tactics when compared to the expert knowledge-based reward function. However, the performance is limited by the existing rules in the rule base, and the absolute optimal behaviour cannot be discovered (Toubman et al., 2015).

Piao et al. (2020) successfully taught BVR air combat tactics were successfully taught to AI agents using end-to-end RL and self-play, without relying on prior human knowledge. The Proximal Policy Optimization algorithm was used in training to improve performance. To address the cold start problem and enable rapid acquisition of basic air combat skills, a Key Air Combat Event Reward Shaping (KAERS) mechanism was developed. Instead of relying

solely on sparse end-of episode win/loss signals, KAERS provides rewards based on objective and sporadically occurring key air combat events such as radar lock, missile launch, and evasion. The simulation environment utilized an RL-focused air combat simulation framework developed by Northwestern Polytechnical University (NPU). By the end of the training, agents trained with KAERS demonstrated higher win rates compared to agents trained with traditional dense reward functions. However, the scope of the study remains limited to 1v1 scenarios only (Piao et al., 2020).

Sun et al. (2021) developed a MARL method was developed for AI-assisted air combat. The proposed Multi-Agent Hierarchical Policy Gradient (MAHPG) algorithm utilizes a PPO based training process. MAHPG features a hierarchical neural network architecture where the high-level makes discrete manoeuvre and firing decisions, while the low-level determines continuous parameters (normal load factor, velocity). RL agents enable learning against each other through self-play. The Key Event based Reward Shaping (KAERS) mechanism helps mitigate the sparse reward problem. The air combat simulation environment named WUKONG, developed by Northwestern Polytechnical University (NPU), was used. WUKONG can simulate both BVR and WVR scenarios. The MAHPG algorithm achieved a higher win rate compared to contemporary algorithms such as P-DQN, MADDPG, PPO, and A2C. MAHPG increased its win rate from 0% to 68% against an expert-based bot as training progressed. However, the framework is only built upon fully observable states (MDP) and has not been adapted for partially observable (POMDP) scenarios (Sun et al., 2021).

Dantas et al. (2025) trained autonomous fighter aircraft agents were trained using DRL for BVR air combat, specifically in two-versus-two (2v2) scenarios. A customized training environment named AsaGym was developed for BVR air combat simulations and DRL-based agent training. AsaGym includes an F-16 aircraft model, radar, missiles, and behaviour trees. For training, PPO, Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (Fujimoto et al., 2018) (TD3), and Advantage Actor Critic (A2C) algorithms were used and compared. In experiments conducted with only the leader using DRL and setups with both leader and wingman using DRL, PPO demonstrated the best performance in terms of average episode reward and episode length metrics. SAC learned stably but with lower performance than PPO, while TD3 and A2C exhibited inconsistent or limited performance. PPO achieved the highest average reward (0.75-0.80) and the shortest engagement times (125-175 steps), showing the best performance. However, PPO also had the longest training time (6.3-6.8 hours). SAC learned with lower rewards (0.5) than PPO but in a stable manner and had a shorter training time (5.0-5.7 hours). TD3 and A2C were the fastest algorithms to train (3.1-4.7 hours), although they achieved significantly lower rewards and had longer/less effective engagement times. The simulations were conducted at a fixed altitude (25,000 feet), and the complexity of three-dimensional manoeuvres was not addressed (Dantas et al., 2025).

Piao et al. (2024) proposed a novel algorithm named deep Excitation Inhibition Factored Manoeuvre (ENACTIVE), which integrates DRL with prior knowledge, to enhance air combat decision-making processes. Inspired by the Excitatory-Inhibitory (E/I) balance in

neuroscience, the algorithm enables the agent to adapt its decision-making timing based on situational awareness, learning when to act. The RL agents were trained on Offensive/Defensive Semantic Manoeuvre (ODSM) and Energy Semantic Manoeuvre (ESM). The experiments in the study were conducted using the WUKONG air combat simulation environment. The proposed ENACTIVE algorithm achieved a 62%-win rate against state-of-the-art algorithms such as MAHPG, PPO, A2C, and TempoRL. The results demonstrated that the ENACTIVE algorithm enables the agent to maintain E/I balance by making timely decisions at critical moments, rather than exhibiting overly reactive or overly passive behaviour. However, the proposed method has so far only been tested in one-on-one and two-on-two scenarios. Its scalability to more crowded scenarios such as 3v3 and 4v4 has not yet been verified (Piao et al., 2024).

2.2.3. BVR & WVR

Toubman et al. (2015) used transfer learning was used to transfer air combat behaviours from a simpler scenario (2v1) to a more complex one (2v2). The Reinforcement Learning-based Dynamic Scripting (DS) technique was used for the training process. Rule weights learned in the source task (2v1) were directly copied to the rule bases of the agents in the target task (2v2). To measure the success of the transfer, metrics such as Jumpstart, Asymptotic Performance, Total Trials Won, Transfer Ratio, and Time to Threshold were used. A specially developed air combat simulation environment was used for training, which modelled F-16-based fighter aircraft. Behaviours were controlled using scripts consisting of if-then rules. In the default, evading, and close-range tactics, the agents that underwent transfer showed higher performance across all five metrics compared to those without transfer. However, in the new and challenging lead-trail tactic scenario, although the transferred agents showed better initial and final performance, their learning process progressed more slowly. Furthermore, they demonstrated lower performance in terms of the total number of trials won and the time to reach the threshold metrics compared to the non-transferred agents (Toubman et al., 2015).

Selmonaj et al. (2025) developed a system that enables real-time interaction between pilots and AI-supported combat pilots. The system aims to facilitate human-AI teamwork, military training scenarios, and the discovery of innovative tactics. The study utilised MARL for training the agents. The model incorporated Multi-Agent Proximal Policy Optimization (MA-PPO), an Actor-Critic network architecture, the Centralized Training with Decentralized Execution (CTDE) paradigm, curriculum learning, and self-play methods to enhance performance. The agents were trained on attack, engage, and defend policies. Additionally, a commander policy was implemented to decide which control policy to activate. A custom 3D flight dynamics simulation environment based on JSBSim was used for the training process. The agents achieved an 80%-win rate in 10-versus-10 air combat scenarios. However, training the AI agents directly within VR-Forces has not yet been fully realised, and a separate training environment was used instead.

3. Discussion

WVR, BVR, and combined WVR&BVR studies have been compared according to defined criteria. Table 1 presents studies focused on WVR air combat scenarios.

Table 1. Studies on air combat scenarios within visual range.

Study	Primary Focus	Method	Simulation Environment	Performance Results
Teng et al., 2012	The development of Adaptive CGF for pilot training.	RL was used together with TD FALCON (ART-based neural network).	CAE STRIVE CGF Studio	Effective strategies were developed against novice pilots over time.
Bae et al., 2023	Manoeuvre generation was enabled in a realistic environment with sensor limitations.	SAC-LSTM (Soft Actor Critic with LSTM networks) and Curriculum Learning algorithms were used together.	JSBSim and OpenAI Gym	It demonstrated significantly better performance than non-LSTM models, with over 75%-win rate at 1 km sensor range and 80%-win rate under sensor noise.
Selmonaj et al., 2024	Explainability of MARL decisions in air combat tactics was achieved.	PPO and HMARL algorithms were used together.	Custom 2D Python simulation environment	The hierarchical commander policy improved performance. Agent behaviour was explained based on detection range and aspect angle.
Selmonaj et al., 2023	HMARL has been used to manoeuvre heterogeneous agents.	Hierarchical MARL, PPO, personal attention and GRU networks were used together.	Custom 2D Python Simulator to model Dassault Rafale	Agents trained in 2v2 were scaled to 5v5. While the commander policy improved performance in small teams, it was challenged by partial observability in larger teams.

Table 1. (Continued)

Kong et al., 2023	Multi plane within range combat studied in variable conditions.	Hierarchical MARL, Repetitive Soft Actor-Critic (RSAC) and Self-Play methods were used.	JSBsim with F-16 Model	75% success rate in defence and a 50%-win rate in offense was achieved.
Zhu et al., 2024	Deep RL agents are trained in a high-quality digital twin environment.	The MCLDPPO (Motivational Curriculum Learning DPPO) method suitable for action cuts is presented.	High Quality F-22 and AIM 120 digital twins created in Unity3D.	A 66%-win rate was achieved against a rule based expert agent. Performance increased by 10% with interrupt rates.
Saldiran et al., 2024	An attempt has been made to ensure the explainability of AI agent tactics.	Deep Q Network with reward decomposition.	3DOF Aircraft Model	The reward decomposition method has been effective in understanding tactical choices.
Israelsen et al., 2018	Optimized AI sparring partner parameters for dogfight training.	Bayesian Optimization with Gaussian Process Surrogate (GPBO) and Hybrid Repeat/Multi point Sampling (HRMS).	Not specified	HRMS has shown more consistent and reproducible results compared to single-point methods.
Israelsen et al., 2018	AI decision makers are adaptively trained in dogfight simulations.	Gaussian Process Bayesian Optimization (GPBO) with Hybrid Repeat/Multi point Sampling (HRMS).	Mission Simulation System	HRMS showed 0.6 times lower variances and required 30-40% less simulation training than classical GPBO.
Pope et al., 2023	AI agents have been developed for simulated dog fighting aerial combat against human pilots.	Hierarchical Deep RL (HRL) and Adversarial Self-Play.	JSBSim	The AI agent defeated an experienced F-16 human pilot 5-0. Demonstrated high accuracy of weapon selection and ability to make quick decisions.

Table 1 shows a clear shift from basic RL (Teng et al., 2012) towards more complex methods such as HMARL (Selmonaj et al., 2024) and advanced actor-critic algorithms combined with LSTM to handle partial observability (Bae et al., 2023). Recent studies address real-world challenges such as sensor noise, limited range, and partial information (Bae et al., 2023). Although dogfight studies (Israelsen et al., 2018) and some WVR studies focus on 1v1 combat, there is a significant research trend towards multi-agent scenarios (Selmonaj et al., 2023). New research areas are emerging that focus not only on performance but also on the explainability and trustworthiness of AI decisions (Saldiran et al., 2024; Selmonaj et al., 2024). The environments used have evolved from proprietary platforms (STRIVE, MSS) to open-source platforms (JSBSim, OpenAI Gym) and high-fidelity digital twins (Unity3D). This evolution has increased both accessibility and realism. From a performance perspective, RL-based methods have demonstrated high adaptability, while PPO and MARL-based models have provided stable learning curves. However, training durations remain long, and computational costs are high. Although computer vision-based models have improved perceptual accuracy, they have shown sensitivity to sensor noise. Overall, WVR research has made significant progress in tactical flexibility and autonomous decision-making; however, there is a future need for hybrid systems that are effectively integrated with real flight data. Studies focusing on BVR air combat scenarios are presented in Table 2.

Table 2. Studies on BVR air combat scenario.

Study	Primary Focus	Method	Simulation Environment	Performance Results
Toubman et al., 2015	An adaptive reinforcement algorithm was used for Computer Generated Forces.	A reinforcement learning-based dynamic script generation technique with a Probability of Kill (Pk) reward was used.	A custom air combat simulation environment built with an F-16 model	The Pk reward improved performance by between 10% and 19.4% compared to the baseline.
Piao et al., 2020	BVR tactics were taught through end-to-end RL without any prior human knowledge.	Key Air Combat Event Reward Shaping (KAERS) mechanism and PPO.	The NPU air combat simulation environment	KAERS trained agents achieved a higher win rate than dense reward agents.
Sun et al., 2021	RL was used to develop multi agent forces for large-scale air combat tactics.	The Multi-Agent Hierarchical Policy Gradient (MAHPG), PPO, and KAERS algorithms were used together.	WUKONG simulation framework	68% win rate was achieved against the expert bot after self-play training.

Table 2. (Continued)

Dantas et al., 2025	An autonomous Deep RL agent with collective situational awareness was trained for 2v2 BVR air combat.	PPO, SAC, TD3, and A2C methods were trained and their performances were compared.	AsaGym simulation environment	PPO achieved the highest win rate, ranging between 75% and 80%, and the fastest convergence times.
Piao et al., 2024	The ENACTIVE algorithm, which combines inhibition balance inspired by neuroscience with Deep RL, was developed.	ENACTIVE (Deep Inhibitory Factorized RL) algorithm was used.	WUKONG air combat simulation environment	ENACTIVE achieved a better win rate, at 62%, compared to the PPO, A2C, and TempoRL algorithms.

As can be seen in Table 2, recent studies on BVR air combat AI demonstrate diversity ranging from rule-based behaviour generation (Toubman et al., 2015) to DRL with multi-agent coordination and neuroscience-inspired decision models (Piao et al., 2024). While early studies largely focused on hand-crafted rewards or scenario constraints, subsequent studies have been adapted to enable the exploration of tactical manoeuvres. Studies using PPO are observed to achieve better results. Comparative studies (Dantas et al., 2025) demonstrate that PPO performs better than SAC, TD3, and A2C in both convergence speed and tactical consistency. The integration of self-play strategies into the training environment (Sun et al., 2021) enhanced generalisation capabilities in dynamic environments. While better results were obtained in single-agent (1v1) scenarios, scalability issues emerged in multi-agent (2v2, 3v3, etc.) environments. Dantas et al. (2025) directly address the coordination problem in 2v2 combat. Electronic warfare, radar jamming, and communication constraints have been included to a limited extent in BVR studies. Most simulation environments simplify some aspects of three-dimensional engagement dynamics. Table 3 presents studies that encompass both BVR and WVR air combat scenarios.

Table 3. Studies on Beyond Visual Range & Within Visual Range air combat scenarios.

Study	Primary Focus	Method	Simulation Environment	Performance Results
Toubman et al., 2015	It was ensured that the learned air combat behaviours were transferred from simple (2v1) scenarios to complex (2v2) scenarios.	Transfer Learning applied to Dynamic Scripting agents	Custom developed F-16 model air combat simulation environment	Transfer-trained agents outperformed in all but the most complex lead trail scenario.

Table 3. (Continued)

Sun et al., 2021	A HMARL framework was developed for tactics that span both within visual-range and beyond visual-range air combat.	MAHPG, PPO, self-play, and KAERS were used together.	WUKONG simulation framework	The framework achieved a 68%-win rate against an expert bot and showed adaptability across both combat regimes.
Selmonaj et al., 2025	A system that enables real-time human-AI air combat interaction and teamwork has been developed.	A human-AI environment has been prepared through the MARL framework.	Custom 3D flight dynamics simulator based on JSBSim	Agents achieved 80%-win rate in 10v10 air combat scenarios.

As seen in Table 3, there is a clear trend towards multi-agent reinforcement learning, transferability, and human-AI collaboration in air combat modelling. The studies demonstrate a progression from single-task tactical learning (Toubman et al., 2015), to hierarchical multi-agent tactical coordination across mixed combat regimes (Sun et al., 2021), and ultimately to human–AI integrated combat systems (Selmonaj et al., 2025). This pattern represents the current frontier in the use of AI technology in air combat.

4. Conclusion

The study presented a comprehensive comparative analysis of AI-based learning systems used in air combat simulation environments. Research in WVR, BVR, and hybrid scenarios were examined. It was demonstrated that there is a transition from basic RL and rule-based systems to advanced DRL and MARL architectures. Hierarchical structures, curriculum learning, self-play, and LSTM memory mechanisms have greatly improved agents' decision-making, adaptability, and realism in dynamic air combat. DRL and MARL demonstrated superior performance in generating complex air combat manoeuvres and ensuring coordination among multiple agents. The importance of explainability has emerged for the operational use of the developed systems, to ensure their decisions are understandable and reliable. In the studies conducted, high computational costs and long training durations are required, especially in complex multi-agent scenarios. The intelligent agents developed for air combat simulation environments have the potential to revolutionize pilot training as adaptive, challenging, and realistic AI opponents and teammates.

The AI-based learning systems examined in this study have been comparatively analysed for their strengths and limitations across different scenarios. In WVR scenarios, LSTM-based algorithms show high robustness against sensor noise and achieve successful results even in limited visual ranges. In contrast, MARL approaches excel particularly in multi-aircraft coordination but suffer from performance degradation in large-scale team scenarios due to

partial observability. In BVR studies, the design of reward functions plays a critical role, directly affecting the realism of missile launch timing and tactical success. While some research in the literature focuses only on WVR or only on BVR scenarios, fewer studies address these two environments together, investigating the algorithms' adaptability to different tactical conditions. Studies that jointly consider both environments demonstrate the scenario-independent generalizability and multidimensional decision-making capacity of AI-based learning systems. Analytically, scenarios combining BVR and WVR offer greater operational realism compared to single-focus studies. However, this approach requires more complex reward functions and more powerful simulation infrastructure. While SAC-LSTM models, which are resistant to sensor noise, are preferred for generating realistic scenarios in pilot training, PPO-based HMARL models have been found to be more efficient in tasks requiring small-team coordination.

Among the reviewed studies, SAC-LSTM models, which exhibit high resilience to sensor noise, stand out for their use in realistic scenarios for pilot training. These models demonstrate stable performance even under sensor errors and limited range conditions, providing reliable results in individual manoeuvre decisions. In contrast, PPO-based HMARL models have shown more efficient performance in tasks requiring team coordination. Thanks to their commander-unit level decision-making structure, these models enhance coordination in multi-aircraft scenarios and strengthen tactical harmony.

Furthermore, it paves the way for the development of advanced autonomous systems for mission planning and execution by enhancing overall combat effectiveness. There is a need to develop more sample-efficient and computationally cost-effective training algorithms in the future. The development of standardized, high-reliability, open-source simulation benchmarks will significantly accelerate progress and enable fair comparisons. Finally, to enhance the reliability of agents, the explainability of AI agents' decisions must be ensured.

Conflict of Interest

The authors declare that they have no conflict of interest related to this study.

Funding

This research received no funding.

Ethical Statements

Not applicable.

Data and Code Availability

Data sharing is not applicable to this article as no new datasets were generated or analysed. No custom code was developed or used in this study.

Supplementary Materials

Not applicable.

References

Bae, J. H., Jung, H., Kim, S., Kim, S., & Kim, Y.-D. (2023). Deep reinforcement learning-based air-to-air combat maneuver generation in a realistic environment. *IEEE Access,* 11, 29249–29265. https://doi.org/10.1109/ACCESS.2023.3257849

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv*. https://doi.org/10.48550/arXiv.1909.07528

Chang, K.-C., Chu, K.-C., Wang, H.-C., Lin, Y.-C., & Pan, J.-S. (2020). Agent-based middleware framework using distributed CPS for improving resource utilization in smart city. *Future Generation Computer Systems,* 108, 445–453. https://doi.org/10.1016/j.future.2020.03.006

Dantas, J. P. A., Medeiros, F. L. L., Samersla, A. R., Botelho, P. L. R., Gomes, V. C. F., Silva, S. R., Ferreira, Y. D., Arantes, A. O., Aquino, M. R. C., & Maximo, M. R. O. A. (2025). Deep reinforcement learning agents with collective situational awareness for beyond visual range air combat. *IEEE Access,* 13, 143052–143069. https://doi.org/10.1109/ACCESS.2025.3597199

Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An introduction to deep reinforcement learning. *arXiv*. https://doi.org/10.48550/arXiv.1811.12560

Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *arXiv*. https://doi.org/10.48550/arXiv.1802.09477

Grondman, I., Busoniu, L., Lopes, G. A. D., & Babuska, R. (2012). A survey of actor critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(6), 1291–1307. https://doi.org/10.1109/TSMCC.2012.2218595

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv.* https://doi.org/10.48550/arXiv.1801.01290

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation,* 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huh, D., & Mohapatra, P. (2023). Multi-agent reinforcement learning: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.2312.10256

Israelsen, B., Ahmed, N., Center, K., Green, R., & Bennett, W. Jr. (2018). Adaptive simulation-based training of artificial-intelligence decision makers using Bayesian optimization. *Journal of Aerospace Information Systems,* 15(2), 38–56. https://doi.org/10.2514/1.I010553

Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. *IEEE Access,* 7, 133653-133667. https://doi.org/10.1109/ACCESS.2019.2941229

Kong, W., Zhou, D., Du, Y., Zhou, Y., & Zhao, Y. (2023). Hierarchical multi-agent reinforcement learning for multi-aircraft close-range air combat. *IET Control Theory & Applications,* 17(13), 1840–1862. https://doi.org/10.1049/cth2.12413

Mei, J., Li, G., & Huang, H. (2024). Deep reinforcement-learning-based air-combat maneuver generation framework. *Mathematics,* 12(19), 3020. https://doi.org/10.3390/math12193020

Mohsen, S., Elkaseer, A., & Scholz, S. G. (2021). Industry 4.0-oriented deep learning models for human activity recognition. *IEEE Access,* 9, 150508–150521. https://doi.org/10.1109/ACCESS.2021.3125733

Neville, K. J., Thom McLean III, A. L., Sherwood, S., Kaste, K., Walwanis, M., & Bolton, A. (2015). Live-Virtual-Constructive (LVC) Training in Air Combat: Emergent Training Opportunities and Fidelity Ripple Effects. In *International Conference on Augmented Cognition* (pp. 82-90). Springer. https://doi.org/10.1007/978-3-319-20816-9_9

Ning, Z., & Xie, L. (2024). A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence,* 3(2), 73–91. https://doi.org/10.1016/j.jai.2024.02.003

Piao, H., Sun, Z., Meng, G., Chen, H., Qu, B., Lang, K., Sun, Y., Yang, S., & Peng, X. (2020). Beyond-visual-range air combat tactics auto-generation by reinforcement learning. *In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE*. https://doi.org/10.1109/IJCNN48605.2020.9207088

Piao, H. Y., Yang, S., Chen, H., Li, J., Yu, J., Peng, X., Yang, X., Yang, Z., Sun, Z., & Chang, Y. (2024). Discovering expert-level air combat knowledge via deep excitatory-inhibitory factorized reinforcement learning. *ACM Transactions on Intelligent Systems and Technology,* 15(4), Article 65, 1–28. https://doi.org/10.1145/3653979

Pope, A. P., Ide, J. S., Mićović, D., Diaz, H., Twedt, J. C., Alcedo, K., Walker, T. T., Rosenbluth, D., Ritholtz, L., & Javorsek, D. II. (2023). Hierarchical reinforcement

learning for air combat at DARPA's AlphaDogfight Trials. I*EEE Transactions on Artificial Intelligence,* 4(6), 1371–1385. https://doi.org/10.1109/TAI.2022.3222143

Saldiran, E., Hasanzade, M., Inalhan, G., & Tsourdos, A. (2024). Towards global explainability of artificial intelligence agent tactics in close air combat. *Aerospace,* 11(6), 415. https://doi.org/10.3390/aerospace11060415

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv.* https://doi.org/10.48550/arXiv.1707.06347

Selmonaj, A., Antonucci, A., Schneider, A., Rüegsegger, M., & Sommer, M. (2025). Explaining strategic decisions in multi-agent reinforcement learning for aerial combat tactics. *arXiv*. https://doi.org/10.48550/arXiv.2505.11311

Selmonaj, A., Del Rio, G., Schneider, A., & Antonucci, A. (2025). Towards human engagement with realistic AI combat pilots. *In Proceedings of the 13th International Conference on Human-Agent Interaction (HAI '25) (pp. 1–3). ACM.* https://doi.org/10.1145/3765766.3765827

Selmonaj, A., Szehr, O., Del Rio, G., Antonucci, A., Schneider, A., & Rüegsegger, M. (2023). Hierarchical multi-agent reinforcement learning for air combat maneuvering. *arXiv*. https://doi.org/10.48550/arXiv.2309.11247

Shakya, A. K., Pillai, G., & Chakrabarty, S. (2023). Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, *231*, 120495. https://doi.org/10.1016/j.eswa.2023.120495

Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6), 1526–1565. https://doi.org/10.1007/s11263-022-01611-x

Sivamayil, K., Rajasekar, E., Aljafari, B., Nikolovski, S., Vairavasundaram, S., & Vairavasundaram, I. (2023). A systematic study on reinforcement learning based applications. *Energies*, 16(3), 1512. https://doi.org/10.3390/en16031512

Sun, Z., Piao, H., Yang, Z., Zhao, Y., Zhan, G., Zhou, D., Meng, G., Chen, H., Chen, X., Qu, B., & Lu, Y. (2021). Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play. *Engineering Applications of Artificial Intelligence*, 98, 104112. https://doi.org/10.1016/j.engappai.2020.104112

Teng, T.-H., Tan, A.-H., Ong, W.-S., & Lee, K.-L. (2012). Adaptive CGF for pilots training in air combat simulation. *In Proceedings of the 2012 15th International Conference on Information Fusion (FUSION)* (pp. 2263–2270). IEEE.

Terven, J. (2025). Deep reinforcement learning: A chronological overview and methods. *AI*, 6(3), 46. https://doi.org/10.3390/ai6030046

Toubman, A., Roessingh, J. J. M., Spronck, P., Plaat, A., & van den Herik, H. J. (2015). Rewarding air combat behavior in training simulations. *In 2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1397–1402). IEEE. https://doi.org/10.1109/SMC.2015.248

Toubman, A., Roessingh, J. J. M., Spronck, P., Plaat, A., & van den Herik, H. J. (2015). Transfer learning of air combat behavior. *In 2015 IEEE 14th International Conference on Machine Learning and Applications* (ICMLA) (pp. 226–231). IEEE. https://doi.org/10.1109/ICMLA.2015.61

Zhu, J., Kuang, M., Zhou, W., Shi, H., & Han, X. (2024). Mastering air combat game with deep reinforcement learning. *Defence Technology*, 34, 295–312. https://doi.org/10.1016/j.dt.2023.08.019