Engineering Applications of Artificial Intelligence 136 (2024) 108911

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Survey paper

A survey on reinforcement learning in aviation applications

Pouria Razzaghi [a,*], Amin Tabrizian [a], Wei Guo [a], Shulu Chen [a], Abenezer Taye [a], Ellis Thompson [a], Alexis Bregeon [b], Ali Baheri [c], Peng Wei [a]

[a] George Washington University, Washington D.C., USA
[b] Ecole Nationale de l'Aviation Civile, Toulouse, France
[c] Rochester Institute of Technology, Rochester, NY, USA

ARTICLE INFO

Keywords:
Reinforcement learning
Deep reinforcement learning
Aviation
Aircraft
Machine learning
Artificial intelligence

ABSTRACT

Reinforcement learning (RL) has emerged as a powerful tool for addressing complex decision making problems in various domains, including aviation. This paper provides a comprehensive overview of RL and its applications in the aviation industry. We begin by introducing the fundamental concepts and algorithms of RL, highlighting their unique advantages in learning from interaction and optimizing decision-making processes. We then delve into a detailed examination of the successful implementation of RL methods in aviation, covering areas such as flight control, air traffic management, airline revenue management, aircraft maintenance scheduling, etc. Furthermore, we discuss the potential benefits of RL in enhancing safety, and sustainability within the aviation sector. Finally, we identify and explore open challenges and areas for future research, emphasizing the need for continued innovation and collaboration between the fields of reinforcement learning and aviation.

1. Introduction

Reinforcement learning (RL) has shown promising performances in complex sequential decision making problems. Compared with classical decision making methods such as control and optimization, RL takes advantages of the availability of large-scale datasets, fast-time simulators, state-of-the-art neural network architectures, and high-performance computing resources to build models and algorithms. In addition, RL methods demonstrate scalability, run-time efficiency, and generalizability in stochastic, non-linear and dynamic decision problems. Many of these problems can be found in aviation and aeronautical applications, ranging from the planning problems such as airline maintenance, air traffic flow management, crew scheduling and aircraft routing, to control problems such as aircraft sequencing and separation, collision avoidance, and flight adaptive control.

Significance and contributions. This paper provides a timely comprehensive survey of reinforcement learning in aviation applications. The objective of this work is to identify state-of-the-art RL algorithms in most successful aviation applications. In addition, we also identify the technical gaps in applying RL in the aviation sector, including the certification concerns of implementing learning-based, neural-network-in-the-loop RL models in safety–critical applications such as aircraft collision avoidance. We expect this paper will serve as an introductory reading material for researchers who plan to start working in this interdiscipline area of RL for aviation.

2. Preliminaries

2.1. Reinforcement Learning in Aviation

Compared with model-based control and optimization methods, RL provides a data-driven, learning-based framework to formulate and solve sequential decision-making problems. Reinforcement learning frameworks have become promising due to largely improved data availability and computing power in the aviation industry. Many aviation applications can be formulated or treated as sequential decision-making problems. Some of them are offline planning problems, while others need to be solved in an online manner and are safety–critical. In this survey paper, we first describe standard RL formulations and solutions. Then we survey the landscape of existing RL-based applications in aviation. Finally, we summarize the paper, identify the technical gaps, and suggest future directions of RL research in aviation.

The RL models and algorithms are comprehensively outlined in the remainder of this section. To begin, we briefly describe the RL problem formulation and a few key concepts. Following that, two classical categories of RL algorithms will be presented: value-based and policy-based leanings. We then will present the more advanced techniques as well as modern actor–critic methods and multi-agent reinforcement learning (MARL). The summarized categories of the RL models and algorithms are shown in Fig. 1.

* Corresponding author.
E-mail address: prazzaghi@smu.edu (P. Razzaghi).

https://doi.org/10.1016/j.engappai.2024.108911
Received 22 November 2022; Received in revised form 31 May 2024; Accepted 28 June 2024
Available online 15 July 2024

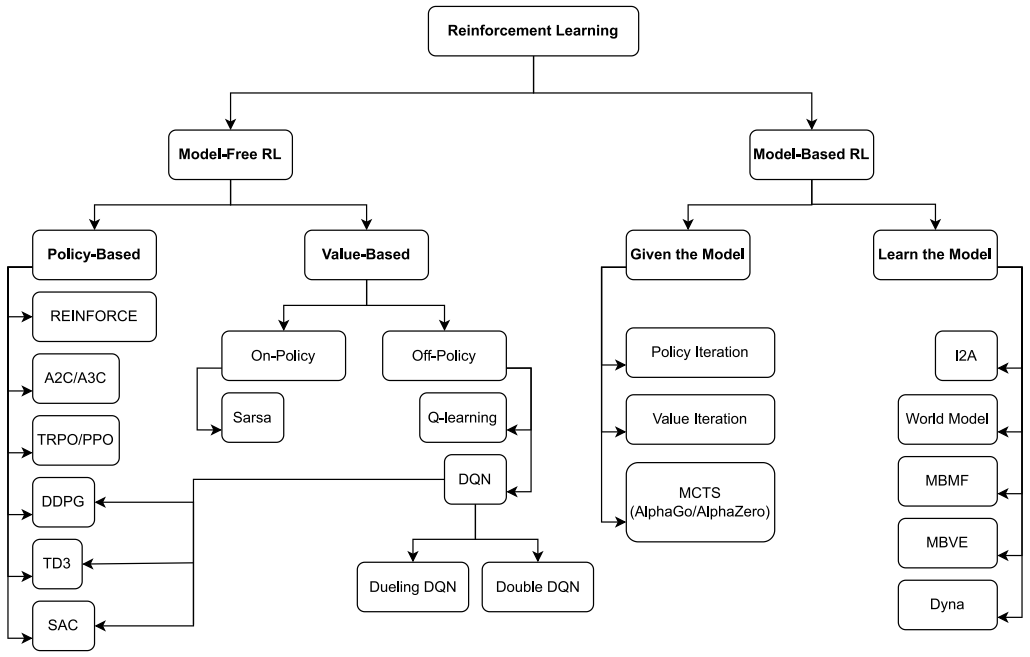

Fig. 1. Categories of the reinforcement learning models and algorithms.

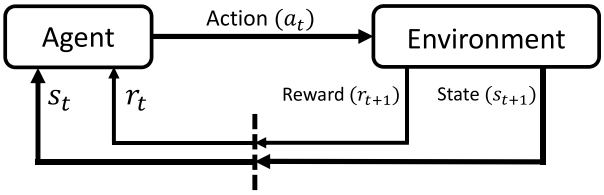

Fig. 2. In RL, an agent selects an action a_t based on its current state s_t, then it will receive a reward from the environment r_t and arrive to the next state s_{t+1}. This process will continue until the agent arrives at a terminal state if any.

2.2. Overview of reinforcement learning

Reinforcement learning is a branch of machine learning that is about an agent interacting with an environment to complete a task or achieve a goal. The environment is stated in the form of a Markov decision process (MDP) used to solve sequential decision-making problems. In an MDP problem, an agent takes a series of actions to maximize the total received reward from an unknown environment. This problem can be represented by a tuple of $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, P is the transition probability function $\left(P\left(s_{t+1}|s_t, a_t\right)\right)$ that maps a state–action (s_t, a_t) pair to distribution of next possible states, R is the received reward at each step, and γ is the discount factor representing the relative importance of future and immediate rewards. The policy, $\pi(.)$, represents a mapping from an agent's state to a distribution on the action space. The optimal policy, $\pi^*(.)$, takes place where the summation of expected rewards, $\left(\sum_{i=0}^{\infty} \gamma^i r^{t+i}\right)$, for the course of action is maximized. Fig. 2 depicts a block diagram of an RL process. An agent observes its current state and reward from the environment; then, the agent selects an action according to its policy. This will change the state of the environment and the new reward and state in the next time step will be pushed back to the agent.

One of the most important differentiation points in the RL algorithms is whether the agent has access to the model or not. Model-based RL algorithms either have direct access or can use a learned model of the environment, i.e., the agent knows $P\left(s_{t+1}|s_t, a_t\right)$, and reward r_t, while those algorithms that do not consider the environment model are known as model-free. In model-free RL, the agent should learn the optimal policy by observing past interactions or by directly interacting with the environment (see details in Fig. 1).

There are a few model-based RL algorithms such as policy iteration, value iteration (Bertsekas, 2012), world models (Ha and Schmidhuber, 2018), and model-based value expansion (MBVE) (Feinberg et al., 2018). Some algorithms have a combination of both model-free and model-based RL like imagination-augmented agents (I2 A) (Racanière et al., 2017), and model-based RL with model-free fine-tuning (MBMF) (Nagabandi et al., 2018). On the contrary, model-free methods have been extensively developed and used recently. In the following, we describe some traditional and modern model-free RL methods.

2.3. Standard formulations of model-free reinforcement learning

2.3.1. Value-based methods

Q-learning is one of the fundamental value-based RL algorithms introduced by Watkins (1989) at the end of the 1980s. A Q-value for every combination of state and action pair in an environment can be defined as Eq. (1). It represents an expected value of the cumulative reward at time step t for an action (a) when it follows a policy π as follows:

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i r_{t+i+1} | s, a\right] \tag{1}$$

where i is the number of steps forward at time step t. After updating the Q-value, the algorithm attempts to determine how valuable it is to take a particular action in a specific state. A Q-table is made by all the stored Q-values of each state–action pair in a discrete space (a Q-function approximator is used for a continuous state space-action). The policy $\pi(s) = \text{argmax}\, Q(s, a)$ yields the highest total reward. An agent selects an action to explore the environment (so-called exploration visiting almost all the state–action pairs a sufficient number of times) and observes the outcome. The Q-value can be updated by the temporal difference (TD) technique (Sutton and Barto, 2018):

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r_t + \gamma \max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right) \tag{2}$$

where $Q' = r_t + \gamma \max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1})$ is considered as the temporal difference target and α denotes the learning rate. We should note

that Eq. (1) is a stochastic approximation scheme for the Bellman optimality equation solution and it will converge to Q^* under certain assumptions (Tsitsiklis, 1994; Qu and Wierman, 2020).

An off-policy method learns the value of the optimal policy independent of the agent's actions. Q-learning is considered an off-policy learning algorithm since it involves updating Q-value based on experiences that are not always generated from the derived policy. Whereas state–action–reward–state–action (SARSA) is an on-policy one that generates experiences using the derived policy. For example, SARSA uses $Q' = r_t + \gamma Q(s_{t+1}, a_{t+1})$ where a_{t+1} is an action generated from the current policy or a given default policy.

A Monte Carlo method can be also used to estimate expected returns in non-Markovian episodic settings by averaging the results of multiple roll-outs. The Monte Carlo and TD methods have been joined and constructed the TD(λ) (Sutton and Barto, 2018). The major problem of the traditional methods (Watkins and Dayan, 1992; Rummery and Niranjan, 1994) is the "curse of dimensionality". These methods rely on storing all the state–action pairs and representing them in a tabular format that will grow exponentially as a factor of the number of states. One approach to solving this problem is to use a deep neural network (DNN) to approximate a parameterized Q-function. This creates a deep Q-networks (DQNs) (Mnih et al., 2015). DQN introduces replay memory and a separate target network, to overcome the problem of the instability and divergence issues in the training process of the approximation. To improve the stability of learning, this method uses a separate network $\hat{Q}$ for generating targets. A specific number of iterations is fixed for each episode. Moreover, by storing all transition experiences $(s_t, a_t, s_{t+1}, a_{t+1})$, the experience replay makes the random sampling for Q-learning updates more efficient (Lin, 1992). In addition, the DQN performance is further improved by several notable variants, such as continuous DQN (cDQN) (Gu et al., 2016), double DQN (Van Hasselt et al., 2016), dueling DQN (Wang et al., 2016b), and quantile regression DQN (QR-DQN) (Duan et al., 2021).

2.3.2. Policy-based methods

Another family of RL algorithms are policy gradient algorithms, which do not calculate value but attempt to determine an optimal policy directly. In these algorithms, a probability distribution over a set of actions ($\pi(a|s,\theta)$) concerning a policy defined as a function of parameters θ will be produced. An agent's likelihood of visiting state s after applying a policy π is described by the discounted state distribution. Using gradient ascent, the policy is optimized for the objective function:

$$J(\theta) = \int_S \rho_\pi(s)\, r(s, \pi_\theta(s))\, ds = \mathbb{E}_{s_{\rho_\pi}}[r(s, \pi_\theta(s))] \tag{3}$$

where ρ_π is the discounted state distribution (Sutton et al., 1999). The gradients are calculated ($\theta \leftarrow \theta + \eta \nabla J(\theta)$, where η is the step size) while the actions are taken following the policy, and rewards are observed. More straightforwardly, the policy gradient methods choose actions directly from a model and then update the model weights to maximize the expected returns. The original policy-based method is called REINFORCE (Williams, 1992), which collects a full trajectory and then updates the policy weights in the Monte Carlo style and indicates that the total return is sampled from the entire trajectory.

In the deterministic policy gradient (DPG) (Silver et al., 2014), instead of using a stochastic policy ($\pi(s,\theta)$), the actions are deterministically selected using policy $\mu(s,\theta)$. DPGs are limited cases of stochastic gradient policies when the variance becomes zero. The major drawback of a deterministic policy is the lack of exploration. For a proper exploration of the environment, the noise needs to be added and the policy becomes stochastic again $\left(\text{adding Gaussian noise } \xi, a = \pi_\theta(s) + \xi\right)$. DPGs are therefore commonly implemented as actor–critic methods to allow off-policy exploration. Consequently, it is possible to add noise to action outputs for additional exploration without the need for a stochastic policy. Over action-value modeling, policy parametrization has the advantage of incorporating knowledge into the learning system in the form of the policy. Deep deterministic policy gradient (DDPG) (Lillicrap et al., 2016) is a model-free off-policy algorithm for learning continuous actions, which combines ideas from DPG and DQN.

One of the problems of policy-based methods is in the gradient update. The policy performance drops if the updated policies deviate largely from previous ones. Trust region policy optimization (TRPO) (Schulman et al., 2015) ensures a monotonic improvement in policy performance by optimizing a surrogate objective function. The policy gradient updates are enforced by approximating the Kullback–Leibler (KL) divergence between the old and new policies using a quadratic approximation to be in a given range. Proximal policy optimization (PPO) (Schulman et al., 2017) achieves the same benefits as TRPO with a simplified implementation and improved sample complexity. It is revised based on TRPO but only uses first-order optimization.

It is worth mentioning that there is no specific way to differentiate and easily define the clusters of RL methods. Most of the aforementioned methods can be pointed out as actor–critic architecture as it is illustrated in Fig. 1.

2.3.3. Actor–critic methods

The actor–critic algorithm is an eminent and widely used architecture combining policy-based and value-based methods, inheriting their advantages (Ke et al., 2017). The actor–critic algorithm can be considered a TD learning method that represents the policy function independent of the value function. It introduces the eponymous components: the actor and the critic; the policy used to select actions is called the *actor*, and the estimated value function known as the *critic* criticizes the actions made by the actor (Sutton and Barto, 2018).

The actor–critic methods achieved great success in many complex tasks; however, they suffer from various problems such as high variance, slow convergence, and local optimum. Hence, many variants have been developed to improve the performance of actor–critic methods. Asynchronous advantage actor–critic (A3C) (Mnih et al., 2016) uses advantage estimates rather than discounted returns in the actor–critic framework and asynchronously updates both the policy and value networks on multiple parallel threads of the environment. The parallel independent environments stabilize the learning process and enable more exploration. Advantage actor–critic (A2C) (Wang et al., 2016a), the synchronous version of A3C, uses a single agent for simplicity or waits for each agent to finish its experience to collect multiple trajectories. This modification can significantly reduce the variance of the policy gradient estimate without changing the expectation. In this method, multiple actors are trained in parallel with different exploration policies, then the global parameters get updated based on all the learning results and synchronized to each actor. Soft actor–critic (SAC) (Haarnoja et al., 2018) with stochastic policies is an off-policy deep actor–critic algorithm based on the maximum entropy RL framework. It benefits from adding an entropy term to the reward function to encourage better exploration.

2.4. Multi-agent reinforcement learning methods

There are some newborn control tasks to regulate the behavior of a multi-agent system interacting in a common environment. Multi-agent reinforcement learning (MARL) will be critical for the development of communication skills and other intellectual capacities, as well as for teaching agents how to cooperate without causing harm to each other. These challenging tasks motivate researchers to use multi-agent RL frameworks (Gronauer and Diepold, 2022). A summary of related algorithms and theories is outlined in Zhang et al. (2021a). In the MARL framework, a set of N agents interact with the same environment. At each time step and for a given state, each agent takes its action, receiving a reward. The system then propagates to the next state. In the MARL framework, multi-task and partial observation are

usually considered (Omidshafiei et al., 2017). The centralized and decentralized multi-agent RL methods attract much attention in aviation applications (Wang et al., 2021b). One popular variant involves each agent adopting a policy, which determines the action based on local observations. As only local observations are required for the execution, this method permits decentralized implementation. However, centralized training is still required since the system's state transition relies on the actions of every agent. Here are some popular MARL methods:

- **Independent Q-Learning (IQL)**: Each agent learns its own Q-function independently, treating other agents as part of the environment. This approach is simple but can lead to non-stationarity issues because the environment appears to change as other agents learn (Tan, 1993).
- **Joint Action Learning (JAL)**: Agents learn a joint action-value function that depends on the actions of all agents. This approach can handle the non-stationarity problem but becomes infeasible as the number of agents and action spaces grows (Li et al., 2023).
- **Cooperative Multi-Agent Q-Learning**: In this approach, agents cooperate to learn a joint policy that maximizes the total reward for the group. This is often used in scenarios where agents share a common goal (Matignon et al., 2007).
- **Counterfactual Multi-Agent (COMA) Policy Gradients**: COMA uses a centralized critic to estimate the Q-function and decentralized actors to optimize the policies of individual agents. It addresses the credit assignment problem by using a counterfactual baseline that marginalizes a single agent's action while keeping others fixed (Foerster et al., 2018).
- **Multi-Agent Deep Deterministic Policy Gradient (MADDPG)**: An extension of the DDPG algorithm to multi-agent settings, where each agent has its own actor–critic pair. The critic is augmented with extra information about the policies of other agents, enabling centralized training with decentralized execution (Li et al., 2019b).
- **Mean Field Reinforcement Learning**: This approach is used for large-scale MARL problems, where interactions with other agents are approximated by interactions with an average agent or a mean field (Yang et al., 2018).
- **Value Decomposition Networks (VDN)**: VDN decomposes the joint value function into individual value functions for each agent, which are then combined additively. This allows for efficient learning while maintaining decentralized policies (Sunehag et al., 2017).
- **QMIX**: An extension of VDN, QMIX uses a mixing network to combine individual value functions in a more complex, non-linear way, subject to a monotonicity constraint. This allows for better representation of joint action values while still enabling decentralized policies (Rashid et al., 2020).

Each of these methods has its strengths and weaknesses, and the choice of method depends on the specific requirements of the multi-agent system, such as the level of cooperation, communication constraints, and the size of the action and state spaces.

3. Selected applications of RL in aviation

With the increasing complexity of airspace and the growth in air traffic, traditional control, optimization, and other decision making methods are being stretched to their limits. Many challenging problems in aviation can now be addressed using data-driven and machine-learning-based methods due to the availability of aviation data and significant increases in computational power. Here is a list of some of these problems: air traffic management (Schmidt, 2017), aircraft sequencing (Ahmed et al., 2018), air traffic flow management (Conde Rocha Murca et al., 2016), taxi-out time prediction (Lee et al., 2016), flight delay prediction (Takeichi et al., 2017; Choi et al., 2016), trajectory prediction (Ayhan and Samet, 2016), and aircraft performance parameter predicting (Alligier et al., 2015). In the following, we present a comprehensive description of each application and RL's role in providing the solution for their corresponding problems.

- Collision avoidance and separation assurance: it ensures that aircraft maintain a safe distance from each other and obstacles both in the air and on the ground. RL methods have emerged as a promising approach to enhance these safety mechanisms by enabling adaptive, real-time decision-making. Collision avoidance systems are designed to prevent aircraft from coming too close to each other or obstacles. These systems typically provide pilots or automated systems with warnings or recommended actions to maintain safe separation. RL can be used to develop more sophisticated collision avoidance algorithms that learn from experience and can adapt to a wide range of scenarios. Separation assurance involves maintaining a minimum safe distance between aircraft in all phases of flight. Air traffic controllers play a crucial role in separation assurance by directing aircraft along safe paths. RL can assist in this process by providing decision support tools that learn from past traffic patterns and optimize separation strategies in real time. This can help controllers manage traffic more efficiently, reduce the risk of close encounters, and minimize delays.
- Air traffic flow management: it ensures the safe and efficient movement of aircraft in the airspace and at airports. It encompasses a range of services and functions, including air traffic control, airspace management, and flow management. With the increasing complexity and volume of air traffic, maintaining efficiency and safety has led to the exploration of RL techniques. RL can be applied to manage the flow of air traffic in congested areas or during peak times, by learning to balance demand and capacity and minimize delays. In response to unforeseen events such as weather disruptions or emergencies, RL-based systems can dynamically reroute aircraft to ensure safety and minimize disruptions to the overall traffic flow. The integration of RL into air traffic management has the potential to enhance the adaptability, efficiency, and safety of the airspace system.
- Airline revenue management: it involves the use of sophisticated strategies to optimize the pricing and allocation of airline seats to maximize revenue. With the complexity of factors influencing ticket pricing, such as demand fluctuations, competition, and operational costs, RL models offer a promising solution by enabling dynamic and adaptive decision-making. Key challenges include accurately forecasting demand, dynamically adjusting prices based on market conditions, and balancing short-term gains with long-term profitability. RL can enhance airline revenue management by learning optimal pricing and inventory control policies through trial and error, using feedback from the market. By continuously interacting with the environment (i.e., the market), an RL agent can learn to make decisions that maximize cumulative revenue over time. Despite its potential, the application of RL in airline revenue management comes with challenges, such as the need for large amounts of data, computational complexity, and the requirement for robust and safe exploration strategies in a highly competitive market. Additionally, the integration of RL with existing revenue management systems and processes requires careful consideration.
- Aircraft flight and attitude control: it involves maintaining the stability and desired trajectory of an aircraft during flight and is responsible for managing the orientation (attitude) of an aircraft. This includes controlling the pitch, roll, and yaw angles, as well as the altitude and speed. Attitude control is particularly important for maintaining stability and ensuring that the aircraft responds correctly to pilot inputs and external disturbances. RL can be applied to flight and attitude control to develop controllers that learn optimal control strategies through interaction with

the environment. By continuously updating their control policies based on feedback from the aircraft's sensors and performance, RL-based controllers can adapt to changing conditions and uncertainties in the aircraft's dynamics. Implementing RL in flight and attitude control systems requires careful consideration of safety and robustness, as any failure could have severe consequences. Therefore, extensive simulation and testing are essential before deploying RL-based controllers in real-world scenarios. Additionally, integrating RL with existing control architectures and ensuring compatibility with aviation standards and regulations are important challenges to address.

- Fault tolerant controller: it ensures the safety and reliability of aircraft, especially in the presence of component failures or unexpected disturbances. These systems are designed to detect faults in the aircraft's control surfaces, engines, or other critical systems and then reconfigure the control strategy to maintain stable flight and safe operation. It can be broadly classified into two categories. *Passive fault-tolerant control*: these systems are designed with inherent robustness to handle faults without the need for detection and reconfiguration. They typically use redundant components and conservative control strategies to ensure stability under a range of fault conditions. *Active fault-tolerant control*: These systems actively detect and isolate faults, and then reconfigure the control strategy to compensate for the fault. This can involve switching to backup systems, adjusting control laws, or using alternative control surfaces. RL can play a significant role in enhancing active fault-tolerant control systems by enabling them to learn and adapt to faults in real time. Algorithms can learn to recognize patterns in sensor data that indicate the onset of a fault. By continuously updating their understanding of normal and faulty conditions, RL-based systems can improve their accuracy in detecting and diagnosing faults. Also, once a fault is detected, an RL agent can learn to select the optimal control strategy to maintain a safe flight. This might involve choosing which backup systems to activate or how to redistribute control authority among the remaining functional components.
- Aircraft flight planning: it involves determining the optimal route, altitude, and speed for a flight to ensure safety, efficiency, and compliance with regulations. It is a complex task that takes into account various factors such as weather conditions, airspace restrictions, fuel consumption, and air traffic control requirements. Traditional flight planning relies on predefined algorithms and models to calculate the best flight path. However, these methods may not always be able to adapt to real-time changes or unexpected events, such as sudden changes in weather or airspace closures. RL can address these challenges by learning from experience and continuously updating the flight plan based on new information. Reinforcement learning's role can be placed into different aspects such as dynamic route optimization, fuel-efficient flight planning, adaptive er-routing, and integration with air traffic control.
- Airline maintenance: it ensures that aircraft are safe, reliable, and available for service. It involves regular inspections, repairs, and overhauls of various aircraft components. Traditional maintenance strategies often rely on fixed schedules or reactive approaches, which may not be optimal in terms of cost, efficiency, or aircraft availability. Effective airline maintenance requires balancing several factors. *Safety*: Ensuring that all aircraft systems and components meet stringent safety standards. *Reliability*: Minimizing unexpected breakdowns and delays. *Cost*: Managing maintenance costs while maintaining high safety and reliability standards. *Availability*: Maximizing the time aircraft are available for service and minimizing downtime. RL methods can be used to enhance the quality of the following sides of airline maintenance. *Predictive maintenance*: RL can be used to develop predictive maintenance models that learn from historical data and sensor readings to predict when a component is likely to fail. This allows airlines to perform maintenance proactively, reducing unexpected failures and downtime. *Optimal scheduling*: RL algorithms can learn to schedule maintenance activities optimally, considering factors such as aircraft usage patterns, maintenance resource availability, and the cost of downtime. This can help airlines minimize maintenance costs while ensuring high levels of safety and reliability. *Resource allocation*: RL can assist in the dynamic allocation of maintenance resources, such as personnel and equipment, based on real-time needs and priorities. This can improve the efficiency of maintenance operations. *Spare parts inventory management*: RL can be used to optimize the inventory levels of spare parts, balancing the cost of holding inventory with the risk of stock-outs that could lead to maintenance delays.

Fig. 3. Taxonomy layout of RL in aviation. Different applications are shown with their corresponding illustrations. Flight planning represents a pre-defined path from the initial point. The revenue management illustration shows an increase in the profit of the airline. Controlling a drone lost one of its motors goes under the adaptive control of an air vehicle. A gimbal shape represents the attitude control of the system. The traffic management sketch depicts the control room monitor to supervise the traffic in the air. A collision avoidance picture is an alarm of avoiding a conflict between two vehicles.

Fig. 3 illustrates the taxonomy of using RL methods in different aviation applications. In the following sections, we try to summarize the utilization of the RL algorithms in these selected applications. To the best of the authors' knowledge, this survey paper is the first study that reviews the RL methods in aviation.

Also, it is worth mentioning that the simulation environments are publicly available. In the field of aviation, there are several public benchmarks and simulation environments available for training and testing various models and algorithms. These resources are crucial for researchers and practitioners to develop, evaluate, and compare their approaches. Here's an overview of some of the notable ones.

3.1. Public benchmarks and datasets

1. *Airline On-Time Performance Data* (Bureau of Transportation Statistics): This dataset contains information on flight arrival and departure details for commercial flights within the United States, sourced from the U.S. Department of Transportation's Bureau of Transportation Statistics. It is widely used for research

in airline operation optimization, delay prediction, and network analysis (US Department of Transportation, 2024).

2. *ACAS Xu Dataset*: The Airborne Collision Avoidance System X (ACAS X) is a family of collision avoidance systems developed by NASA. The ACAS Xu dataset is used for developing and evaluating collision avoidance models for unmanned aircraft (EKim and Bak, 2019).
3. *EUROCONTROL's DDR2 Dataset*: The Demand Data Repository 2 (DDR2) by EUROCONTROL provides air traffic management (ATM) related data, including flight plans, sector configurations, and traffic counts, which are useful for ATM research and simulations (EUROCONTROL, 2022).
4. *Flight Quest Dataset*: As part of the GE Flight Quest challenge on Kaggle, this dataset contains flight data aimed at improving flight efficiency, including information on routes, weather, and airspace constraints.

3.2. Simulation environments

1. *BlueSky ATC Simulator*: BlueSky is an open-source air traffic control simulator that is used for research and education in air traffic management. It allows for the simulation of airspace, aircraft, and ATC operations (Hoekstra and Ellerbroek, 2016).
2. *JSBSim Flight Dynamics Model*: JSBSim is an open-source flight dynamics model that can be used for flight control analysis, aircraft design, and flight simulation. It provides a flexible framework for modeling the dynamics of fixed-wing and rotary-wing aircraft (Berndt, 2004).
3. *FlightGear Flight Simulator*: FlightGear is a free, open-source flight simulator that can be used for research, education, and pilot training. It offers a realistic flight dynamics model and supports a wide range of aircraft (FLIGHTGEAR, 2023).
4. *OpenAI Gym*: While not specific to aviation, OpenAI Gym provides a standardized interface for reinforcement learning environments, including some that can be adapted for aviation-related tasks, such as control and navigation.

These resources provide valuable data and simulation capabilities for various applications in aviation, including flight dynamics, air traffic control, route optimization, and collision avoidance. Researchers and developers can leverage these tools to advance the state-of-the-art in aviation technology and safety.

3.3. Collision avoidance and separation assurance

Air traffic control (ATC) plays a crucial role as it is responsible for maintaining flight safety and efficiency. Collision avoidance is the last layer of defense against mid-air collision. Air traffic controllers must maintain a safe separation distance between any two aircraft at all times. This function is called conflict resolution or separation assurance. An early adaptation of an in-air collision avoidance system was the Traffic Alert and Collision Avoidance System (TCAS) (Harman, 1989) and more recently Next-Generation Airborne Collision Avoidance System (ACAS-X) (Jeannin et al., 2015; Kochenderfer et al., 2012). The latter was built upon TCAS, introducing a partially observable Markov decision process (POMDP) for the problem formulation. It provides audible and visual warnings to pilots by evaluating the time to closest approach, to determine if a collision is likely to occur. Many studies have been recently conducted on RL-based collision avoidance and separation assurance, which a selection is presented in Table 1.

A MDP collision avoidance in free-flight airspace was introduced in Bertram and Wei (2020). In a 3D environment with both cooperative (aircraft actively trying to avoid others) and non-cooperative aircraft (those not concerned with collision avoidance), the MDP formulation in a free flight was able to avoid collision between aircraft. In Li et al. (2019a), a Deep Reinforcement Learning (DRL) method was implemented as an optimization to a collision avoidance problem.

Showing beyond human-level performance in many challenging problems, the collision avoidance problem of unmanned aerial vehicles (UAV) has been solved by implementing a DQN algorithm (Wulfe, 2017). Deep Q-Learning from demonstrations (DQfD) and reward decomposition were implemented to provide interpretable aircraft collision avoidance solutions in Herman (2021). A DQN technique was also applied for the collision avoidance of UAVs (Wulfe, 2017), change routes and speeds in NASA's Sector 33 (Brittain and Wei, 2018), compute corrections on top of the existing collision avoidance approaches (Harman, 1989; Kochenderfer et al., 2012), and unmanned free flight traffic in dense airspace (Li et al., 2019a). A framework using RL and GPS waypoints to avoid collisions was suggested in Jacob et al. (2022). A double deep Q-network (DDQN) was applied to guide the aircraft through terminal sectors without collision in Xu et al. (2021). The approach tackles the cases where traditional collision avoidance methods fail namely in dense airspace, those expected to be occupied by UAVs, and demonstrated the ability to provide reasonable corrections to maintain sufficient safety among aircraft systems. An Intelligent and Safe Urban Air Mobility (ISUAM) system, which leverages DRL models to execute tactical deviation maneuvers within urban air mobility environments was introduced in Garcia et al. (2023), which the Dueling DQN showed the best performance in terms of satisfying the conflict resolution.

PPO methods are widely used in aircraft collision avoidance and have shown promising success. The problem of collision avoidance in structured airspace using PPO networks (Brittain and Wei, 2019) was addressed using a Long Short-term Memory (LSTM) network (Brittain et al., 2021), and attention networks (Brittain and Wei, 2021) to handle a variable number of aircraft. While these algorithms show high performance in the training environment, a slight change in the evaluation environment can decrease the performance of these PPO models. A safety module based on Monte-Carlo Dropout (Gal and Ghahramani, 2016) and execution-time data augmentation was proposed to solve the collision avoidance problem in environments, which are different from the training environments (Guo et al., 2021). A PPO network was proposed for unmanned aircraft to provide safe and efficient computational guidance of operations (Hu and Liu, 2020) and guided UAV in continuous state and action spaces to avoid collision with obstacles (Hu et al., 2022). A message-passing network (Dalmau and Allard, 2020) was introduced to support collision avoidance.

A prior physical information of airplanes was injected to build a physics-informed DRL algorithm for aircraft collision avoidance (Zhao and Liu, 2021). A reward engineering approach was proposed in Panoutsakopoulos et al. (2022) to support the PPO network to solve the collision avoidance problem in a 2D airspace.

Several studies have applied DDPG (Lillicrap et al., 2016) to aircraft collision avoidance problems. A DRL method was applied to resolve the conflict between two aircraft with continuous action space in the presence of uncertainty based on DPG in Pham et al. (2019b). Also, an intelligent interactive conflict solver was used to acquire ATCs' preferences and an RL agent to suggest conflict resolutions capturing those preferences (Tran et al., 2019). Later, the DDPG algorithm dealt with air sectors with increased traffic (Tran et al., 2020). A proper heading angle was obtained by the DDPG algorithm before the aircraft reached the boundary of the sector to avoid collisions (Wen et al., 2019). DDPG method was also proposed to mitigate collisions in high-density scenarios and uncertainties in Pham et al. (2019a). A mixed approach, which combines the traditional geometric resolution and the DDPG model, was proposed to avoid the conflicts (Ribeiro et al., 2020). Multi-agent deep deterministic policy gradient (MADDPG) was applied to pair-wisely solve the collisions between two aircraft (Isufaj et al., 2021). Another MADDPG-based conflict resolution method reduced the workloads of ATC and pilots in operation (Lai et al., 2021).

Table 1
Selection from the literature on RL in collision avoidance. State/Action space (S/A Space) can be continuous (C), discrete (D), or mixed (M).

Reference	S/A space	Algorithm	Policy class	Key features
Wulfe (2017)	D/D	Double DQN	ϵ-greedy	Prioritized sampling, regularization, and discretization of dynamics.
Herman (2021)	M/D	DQN		Deep Q-learning from demonstrations, reward decomposition.
Brittain and Wei (2019), Brittain et al. (2021) and Brittain and Wei (2021)	M/D	PPO, Attention network, and LSTM	ANN	1. Adopting a multi-agent framework to handle collision avoidance. 2. Using LSTM to enhance the performance of PPO.
Guo et al. (2021)	M/D	PPO, Dropout	ANN	Using Monte-Carlo Dropout and data augmentation to improve the safety in unseen environments.
Hu et al. (2022)	C/C	PPO	ANN	Developing continuous control for unmanned aircraft system.
Pham et al. (2019b)	C/C	DPG	ANN	Developing an air traffic scenario simulator.
Wang et al. (2019a)	C/C	K-control actor–critic	ANN	Two-dimensional continuous action selection.
Li et al. (2019a)	C/C	DPG	ANN	1. Building on ACAS to provide corrections for dense airspace. 2. Handling dense airspace.
Bertram and Wei (2020)	C/C	PPO	ANN	1. Introducing solution for high-density UAM airspace. 2. Using an MDP-based trajectory planner to avoid cooperative and non-cooperative aircraft. 3. Adopting a multi-agent system to handle large numbers of aircraft.
Harman (1989)	D/D	DDQN	ϵ-greedy	1. Introducing an onboard collision avoidance tool for pilots. 2. Interrogating an airspace with rule-based logic.
Jeannin et al. (2015)	M/D	DQN	–	Formal verification of ACAS-X.
Kochenderfer et al. (2012)	M/D	PPO, Dropout	ANN	Building on TCAS using a numeric lookup optimized to a probabilistic model.

The actor–critic algorithms are also popular in this application. K-control actor–critic algorithm was proposed to detect conflict and resolution with a 2D continuous action space in Wang et al. (2019a). A policy function returns a probability distribution over the actions that the agent can take based on the given state. A graph-based network for ATC in 3D unstructured airspace was built in Mollinga and van Hoof (2020) to manage the airspace by avoiding potential collisions and conflicts. A multi-layer RL model was proposed to guide an aircraft in a multi-dimensional goal problem (Zu et al., 2021). Also, an LSTM network and an actor–critic model were used to avoid collisions for fixed-wing UAVs (Zhao et al., 2021). A recent study aimed to enhance autonomous self-separation in Advanced Air Mobility (AAM) corridors by implementing speed and vertical maneuvers (Alvarez et al., 2023; Brittain et al., 2024). The research utilized a sample-efficient, off-policy soft actor–critic algorithm to guarantee both safe and efficient aircraft separation.

Besides the popular models, other RL methods were also implemented for collision avoidance. A message-passing-based decentralized computational guidance algorithm was proposed in Yang and Wei (2020), which used a multi-agent Monte Carlo tree search (MCTS) (Chaslot et al., 2021) formulation. It was also able to prevent loss of separation (LOS) for UAVs in an urban air mobility (UAM) setting. A highly efficient MDP-based decentralized algorithm was established to prevent conflict with cooperative and non-cooperative UAVs in the free flight airspace in Bertram and Wei (2020). The MuZero algorithm (Schrittwieser et al., 2020) was proposed to mitigate a collision in Yilmaz et al. (2021). Difference rewards tool was applied in Singh et al. (2021) and a graph convolutional reinforcement learning algorithm solved the multi-UAV conflict resolution problem (Isufaj et al., 2022). An MARL approach was proposed to mange high-density AAM structured airspace (Deniz et al., 2024).

Incorporating a DRL model to learn a collision avoidance strategy while training an NN simultaneously could reduce the learning time and execute a more accurate model due to removing the discretization problem (Julian et al., 2016). Though DRL has shown great success in aircraft separation assurance, there are still a lot of unsolved problems. These problems create crucial obstacles to building DRL models in this safety–critical application in the real world. One major problem is validation. DRL models for aircraft separation have deep structures and complex input states. The complex architecture makes it difficult to verify the properties of DRL models using traditional formal methods. Current work with formal methods can only validate very simple properties with shallow DRL models. The lack of validation limits the trustworthiness of these DRL models and their use in real-world applications.

Another important question is the gap between simulation and reality. RL for aircraft separation assurance is trained with simulators because real-world training is too expensive considering the potential damage. However, it is not possible to have a simulation mimic reality perfectly. The distribution shift between the simulation and reality may constrain the learning performance of the RL models.

Besides these two issues, RL for aircraft separation assurance also faces the problems of general RL models. For example, RL for separation assurance currently has a low sampling efficiency, which highly restricts the training speed. Also, the RL for the separation assurance model works as a black-box. It cannot provide explainable decision-making in this process.

3.4. Air traffic flow management

Air traffic management is an encompassing term for a system that directly affects or is used to decide air traffic movements. The overarching aim of these systems is to reduce delay while maintaining the operational safety of the airspace. Generally, air traffic flow and capacity management are part of a common air traffic service (ATS) and interface with either pilots directly or through ATC. These systems can be considered through two classifications; systems for unmanned traffic management (UTM) and UAS operations, and those for more conventional operations (see Table 2).

Air Traffic Flow and Management (ATFM) is a subset of traffic management that focuses on ensuring the available airspace capacity is used efficiently. The capacity can be influenced by the sector's geometry: size, shape, or altitudes as well as stochastic variables like wind, weather, and emergencies, or more constant variables like airport capacity and throughput. In choosing the role of an autonomous system, existing workflows must be observed with human actors to identify what could maximize performance. Using a MARL approach (Schrittwieser et al., 2020) presents two reward functions for both the Ground Holding Problem (GHP) and Air Holding Problem (AHP). The approach focuses on optimizing six objectives: minimize delays in the ATC sector; minimize delays in the terminal sector; minimize financial cost to airlines; improve fairness among airlines; reduce impact in ATC sectors by avoiding unnecessary actions; and improve safety in the ATFM. Based on these factors, the reward functions, when tested in Brazilian airspace, maintained safety standards when aiding air traffic controller decision-making and also improved efficiency and fairness among aircraft.

Table 2
Selection from the literature on RL in air traffic flow management.

Reference	S/A space	Algorithm	Policy class	Key features
Cruciol et al. (2013)	C/C	Agent-based MARL	Q-learning	1. Reward function considering safety and fairness in Ground Holding Problem (GHP). 2. Reward function considering safety and separation in Air Holding Problem.
Xu et al. (2020)	C/C	K-control and actor–critic	ANN	1. A collaborative approach to DCB. 2. Relax constraints of airspace configurations to optimize airspace utilization.
Huang and Xu (2021a)	C/C	MAA3C and LSTM	ANN	Unsupervised and supervised frameworks for ATFM
Xie et al. (2021)	C/C	DQN	ANN	1. Flow management for UAM airspace. 2. Using a DQN with genetic algorithm to solve DCB problem.
Tang and Xu (2021)	C/C	K-control and actor–critic	ANN	1. Rule-based time-step environment to mimic the DCB process. 2. MARL framework to address credit assignment problem.
Kravaris et al. (2019) and Spatharis et al. (2021)	C/C	Hierarchical RL	ϵ-greedy	1. Hierarchical approach partitions task into hierarchies of states and actions. 2. Hierarchical methods improve on DCB in the pre-tactical stage. 3. Agent-based MARL.
Duong et al. (2019)	C/C	Block-chain based RL	ϵ-greedy	Introduces decentralized a blockchain-based RL agent.
Chen et al. (2021)	C/C	DDQN, experience replay	Adaptive ϵ-greedy	1. Comparison with the actual method used in operations 2. Decentralized training with decentralized execution

Demand capacity balancing (DCB) is a predictive method to ensure the efficient operation of airspace or ground operations. A collaborative approach was introduced to DCB utilizing: assigning delays, allowing alternative trajectories, using fixed airspace sectorization, or adjusting airspace sectorization to efficiently manage airspace (Xu et al., 2020). Unlike other solutions, synchronized collaborative-demand capacity balancing (SC-DCB) seeks to relax the constraints of airspace configurations, with the outcome demonstrating a reduction in active sectors resulting in better utilization of the active ones. Further modeling the DCB problem as a POMDP, Huang and Xu (2021a) proposes a multi-agent asynchronous advantage actor–critic (MAA3C) network to manage delaying aircraft based on an unsupervised training approach. A combination of DCB for strategic conflict management and RL for tactical separation was proposed in Chen et al. (2024). A better tactical safety separation performance was achieved by using DCB to precondition traffic to proper density levels. Also, a recent study was shown the efficiency of RL methods by changing the traffic density in the training process (Groot et al., 2024).

By improving the computational capabilities, flights have been considered as agents, and MARL methods (Spatharis et al., 2018) were proposed to solve the capacity problems. Various algorithms were also studied: independent learners, edge-MARL, and agent-based-MARL, based on Q-learning techniques. Inspired by supervised learning, multiple supervised-MARL frameworks built on PPO were suggested (Tang and Xu, 2021), where the agents representing the flights have three actions: hold their departure, take-off, or cooperate. This study indicated that adding supervisors can help improve search and generalization abilities. DQN and decentralized training and decentralized execution (DTDE) combined with replay experience (Chen et al., 2021) were also used to solve the DCB problem.

In recent work (Xie et al., 2021), RL techniques have been utilized to examine their efficiency in UAM flow management, using a state space consisting of data retrieved from aircraft, weather, airspace capacity, and traffic density surveillance, and training data constructed through a Post-Hoc system. Multi-agent approaches in flow management also emerged (Duong et al., 2019; Tang and Xu, 2021; Spatharis et al., 2021) to demonstrate that a MARL approach can successfully resolve hot spots in dense traffic areas by taking holding, departure, or cooperation actions, resulting in an overall reduction in delay.

Ground delay programs (GDP) deal with an excessive number of flights reaching an airport serving as another air traffic flow management mechanism. Airports' ability to handle arrivals may be adversely affected by weather conditions. Hot spots were solved using GDP, in which flight departures are delayed, to shift the whole trajectory (Kravaris et al., 2019). The results show that collaborative methods yield better results. To reduce the search space, a hierarchical MARL scheme was proposed to solve the DCB problem with GDP (Spatharis et al., 2021), thus allowing the abstraction of time and state–action.

Issuing terminal traffic management initiatives (TMIs) is a technique for reducing the number of incoming aircraft to an airport for a short period. One type of this technique is the ground delay program. A data-driven approach based on a multi-armed bandit framework was proposed for suggesting TMI actions (Estes and Ball, 2017). This would be beneficial for human decision-makers to evaluate whether a suggested solution is reasonable or not. The suggestions were based on historical data of forecasted and observed demand and capacity, chosen TMI actions, and observed performance. The results showed that almost all proposed algorithms slightly outperform the historical actions. Jones et al. (2021) proposed four methods for recommending strategic TMI parameters during uncertain weather conditions. The first two methods were based on random exploration, while the others were using an ϵ-greedy approach and a Softmax algorithm. The fast-time simulation results demonstrated the strong performance of the two latter methods relative to the others, and their potential to help with dealing with weather uncertainty.

A comparison between behavioral cloning (BC) and inverse reinforcement learning (IRL) in predicting hourly expert GDP implementation actions was made in Bloem and Bambos (2015). Historical data was used to predict GDP decisions on San Francisco and Newark international airports. The IRL method was proposed to reduce the complexity by only exploring the states in the data. The results demonstrated that BC has a more robust predictive performance than the IRL GDP-implemented models. The experiments also suggested that neither the BC nor the IRL models predict the relatively infrequent GDP initialization or cancellation events well, unlike Q-learning, which tends to provide accurate predicted times (George and Khan, 2015). Better prediction of taxi-out times will improve taxiing management, which can benefit trajectory planning by using GDP to reduce congestion.

Runway Configuration Management (RCM) plays a crucial role in optimizing runway usage, considering factors such as traffic flow and weather conditions. This complex facet of air traffic management is challenging due to its reliance on fluctuating operational and environmental conditions. A Runway Configuration Assistance (RCA) decision-support tool was developed by employing offline model-free reinforcement learning (Memarzadeh et al., 2023; Nethi et al., 2024). An innovative integration of predictive data from the Localized Aviation Model Output Statistics Program (LAMP) and Terminal Area Forecast (TAF) were introduced, where significantly improving the tool's precision and its responsiveness to rapid changes in wind conditions.

With airspace becoming denser due to higher traffic and the introduction of emerging UAS/UTM technologies, traffic management solutions will be needed to demonstrate the ability of adaptation to accommodate not only higher volumes and densities of air traffic but

Table 3
Selection from the literature on RL in airline revenue management.

Reference	Problem type	Algorithm	Policy class	Key features
Gosavii et al. (2002)	Single leg	Q-learning	ϵ-greedy	Infinite time horizon under the average reward optimizing criterion.
Lawhead and Gosavi (2019)	Single leg	Bounded actor–critic	ϵ-greedy	Test two types of reward: discounted reward MDP and the average reward SMDP.
Bondoux et al. (2020)	Single leg	DQN	ϵ-greedy	Comparison between DQL and RMS
Shihab and Wei (2021)	Single leg	DQN	ϵ-greedy	Considering both cancellation and overbooking in the environment.
Wang et al. (2021a)	Single leg	DQN actor–critic	ϵ-greedy	Combining quantity-based RM and price-based RM together.
Alamdari and Savard (2021)	Single leg & Network	DQN	AGen	Greedily generate a set of "effective" actions to replace the original action space.

also any new requirements imposed by this new classification of air traffic. Additionally, the safety and capacity of these systems will require formal verification and standardized validation, moving the field of RL in ATM away from the laboratory and being ready to be accepted by official bodies. Finally, there are still many unknowns about how the UTM/UAS airspace will be constructed, adding a further layer of complexity to solution design; new systems should entertain this notion and provide flexibility while the airspace is still being defined.

3.5. Airline revenue management

In 1970s, there was limited control over ticket pricing and network scheduling. If one airline company wanted to increase its fare, permission from a federal agency called the Civil Aeronautics Board (CAB) was needed. The pricing regulation at that time always led to a higher fare. Airline deregulation happened in 1979, which allowed companies to schedule and price freely. Consequently, airline revenue management (ARM) came out as a business practice to set prices when there is perishable inventory. The ARM is an airline company's strategy to maximize revenue by optimizing ticket prices and product availability. The classic ARM problem could be divided into two types, quantity-based and price-based revenue management (RM) (Talluri et al., 2004) (see Table 3).

Quantity-based RM works on a predefined n-class fare structure and determines how many tickets are protected for each fare class. Also, it focuses on the capacity control of single and network flight legs. As a representative of the quantity-based RM, the expected marginal seat revenue (EMSR) models (Belobaba, 1987) are wildly used in the modern airline industry. The price-based RM focuses more on the dynamic pricing situation.

Traditional and widely used approaches for ARM systems are model-based and data-driven, which heavily depend on the accuracy of forecasting data such as passenger arrival distribution, willingness to pay (WTP), and cancellation rate. Recently, researchers have been considering applying model-free learning-based methods on ARM, such as optimal control theory or RL. A research direction of using RL in ARM started in 2002 (Gosavii et al., 2002), where the λ-smart algorithm was designed to cast the single-leg ARM problem as a semi-Markov decision problem (SMDP) over an infinite time horizon under the average reward optimizing criterion. Later, a bounded actor–critic approach was applied on the same problem (Lawhead and Gosavi, 2019). Both studies claimed that the model's performance was better than the EMSR model. A DRL model on ARM has been introduced to integrate the domain knowledge with a DNN trained on graphical processing units (GPUs) (Bondoux et al., 2020). A DRL model was also applied to the inventory control problem, using DQN and considering both cancellation and overbooking in their environment (Shihab and Wei, 2021). Some other improvements to DRL models have also appeared in recent years. For example, an ARM problem was studied by combining quantity-based RM, and price-based RM (Wang et al., 2021a), while the DRL was applied to both the single leg and network leg problems (Alamdari and Savard, 2021).

The previous learning-based approaches consider the game between passengers and airline companies. However, there is limited work regarding the competitive pricing process among different airline companies. We believe it will be an exciting topic with the development of MARL.

3.6. Aircraft flight and attitude control

Attitude control of an aircraft can be challenging due to the system's nonlinearities, uncertainties, and noises acting upon the system, which are intrinsically present in the environment. Recently, researchers have aimed to develop advanced controllers based on RL algorithms. A selection of RL methods in attitude control applications is presented in Table 4. This section focuses on controllers based on RL algorithms. A comprehensive survey on NN-based flight control systems was done in Emami et al. (2022).

These proposed controllers have been used in target tracking (Li et al., 2022, 2021), single/multi-agent obstacle avoidance (Li et al., 2021; Zhao et al., 2020), vision-based landing (Lee et al., 2018), stabilization (Xian et al., 2021; Zhen et al., 2020; Huang et al., 2019, 2020), visual servoing (Shi et al., 2016), and flat spin recovery (Kim et al., 2017). In Huang et al. (2019), it was shown that training a controller directly by RL, based on a nonlinear or unknown model, is feasible. The performance of the controllers based on different RL algorithms was also compared in Zuo et al. (2019). The results showed that a DQN is more suitable for discrete tasks than policy gradient or DDPG, whereas DDPG was shown to perform better in more complex tasks. Also, a DQN method was used to design attitude control systems for aircraft (Huang et al., 2019; Zuo et al., 2019). In addition, the DDPG-based controllers were established in Li et al. (2021), Zuo et al. (2019), Wang et al. (2020), Zhang et al. (2020) and Al-Gabalawy (2019). An improved DDPG method was combined with transfer learning and a control system was developed to perform autonomous maneuvering target tracking (Li et al., 2021). A DDPG-based controller was also studied, guiding a UAV to a fixed position in a horizontal plane from any position, and attitude (Zhang et al., 2020).

Other studies have been conducted using PPO methods (Zhao et al., 2020; Zhen et al., 2020; Bøhn et al., 2019). An improved MARL algorithm was developed, named multi-agent joint proximal policy optimization (MAJPPO), to perform formation and obstacle avoidance. The controller has used a moving averaging method to make each agent obtain a centralized state value function (Zhao et al., 2020). By performing the experimental comparison, it was shown that the MAJPPO algorithm could better deal with partially observable environments. A PPO-based controller was designed for stabilizing a fixed-wing UAV (Zhen et al., 2020). It was shown that the RL controller could stabilize the system in the presence of disturbances in the environment more precisely compared to a PID controller.

Since RL has achieved significant progress in attitude control, it has been considered a promising approach for designing optimal and robust controllers. However, there are still some challenges that should be addressed. The gap between simulations and natural environments was experimentally demonstrated (Wada et al., 2021), which required a new training approach. A controller learned to adapt to the difference between training models and real environments. Exploration and exploitation balance is another dilemma in RL. A normal distribution noise for exploring the environment was used at the start of the training process (Wang et al., 2020). It also proposed using Uhlenbeck–Ornstein stochastic noise for future works.

Table 4
Selection from literature on RL in attitude control.

Reference	S/A space	Algorithms	Policy class	Key features
Li et al. (2022)	C/C	Actor–critic	ANN	1. Compensating for the actuator fault and system input saturation. 2. Proving system stability by Lyapunov theory.
Li et al. (2021)	C/C	MMN-DDPG transfer learning	ANN	1. Introducing exploratory noises and parameter-based transfer learning to improve speed and generalization. 2. Performing target tracking and obstacle avoidance precisely in uncertain environments.
Zhao et al. (2020)	–	Multi-agent joint PPO (MAJPPO)	ANN	1. Using a moving window averaging of state-valued function to deal with multi-agent coordination problems. 2. MAJPPO, a centralized training and distributed execution.
Lee et al. (2018)	C/C	Actor–critic	ANN	Using a simple PID controller for handling attitude and position of UAV and a DRL algorithm to generate proper commands.
Xian et al. (2021)	C/C	Actor–critic	ANN	1. Compensating the error of actor–critic network by a robust nonlinear sliding mode control method. 2. Achieving a better control performance compared to LQR.
Zhen et al. (2020)	–	PPO	ANN	Achieving more precise control comparing to a PID controller in the presence of disturbance.
Huang et al. (2020)	C/C	Actor–critic	ANN	Introducing an NN approximation to learn the optimal controller online with no information of model.
Huang et al. (2019)	C/C	DDQN	ANN/ϵ – greedy	Proposing model can train the controller in time domain directly on nonlinear or unknown model.
Shi et al. (2016)	D/D	Q-learning	TD	Taking Q-learning for adaptive servoing gain adjustment.
Kim et al. (2017)	D/C	DQN	ANN	Covering both unusual attitude and stable spin mode recoveries.
Zuo et al. (2019)	C/C	DQN, PG, DDPG	ANN/ϵ-greedy	1. Being more efficient and faster. 2. Handling continuous action space but not efficient enough.
Wang et al. (2020)	C/C	DDPG	ANN	Using a normal distribution for having better exploration.
Bøhn et al. (2019)	C/C	PPO	ANN	1. Converging faster than PID. 2. Generalizing to turbulent wind conditions.
Wada et al. (2021)	C/C	actor–critic(A3C) LSTM	ANN	1. Stability of the NNs in different delays. 2. Experimentally demonstrating the reality and simulation gap.

3.7. Fault tolerant controller

A fault is a change in a system's property or parameters that causes the system to behave differently from its design. In other words, failure is a condition that prevents a system from functioning. A fault-tolerant controller (FTC) is a control strategy that aims to improve the performance of a system operating in degraded performance due to a fault (Blanke et al., 2006). FTCs are characterized as model-based or data-driven, based on the method used to develop the controllers. Model-based techniques necessitate knowledge of the system's model and parameters to design a fault-tolerant controller. On the contrary, data-driven approaches learn the FTC directly from system data. The fundamental problem of a model-based FTC approach is that its effectiveness depends on the system model's correctness, which is difficult to establish when system parameters can vary due to faults. Furthermore, complex systems necessitate complicated controllers, which, in turn, impacts the controllers' robustness. On the other hand, data-driven techniques utilize data to design FTC without knowing the system's dynamics. As a result, data-driven methods, particularly RL-based techniques, have recently gained a lot of attention.

Several approaches have been proposed in the literature to solve the FTC controller using RL. Different RL algorithms, including DDPG, TRPO, and PPO, have been used to develop FTC techniques for quadrotor attitude control (Koch et al., 2019). The results indicated that among the developed RL-based fault-tolerant controllers, the trained PPO-based attitude controller outperformed a fully tuned PID controller in terms of rising time, peak velocities achieved, and total error among the trained set of controllers. A DPG-based technique with an integral compensator was adopted to develop a position-tracking controller for the quadrotor (Wang et al., 2019b). The approach employed a two-phased learning scheme, with a simplified model being utilized for offline learning and the learned policy being refined during flight. The results showed that the learned FTC is sufficiently robust to model errors and external disturbances. A DDPG-based fault-tolerant policy for position tracking of quadcopters was proposed in Fei et al. (2020). The framework operates so that it runs simultaneously with the model-based controller and only becomes active when the system's behavior changes from the normal operating condition.

One of the significant drawbacks of model-free RL-based FTC methods is that there is no guarantee of convergence. To overcome this problem, a model-based framework for position tracking of octocopters was proposed (Bhan et al., 2021). Four RL algorithms were proposed, PPO, DDPG, Twin-Delayed DDPG (TD3), and soft actor–critic (SAC). The results showed that PPO is more suitable for a fault-tolerant task (see Table 5).

3.8. Aircraft flight planning

Flight and trajectory planning is a well-known aviation problem and is crucial. While airspace users want the most optimal trajectory to minimize a cost function, many constraints, such as ground obstacles, capacity limitations, or environmental threats, make this problem difficult to solve. Several techniques, including rerouting or ground delay are proposed to mitigate traffic congestion in most cases. The ATM domain is essentially based on temporal operations, with a capacity supply and demand model to manage air traffic flows. This operation can lead to capacity imbalances and create hot spots in sectors when capacity (defined as the number of aircraft accepted in a given sector during a given period) is exceeded. The planning of a trajectory or flight of an aircraft can be done in several stages defined in the ATM domain; the strategic phase includes the planning of flights performed between one year and D-7, the pre-tactical phase takes place between D-7 and D-1, and finally, the tactical phase takes place on D-day. An RL planner has shown to be a promising tool to solve pre-flight planning problems in dangerous environments (Wickman, 2021).

UAV's versatility in performing tasks ranging from terrain mapping to surveillance and military missions makes this problem a fundamental part of aircraft operations. One of many defined missions for UAVs is to fly over ground targets. The theory of POMDP was presented for military use, and nominal belief-state optimization (NBO) was used to

Table 5
Selection from the literature of RL on fault-tolerant controller.

Reference	Problem type	S/A space	Algorithm	Key features
Koch et al. (2019)	Attitude control	C/C	DDPG, TRPO, PPO	Training RL algorithms to perform end-to-end attitude control.
Wang et al. (2019b)	Position tracking	C/C	DPG	Integrating DPG with integral compensator and adopting a two-phased approach.
Fei et al. (2020)	Position tracking	C/C	DDPG	Running simultaneously with the model-based controller.
Bhan et al. (2021)	Position tracking	C/C	DDPG, TD3 SAC, PPO	1. Estimating fault-related parameters using an estimator. 2. Training several RL algorithms using the estimated parameters.

Table 6
Selection from the literature on RL in flight planning.

Reference	Problem type	Algorithms	Policy class	Key features
Ragi and Chong (2013)	UAV path planning	POMDPNBO approximation	–	1. Dynamical environment. 2. Wind effects are taken into account.
Zhang et al. (2015)	UAV path planning	Geometric RL	–	Convergence of calculating the reward matrix theoretically proven.
Yan et al. (2020)	UAV path planning	D3QN, DDQN, DQN	ϵ-greedy	1. Stage scenario for simulation. 2. DRL approach and comparison of methods.
Bertram et al. (2021)	Flight plan scheduling	FastMDP	ϵ-greedy	1. Centralized or distributed flight plan scheduling. 2. Parallelization for large-scale scheduling.

find the optimal trajectory considering threats, wind effects, or other agents (Ragi and Chong, 2013). Also, an RL approach was proposed to use geometric information from the drone's environment and produce smoother and more feasible trajectories in real-time planning (Zhang et al., 2015). The dueling double deep Q-networks (D3QN), DDQN, and DQN methods have been compared in Yan et al. (2020) to solve the path planning problem for an agent in the context of a dynamic environment where it faces an environmental threat.

An RL method was used to resolve these hot spots with traffic speed regulation (Tumer and Agogino, 2007). An agent representing a fix (a 2D point in the sector) can regulate the flows. In addition, a multi-agent asynchronous advantage actor–critic (MAA3C) framework was constructed to resolve airspace hot spots within a proper ground delay (Huang and Xu, 2021b) (see Table 6).

All these works aim to reduce hot spots by delaying flights while minimizing average delays and ensuring good distribution. Still, they have not studied other trajectory planning techniques. An RL approach was proposed to select a low-level heuristic to mitigate the air traffic complexity (Juntama et al., 2022). Flight level allocation, staggered departure times, and en route path deviation reduced congestion. In a UAM concept, the pre-departure airspace reservation problem as an MDP was formulated (Bertram et al., 2021). The first-in-first-out (FIFO) principle and the fast-MDP algorithm provided a conflict-free trajectory at the strategic stage. The scheduler, allowing both centralized and decentralized flight planning, takes advantage of the computing power and parallelization of GPUs to process a large number of flights. A Learning-to-Dispatch algorithm was proposed to maximize the air capacity under emergencies such as hurricane disasters (Zhang et al., 2021b).

3.9. Airline maintenance

Maintenance scheduling is the process of planning when and what type of maintenance check should be performed on an aircraft. The maintenance tasks of airlines are usually grouped into four-letter checks (A, B, C, and D). The level of detail in the maintenance check of these groups is different. For example, A- and B-checks are considered light maintenance, and C- and D-check as heavy maintenance and more detailed inspection. Usually, weather conditions and flight disruptions cause deviation from the scheduled plan. These uncertainties make aircraft maintenance scheduling a challenging task.

A look-ahead approximate dynamic programming methodology was developed for aircraft maintenance check (Deng and Santos, 2022). Its schedules minimized the wasted utilization interval between maintenance checks while reducing the need for additional maintenance slots. The methodology was tested with two case studies of maintenance data of an A320 family fleet. The developed method showed significant changes in scheduled maintenance times; it reduced the number of A-checks by 1.9%, the number of C-check by 9.8%, and the number of additional slots by 78.3% over four years.

An RL-based approach was proposed in Hu et al. (2021) to solve the aircraft's long-term maintenance optimization problem. The proposed method uses information about the aircraft's future mission, repair cost, prognostics and health management, etc., to provide real-time, sequential maintenance decisions. The RL-driven approach outperforms three existing commonly used strategies in adjusting its decision principle based on the diverse data in several simulated maintenance scenarios. The integration of an RL model for Human–AI collaboration in maintenance planning and the visualization of the Condition-Based Maintenance indicators were proposed in Ribeiro et al. (2022). Optimal maintenance decision-making in the presence of unexpected events was also developed.

3.10. Safety and certification of reinforcement learning

Safety is of utmost importance in safety–critical applications such as aviation systems. Recent promising results in RL have encouraged researchers to apply such techniques to many real-world applications. However, the certification of learning-based approaches, including RL in safety–critical applications, remains an open research question (Van Wesel and Goodloe, 2017; Baheri et al., 2022). Recent surveys provide a comprehensive overview of efforts toward safe RL for safety–critical applications (Garcıa and Fernández, 2015). While there has been a lot of research interest in safe RL, especially in the autonomous driving community (Kiran et al., 2021; Baheri et al., 2020; Baheri, 2022), the safe RL problem is still underexplored in the aviation research community. The application of safe RL in aviation systems has been studied from different angles. For instance, recently, a safe DRL approach was proposed for autonomous airborne collision avoidance systems (Panoutsakopoulos et al., 2022). From the conflict resolution perspective, soft actor–critic models were used during vertical maneuvers in layered airspace (Groot, 2021). In a similar line of research, a safe deep MARL framework can identify and resolve conflicts between aircraft in a high-density (Brittain and Wei, 2019).

From the run-time assurance perspective, a run-time safety assurance approach casts the problem as an MDP framework and uses RL to solve it (Lazarus et al., 2020). Similarly, the path planning problem was framed as MDP and utilized MCTS for safe and assured path planning (Wu and Chen, 2022). To guarantee the safety of real-time autonomous flight operations, an MCTS algorithm was proposed

along with Gaussian process regression and Bayesian optimization to discretize the continuous action space (Wu et al., 2022). Furthermore, a reinforcement learning framework predicts and mitigates the potential loss of separation events in congested airspace (Hawley et al., 2019). Recently, a safety verification framework was presented for design-time and run-time assurance of learning-based components in aviation systems (Baheri et al., 2022).

4. Conclusion

In this paper, after a review of the most common RL techniques and the overall methodology and principles, a survey of the applications of RL in aviation is proposed. Ranging from airline revenue management to aircraft altitude control, the use of RL methods has shown a great interest in the literature in the last decade. Indeed, with the increase in computational power and access to a large source of data, this data-driven approach has become widely studied. Whether it is collision avoidance, traffic management, or other aviation-related problems, these learning-based frameworks show promising results and a variety of algorithms and techniques are often studied for a specific problem. The most advanced techniques such as DRL or DPG are used to deal with critical systems such as collision avoidance or to handle the increase of growing air traffic in traffic management and flight planning. However, differences between the simulated environment and real-world application or its black-box scheme can still be a hindrance to implementation in the aviation industry, constrained by numerous safety measures. The certification of such methods is then a crucial point for these innovative and disruptive applications in aviation and should be one of the focuses of research in this area.

CRediT authorship contribution statement

Pouria Razzaghi: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Conceptualization. **Amin Tabrizian:** Writing – review & editing, Writing – original draft, Investigation. **Wei Guo:** Writing – original draft. **Shulu Chen:** Writing – original draft. **Abenezer Taye:** Writing – original draft. **Ellis Thompson:** Writing – original draft. **Alexis Bregeon:** Writing – original draft. **Ali Baheri:** Writing – original draft, Supervision, Conceptualization. **Peng Wei:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pouria Razzaghi reports financial support was provided by US Department of Transportation Office of Aviation Analysis.

Data availability

No data was used for the research described in the article.

Acknowledgments

The paper is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this paper. The findings and conclusions are those of the authors and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. This research was supported by the FAA, United States under contract No. 692M15-21-T-00022.

References

Ahmed, M.S., Alam, S., Barlow, M., 2018. A cooperative co-evolutionary optimisation model for best-fit aircraft sequence and feasible runway configuration in a multi-runway airport. Aerospace 5 (3), 85.

Al-Gabalawy, M., 2019. Machine learning for aircraft control. J. Adv. Res. Dyn. Control Syst. 11, 3165–3191.

Alamdari, N.E., Savard, G., 2021. Deep reinforcement learning in seat inventory control problem: an action generation approach. J. Revenue Pricing Manag. 20 (5), 566–579.

Alligier, R., Gianazza, D., Durand, N., 2015. Machine learning and mass estimation methods for ground-based aircraft climb prediction. IEEE Trans. Intell. Transp. Syst. 16 (6), 3138–3149.

Alvarez, L.E., Brittain, M., Breeden, K., 2023. Towards a standardized reinforcement learning framework for AAM contingency management. In: 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference. DASC, IEEE, pp. 1–6.

Ayhan, S., Samet, H., 2016. Aircraft trajectory prediction made easy with predictive analytics. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 21–30.

Baheri, A., 2022. Safe reinforcement learning with mixture density network, with application to autonomous driving. Results Control Optim. 6, 100095.

Baheri, A., Nageshrao, S., Tseng, H.E., Kolmanovsky, I., Girard, A., Filev, D., 2020. Deep reinforcement learning with enhanced safety for autonomous highway driving. In: 2020 IEEE Intelligent Vehicles Symposium. IV, IEEE, pp. 1550–1555.

Baheri, A., Ren, H., Johnson, B., Razzaghi, P., Wei, P., 2022. A verification framework for certifying learning-based safety-critical aviation systems. In: AIAA AVIATION 2022 Forum. http://dx.doi.org/10.2514/6.2022-3965.

Belobaba, P., 1987. Air Travel Demand and Airline Seat Inventory Management (Ph.D. thesis). Massachusetts Institute of Technology.

Berndt, J., 2004. JSBSim: An open source flight dynamics model in C++. In: AIAA Modeling and Simulation Technologies Conference and Exhibit. p. 4923.

Bertram, J., Wei, P., 2020. Distributed computational guidance for high-density urban air mobility with cooperative and non-cooperative collision avoidance. p. 1371,

Bertram, J., Wei, P., Zambreno, J., 2021. Scalable FastMDP for pre-departure airspace reservation and strategic de-conflict. In: AIAA Scitech 2021 Forum. p. 0779.

Bertsekas, D., 2012. Dynamic Programming and Optimal Control: Volume I, vol. 1, Athena scientific.

Bhan, L., Quinones-Grueiro, M., Biswas, G., 2021. Fault tolerant control combining reinforcement learning and model-based control. In: 2021 5th International Conference on Control and Fault-Tolerant Systems. SysTol, IEEE, pp. 31–36.

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., Schröder, J., 2006. Diagnosis and Fault-Tolerant Control, vol. 2, Springer.

Bloem, M., Bambos, N., 2015. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. J. Aerosp. Inf. Syst. 12 (3), 299–313.

Bøhn, E., Coates, E.M., Moe, S., Johansen, T.A., 2019. Deep reinforcement learning attitude control of fixed-wing uavs using proximal policy optimization. In: 2019 International Conference on Unmanned Aircraft Systems. ICUAS, IEEE, pp. 523–533.

Bondoux, N., Nguyen, A.Q., Fiig, T., Acuna-Agost, R., 2020. Reinforcement learning applied to airline revenue management. J. Revenue Pricing Manag. 19 (5), 332–348.

Brittain, M.W., Alvarez, L.E., Breeden, K., 2024. Improving autonomous separation assurance through distributed reinforcement learning with attention networks. Proc. AAAI Conf. Artif. Intell. 38 (21), 22857–22863.

Brittain, M.W., Wei, P., 2018. Towards autonomous air trac control for sequencing and separation-a deep reinforcement learning approach. In: 2018 Aviation Technology, Integration, and Operations Conference. p. 3664.

Brittain, M., Wei, P., 2019. Autonomous separation assurance in an high-density en route sector: A deep multi-agent reinforcement learning approach. In: 2019 IEEE Intelligent Transportation Systems Conference. ITSC, pp. 3256–3262.

Brittain, M.W., Wei, P., 2021. One to any: Distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory. In: AIAA Scitech 2021 Forum. p. 1952.

Brittain, M., Yang, X., Wei, P., 2021. A deep multi-agent reinforcement learning approach to autonomous separation assurance. AIAA J. Aerosp. Inf. Syst. 18 (12).

Chaslot, G., Bakkes, S., Szita, I., Spronck, P., 2021. Monte-Carlo tree search: A new framework for game ai. In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Vol. 4, pp. 216–217, (1).

Chen, S., Evans, A.D., Brittain, M., Wei, P., 2024. Integrated conflict management for UAM with strategic demand capacity balancing and learning-based tactical deconfliction. IEEE Trans. Intell. Transp. Syst..

Chen, Y., Xu, Y., Hu, M., Yang, L., 2021. Demand and capacity balancing technology based on multi-agent reinforcement learning. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–9.

Choi, S., Kim, Y.J., Briceno, S., Mavris, D., 2016. Prediction of weather-induced airline delays based on machine learning algorithms. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–6.

Conde Rocha Murca, M., DeLaura, R., Hansman, R.J., Jordan, R., Reynolds, T., Balakrishnan, H., 2016. Trajectory clustering and classification for characterization of air traffic flows. In: 16th AIAA Aviation Technology, Integration, and Operations Conference. p. 3760.

Cruciol, L.L., de Arruda, Jr., A.C., Weigang, L., Li, L., Crespo, A.M., 2013. Reward functions for learning to control in air traffic flow management. Transp. Res. C 35, 141–155.

Dalmau, R., Allard, E., 2020. Air traffic control using message passing neural networks and multi-agent reinforcement learning. In: Proceedings of the 10th SESAR Innovation Days, Virtual Event. pp. 7–10.

Deng, Q., Santos, B.F., 2022. Lookahead approximate dynamic programming for stochastic aircraft maintenance check scheduling optimization. European J. Oper. Res. 299 (3), 814–833.

Deniz, S., Wu, Y., Shi, Y., Wang, Z., 2024. A reinforcement learning approach to vehicle coordination for structured advanced air mobility. In: Green Energy and Intelligent Transportation. Elsevier, 100157.

Duan, J., Guan, Y., Li, S.E., Ren, Y., Sun, Q., Cheng, B., 2021. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. IEEE Trans. Neural Netw. Learn. Syst..

Duong, T., Todi, K.K., Chaudhary, U., Truong, H.-L., 2019. Decentralizing air traffic flow management with blockchain-based reinforcement learning. In: 2019 IEEE 17th International Conference on Industrial Informatics. INDIN, Vol. 1, IEEE, pp. 1795–1800.

EKim, d., Bak, S., 2019. ACASXu closed loop simulation falsification benchmark. URL https://github.com/stanleybak/acasxu_closed_loop_sim.

Emami, S.A., Castaldi, P., Banazadeh, A., 2022. Neural network-based flight control systems: Present and future. Annu. Rev. Control 53, 97–137. http://dx.doi.org/10.1016/j.arcontrol.2022.04.006.

Estes, A., Ball, M., 2017. Data-driven planning for ground delay programs. Transp. Res. Rec. 2603 (1), 13–20.

EUROCONTROL, 2022. ADemand data repository. URL https://www.eurocontrol.int/ddr.

Fei, F., Tu, Z., Xu, D., Deng, X., 2020. Learn-to-recover: Retrofitting uavs with reinforcement learning-assisted flight control under cyber-physical attacks. In: 2020 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 7358–7364.

Feinberg, V., Wan, A., Stoica, I., Jordan, M.I., Gonzalez, J.E., Levine, S., 2018. Model-based value estimation for efficient model-free reinforcement learning. arXiv preprint arXiv:1803.00101.

FLIGHTGEAR FLIGHT SIMULATOR sophisticated, professional, open-source. URL https://www.flightgear.org/, 2023.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S., 2018. Counterfactual multi-agent policy gradients. Proc. AAAI Conf. Artif. Intell. 32 (1).

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Garcıa, J., Fernández, F., 2015. A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. 16 (1), 1437–1480.

Garcia, C.P., Weigang, L., Hirata, N.S., Neumann, C., 2023. ISUAM: Intelligent and safe UAM with deep reinforcement learning. In: 2023 IEEE 29th International Conference on Parallel and Distributed Systems. ICPADS, IEEE, pp. 378–383.

George, E., Khan, S.S., 2015. Reinforcement learning for taxi-out time prediction: An improved Q-learning approach. In: 2015 International Conference on Computing and Network Communications. CoCoNet, IEEE, pp. 757–764.

Gosavii, A., Bandla, N., Das, T.K., 2002. A reinforcement learning approach to a single leg airline revenue management problem with multiple fare classes and overbooking. IIE Trans. 34 (9), 729–742.

Gronauer, S., Diepold, K., 2022. Multi-agent deep reinforcement learning: a survey. Artif. Intell. Rev. 55 (2), 895–943.

Groot, J., 2021. Improving Safety of Vertical Manoeuvres in a Layered Airspace with Deep Reinforcement Learning (Master's thesis). Delft University of Technology, Aerospace Engineering.

Groot, D., Ellerbroek, J., Hoekstra, J., 2024. Analysis of the impact of traffic density on training of reinforcement learning based conflict resolution methods for drones. Eng. Appl. Artif. Intell. 133, 108066.

Gu, S., Lillicrap, T., Sutskever, I., Levine, S., 2016. Continuous deep q-learning with model-based acceleration. In: International Conference on Machine Learning. PMLR, pp. 2829–2838.

Guo, W., Brittain, M., Wei, P., 2021. Safety enhancement for deep reinforcement learning in autonomous separation assurance. In: 2021 IEEE International Intelligent Transportation Systems Conference. ITSC, IEEE, pp. 348–354.

Ha, D., Schmidhuber, J., 2018. World models. arXiv preprint arXiv:1803.10122.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. PMLR, pp. 1861–1870.

Harman, W.H., 1989. TCAS- A system for preventing midair collisions. Lincoln Lab. J. 2 (3), 437–457.

Hawley, M., Bharadwaj, R., Venkataraman, V., 2019. Real-time mitigation of loss of separation events using reinforcement learning. In: 2019 IEEE/AIAA 38th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–6.

Herman, M., 2021. Towards Explainable Automation for Air Traffic Control Using Deep Q-learning from Demonstrations and Reward Decomposition (Master's thesis). Delft University of Technology, Aerospace Engineering.

Hoekstra, J.M., Ellerbroek, J., 2016. Bluesky ATC simulator project: an open data and open source approach. In: Proceedings of the 7th International Conference on Research in Air Transportation. Vol. 131, FAA/Eurocontrol USA/Europe, p. 132.

Hu, J., Liu, Y., 2020. UAS conflict resolution integrating a risk-based operational safety bound as airspace reservation with reinforcement learning. In: AIAA Scitech 2020 Forum. p. 1372.

Hu, Y., Miao, X., Zhang, J., Liu, J., Pan, E., 2021. Reinforcement learning-driven maintenance strategy: A novel solution for long-term aircraft maintenance decision optimization. Comput. Ind. Eng. 153, 107056.

Hu, J., Yang, X., Wang, W., Wei, P., Ying, L., Liu, Y., 2022. Obstacle avoidance for uas in continuous action space using deep reinforcement learning. IEEE Access 10, 90623–90634.

Huang, D., Hu, J., Peng, Z., Chen, B., Hao, M., Ghosh, B.K., 2020. Model-free based reinforcement learning control strategy of aircraft attitude systems. In: 2020 Chinese Automation Congress. CAC, IEEE, pp. 743–748.

Huang, X., Luo, W., Liu, J., 2019. Attitude control of fixed-wing UAV based on DDQN. In: 2019 Chinese Automation Congress. CAC, IEEE, pp. 4722–4726.

Huang, C., Xu, Y., 2021a. Integrated frameworks of unsupervised, supervised and reinforcement learning for solving air traffic flow management problem. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, pp. 1–10. http://dx.doi.org/10.1109/DASC52595.2021.9594397.

Huang, C., Xu, Y., 2021b. Integrated frameworks of unsupervised, supervised and reinforcement learning for solving air traffic flow management problem. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–10.

Isufaj, R., Aranega Sebastia, D., Piera, M.A., 2021. Towards conflict resolution with deep multi-agent reinforcement learning. In: Proceedings of the 14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021), New Orleans, LA, USA. pp. 20–24.

Isufaj, R., Omeri, M., Piera, M.A., 2022. Multi-UAV conflict resolution with graph convolutional reinforcement learning. Appl. Sci. 12 (2), 610.

Jacob, B., Kaushik, A., Velavan, P., Sharma, M., 2022. Autonomous drones for medical assistance using reinforcement learning. In: Advances in Augmented Reality and Virtual Reality. Springer, pp. 133–156.

Jeannin, J.-B., Ghorbal, K., Kouskoulas, Y., Gardner, R., Schmidt, A., Zawadzki, E., Platzer, A., 2015. Formal verification of ACAS X, an industrial airborne collision avoidance system. In: 2015 International Conference on Embedded Software. EMSOFT, IEEE, pp. 127–136.

Jones, J., Ellenbogen, Z., Glina, Y., 2021. Recommending strategic air traffic management initiatives in convective weather. In: Fourteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2021), Lexington, MA 02421, USA.

Julian, K.D., Lopez, J., Brush, J.S., Owen, M.P., Kochenderfer, M.J., 2016. Policy compression for aircraft collision avoidance systems. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–10.

Juntama, P., Delahaye, D., Chaimatanan, S., Alam, S., 2022. Hyperheuristic approach based on reinforcement learning for air traffic complexity mitigation. In: AIAA Journal of Aerospace Information Systems. American Institute of Aeronautics and Astronautics, pp. 1–16.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–3154.

Kim, D., Oh, G., Seo, Y., Kim, Y., 2017. Reinforcement learning-based optimal flat spin recovery for unmanned aerial vehicle. J. Guid. Control Dyn. 40 (4), 1076–1084.

Kiran, B.R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A.A., Yogamani, S., Pérez, P., 2021. Deep reinforcement learning for autonomous driving: A survey. IEEE Trans. Intell. Transp. Syst..

Koch, W., Mancuso, R., West, R., Bestavros, A., 2019. Reinforcement learning for UAV attitude control. ACM Trans. Cyber-Phys. Syst. 3 (2), 1–21.

Kochenderfer, M.J., Holland, J.E., Chryssanthacopoulos, J.P., 2012. Next-Generation Airborne Collision Avoidance System. Tech. Rep., Massachusetts Institute of Technology-Lincoln Laboratory Lexington United States.

Kravaris, T., Spatharis, C., Bastas, A., Vouros, G.A., Blekas, K., Andrienko, G., Andrienko, N., Garcia, J.M.C., 2019. Resolving congestions in the air traffic management domain via multiagent reinforcement learning methods. arXiv preprint arXiv:1912.06860.

Lai, J., Cai, K., Liu, Z., Yang, Y., 2021. A multi-agent reinforcement learning approach for conflict resolution in dense traffic scenarios. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–9.

Lawhead, R.J., Gosavi, A., 2019. A bounded actor–critic reinforcement learning algorithm applied to airline revenue management. Eng. Appl. Artif. Intell. 82, 252–262.

Lazarus, C., Lopez, J.G., Kochenderfer, M.J., 2020. Runtime safety assurance using reinforcement learning. In: 2020 AIAA/IEEE 39th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–9.

Lee, H., Malik, W., Jung, Y.C., 2016. Taxi-out time prediction for departures at charlotte airport using machine learning techniques. In: 16th AIAA Aviation Technology, Integration, and Operations Conference. p. 3910.

Lee, S., Shim, T., Kim, S., Park, J., Hong, K., Bang, H., 2018. Vision-based autonomous landing of a multi-copter unmanned aerial vehicle using reinforcement learning. In: 2018 International Conference on Unmanned Aircraft Systems. ICUAS, IEEE, pp. 108–114.

Li, Z., Chen, X., Xie, M., Zhao, Z., 2022. Adaptive fault-tolerant tracking control of flying-wing unmanned aerial vehicle with system input saturation and state constraints. Trans. Inst. Meas. Control 44 (4), 880–891.

Li, S., Egorov, M., Kochenderfer, M., 2019a. Optimizing collision avoidance in dense airspace using deep reinforcement learning. In: Thirteenth USA/Europe Air Traffic Management Research and Development Seminar. ATM.

Li, C., Liu, J., Zhang, Y., Wei, Y., Niu, Y., Yang, Y., Liu, Y., Ouyang, W., 2023. Ace: Cooperative multi-agent q-learning with bidirectional action-dependency. Proc. AAAI Conf. Artif. Intell. 37 (7), 8536–8544.

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., Russell, S., 2019b. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. Proc. AAAI Conf. Artif. Intell. 33 (1), 4213–4220.

Li, B., Yang, Z.-p., Chen, D.-q., Liang, S.-y., Ma, H., 2021. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning. Defence Technol. 17 (2), 457–466.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2016. Continuous control with deep reinforcement learning. In: ICRA (Poster).

Lin, L.-J., 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach. Learn. 8 (3), 293–321.

Matignon, L., Laurent, G.J., Le Fort-Piat, N., 2007. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 64–69.

Memarzadeh, M., Puranik, T.G., Kalyanam, K.M., Ryan, W., 2023. Airport runway configuration management with offline model-free reinforcement learning. In: AIAA SciTech 2023 Forum. p. 0504.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533.

Mollinga, J., van Hoof, H., 2020. An autonomous free airspace en-route controller using deep reinforcement learning techniques. In: International Conference for Research in Air Transportation. ICRAT.

Nagabandi, A., Kahn, G., Fearing, R.S., Levine, S., 2018. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In: 2018 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 7559–7566.

Nethi, S., Memarzadeh, M., Kalyanam, K., 2024. Optimization of runway configurations with forecast-augmented offline reinforcement learning. In: AIAA SCITECH 2024 Forum. p. 0533.

Omidshafiei, S., Pazis, J., Amato, C., How, J.P., Vian, J., 2017. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In: International Conference on Machine Learning. PMLR, pp. 2681–2690.

Panoutsakopoulos, C., Yuksek, B., Inalhan, G., Tsourdos, A., 2022. Towards safe deep reinforcement learning for autonomous airborne collision avoidance systems. In: AIAA SCITECH 2022 Forum. p. 2102.

Pham, D.-T., Tran, N.P., Alam, S., Duong, V., Delahaye, D., 2019a. A machine learning approach for conflict resolution in dense traffic scenarios with uncertainties. In: ATM Seminar 2019, 13th USA/Europe ATM R&D Seminar.

Pham, D.-T., Tran, N.P., Goh, S.K., Alam, S., Duong, V., 2019b. Reinforcement learning for two-aircraft conflict resolution in the presence of uncertainty. In: 2019 IEEE-RIVF International Conference on Computing and Communication Technologies. RIVF, IEEE, pp. 1–6.

Qu, G., Wierman, A., 2020. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In: Conference on Learning Theory. PMLR, pp. 3185–3205.

Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Jimenez Rezende, D., Puigdomènech Badia, A., Vinyals, O., Heess, N., Li, Y., et al., 2017. Imagination-augmented agents for deep reinforcement learning. Adv. Neural Inf. Process. Syst. 30.

Ragi, S., Chong, E.K., 2013. UAV path planning in a dynamic environment via partially observable Markov decision process. IEEE Trans. Aerosp. Electron. Syst. 49 (4), 2397–2412.

Rashid, T., Farquhar, G., Peng, B., Whiteson, S., 2020. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. Adv. Neural Inf. Process. Syst. 33, 10199–10210.

Ribeiro, J., Andrade, P., Carvalho, M., Silva, C., Ribeiro, B., Roque, L., 2022. Playful probes for design interaction with machine learning: A tool for aircraft condition-based maintenance planning and visualisation. Mathematics 10 (9), 1604.

Ribeiro, M., Ellerbroek, J., Hoekstra, J., 2020. Determining optimal conflict avoidance manoeuvres at high densities with reinforcement learning. In: Proceedings of the Tenth SESAR Innovation Days, Virtual Conference. pp. 7–10.

Rummery, G.A., Niranjan, M., 1994. On-Line Q-Learning Using Connectionist Systems, vol. 37, Citeseer.

Schmidt, M., 2017. A review of aircraft turnaround operations and simulations. Prog. Aerosp. Sci. 92, 25–38.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al., 2020. Mastering atari, go, chess and shogi by planning with a learned model. Nature 588 (7839), 604–609.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P., 2015. Trust region policy optimization. In: International Conference on Machine Learning. PMLR, pp. 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Shi, H., Li, X., Hwang, K.-S., Pan, W., Xu, G., 2016. Decoupled visual servoing with fuzzy Q-learning. IEEE Trans. Ind. Inform. 14 (1), 241–252.

Shihab, S.A., Wei, P., 2021. A deep reinforcement learning approach to seat inventory control for airline revenue management. J. Revenue Pricing Manag. 1–17.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M., 2014. Deterministic policy gradient algorithms. In: International Conference on Machine Learning. PMLR, pp. 387–395.

Singh, A.J., Kumar, A., Lau, H.C., 2021. Approximate difference rewards for scalable multigent reinforcement learning. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1655–1657.

Spatharis, C., Bastas, A., Kravaris, T., Blekas, K., Vouros, G.A., Cordero, J.M., 2021. Hierarchical multiagent reinforcement learning schemes for air traffic management. Neural Comput. Appl. 1–13.

Spatharis, C., Kravaris, T., Vouros, G.A., Blekas, K., Chalkiadakis, G., Garcia, J.M.C., Fernandez, E.C., 2018. Multiagent reinforcement learning methods to resolve demand capacity balance problems. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. pp. 1–9.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K., et al., 2017. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296.

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. MIT Press.

Sutton, R.S., McAllester, D., Singh, S., Mansour, Y., 1999. Policy gradient methods for reinforcement learning with function approximation. Adv. Neural Inf. Process. Syst. 12.

Takeichi, N., Kaida, R., Shimomura, A., Yamauchi, T., 2017. Prediction of delay due to air traffic control by machine learning. In: AIAA Modeling and Simulation Technologies Conference. p. 1323.

Talluri, K.T., Van Ryzin, G., Van Ryzin, G., 2004. The Theory and Practice of Revenue Management, vol. 1, Springer.

Tan, M., 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the Tenth International Conference on Machine Learning. pp. 330–337.

Tang, Y., Xu, Y., 2021. Multi-agent deep reinforcement learning for solving large-scale air traffic flow management problem: A time-step sequential decision approach. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–10.

Tran, N.P., Pham, D.-T., Goh, S.K., Alam, S., Duong, V., 2019. An intelligent interactive conflict solver incorporating air traffic controllers' preferences using reinforcement learning. In: 2019 Integrated Communications, Navigation and Surveillance Conference. ICNS, IEEE, pp. 1–8.

Tran, P.N., Pham, D.-T., Goh, S.K., Alam, S., Duong, V., 2020. An interactive conflict solver for learning air traffic conflict resolutions. J. Aerosp. Inf. Syst. 17 (6), 271–277.

Tsitsiklis, J.N., 1994. Asynchronous stochastic approximation and Q-learning. Mach. Learn. 16 (3), 185–202.

Tumer, K., Agogino, A., 2007. Distributed agent-based air traffic flow management. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 1–8.

US Department of Transportation, 2024. Bureau of transportation statistics. URL https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp.

Van Hasselt, H., Guez, A., Silver, D., 2016. Deep reinforcement learning with double q-learning. Proc. AAAI Conf. Artif. Intell. 30 (1).

Van Wesel, P., Goodloe, A.E., 2017. Challenges in the Verification of Reinforcement Learning Algorithms. Tech. Rep., NASA.

Wada, D., Araujo-Estrada, S.A., Windsor, S., 2021. Unmanned aerial vehicle pitch control under delay using deep reinforcement learning with continuous action in wind tunnel test. Aerospace 8 (9), 258.

Wang, R., Gan, X., Li, Q., Yan, X., 2021a. Solving a joint pricing and inventory control problem for perishables via deep reinforcement learning. Complexity 2021.

Wang, J.X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D., Botvinick, M., 2016a. Learning to reinforcement learn. arXiv preprint arXiv:1611.05763.

Wang, Z., Li, H., Wang, J., Shen, F., 2019a. Deep reinforcement learning based conflict detection and resolution in air traffic control. IET Intell. Transp. Syst. 13 (6), 1041–1047.

Wang, W., Liu, Y., Srikant, R., Ying, L., 2021b. 3M-RL: Multi-resolution, multi-agent, mean-field reinforcement learning for autonomous UAV routing. IEEE Trans. Intell. Transp. Syst..

Wang, Z., Luo, W., Gong, Q., Cui, Y., Tao, R., Wang, Q., Liang, Q., Wang, S., 2020. Attitude controller design based on deep reinforcement learning for low-cost aircraft. In: 2020 Chinese Automation Congress. CAC, IEEE, pp. 463–467.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., Freitas, N., 2016b. Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 1995–2003.

Wang, Y., Sun, J., He, H., Sun, C., 2019b. Deterministic policy gradient with integral compensator for robust quadrotor control. IEEE Trans. Syst. Man Cybern. Syst. 50 (10), 3713–3725.
Watkins, C.J.C.H., 1989. Learning from Delayed Rewards (Ph.D. thesis). Cambridge United Kingdom, King's College.
Watkins, C.J., Dayan, P., 1992. Q-learning. Mach. Learn. 8 (3), 279–292.
Wen, H., Li, H., Wang, Z., Hou, X., He, K., 2019. Application of DDPG-based collision avoidance algorithm in air traffic control. In: 2019 12th International Symposium on Computational Intelligence and Design. ISCID, Vol. 1, IEEE, pp. 130–133.
Wickman, A., 2021. Exploring Feasibility of Reinforcement Learning Flight Route Planning (Bachelor's thesis). Department of Computer and Information Science.
Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. 8 (3), 229–256.
Wu, P., Chen, J., 2022. Comparisons of RRT and MCTS for safe assured path planning in urban air mobility. p. 1841,
Wu, P., Yang, X., Wei, P., Chen, J., 2022. Safety assured online guidance with airborne separation for urban air mobility operations in uncertain environments. IEEE Trans. Intell. Transp. Syst..
Wulfe, B., 2017. UAV collision avoidance policy optimization with deep reinforcement learning.
Xian, B., Zhang, X., Zhang, H., Gu, X., 2021. Robust adaptive control for a small unmanned helicopter using reinforcement learning. IEEE Trans. Neural Netw. Learn. Syst..
Xie, Y., Gardi, A., Sabatini, R., 2021. Reinforcement learning-based flow management techniques for urban air mobility and dense low-altitude air traffic operations. In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–10.
Xu, Q., Huang, J., Liu, Z., Ding, H., 2021. A method based on deep reinforcement learning to generate control strategy for aircrafts in terminal sector. In: Artificial Intelligence in China. Springer, pp. 356–363.
Xu, Y., Prats, X., Delahaye, D., 2020. Synchronised demand-capacity balancing in collaborative air traffic flow management. Transp. Res. C 114, 359–376.
Yan, C., Xiang, X., Wang, C., 2020. Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments. J. Intell. Robot. Syst. 98 (2), 297–309.
Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., Wang, J., 2018. Mean field multi-agent reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 5571–5580.
Yang, X., Wei, P., 2020. Scalable multi-agent computational guidance with separation assurance for autonomous urban air mobility. J. Guid. Control Dyn. 43 (8), 1473–1486.
Yilmaz, E., Sanni, O., Kotwicz Herniczek, M., German, B., 2021. Deep reinforcement learning approach to air traffic optimization using the MuZero algorithm. In: AIAA AVIATION 2021 FORUM. p. 2377.
Zhang, K., Li, K., Shi, H., Zhang, Z., Liu, Z., 2020. Autonomous guidance maneuver control and decision-making algorithm based on deep reinforcement learning UAV route. Syst. Eng. Electron. 42 (7), 1567–1574.
Zhang, B., Mao, Z., Liu, W., Liu, J., 2015. Geometric reinforcement learning for path planning of UAVs. J. Intell. Robot. Syst. 77 (2), 391–409.
Zhang, K., Yang, Z., Başar, T., 2021a. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In: Handbook of Reinforcement Learning and Control. Springer, pp. 321–384.
Zhang, K., Yang, Y., Xu, C., Liu, D., Song, H., 2021b. Learning-to-dispatch: Reinforcement learning based flight planning under emergency. In: 2021 IEEE International Intelligent Transportation Systems Conference. ITSC, IEEE, pp. 1821–1826.
Zhao, W., Chu, H., Miao, X., Guo, L., Shen, H., Zhu, C., Zhang, F., Liang, D., 2020. Research on the multiagent joint proximal policy optimization algorithm controlling cooperative fixed-wing UAV obstacle avoidance. Sensors 20 (16), 4546.
Zhao, Y., Guo, J., Bai, C., Zheng, H., 2021. Reinforcement learning-based collision avoidance guidance algorithm for fixed-wing UAVs. Complexity 2021.
Zhao, P., Liu, Y., 2021. Physics informed deep reinforcement learning for aircraft conflict resolution. IEEE Trans. Intell. Transp. Syst..
Zhen, Y., Hao, M., Sun, W., 2020. Deep reinforcement learning attitude control of fixed-wing UAVs. In: 2020 3rd International Conference on Unmanned Systems. ICUS, IEEE, pp. 239–244.
Zu, W., Yang, H., Liu, R., Ji, Y., 2021. A multi-dimensional goal aircraft guidance approach based on reinforcement learning with a reward shaping algorithm. Sensors 21 (16), 5643.
Zuo, Y., Deng, K., Yang, Y., Huang, T., 2019. Flight attitude simulator control system design based on model-free reinforcement learning method. In: 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference. IMCEC, IEEE, pp. 355–361.