Journal of Aerospace Science and Technology 国际航空航天科学, 2026, 14(1), 21-29
Published Online March 2026 in Hans. https://www.hanspub.org/journal/jast
https://doi.org/10.12677/jast.2026.141004

Hans汉斯

关于深度强化学习方法在航空装备维修保障中应用的可行性分析

孟庆骁，王远锁，葛明磊

91331部队，辽宁 兴城

收稿日期：2026年1月26日；录用日期：2026年3月17日；发布日期：2026年3月24日

摘　要

当前，航空装备维修保障策略主要依靠于人工决策，对于航空装备寿命梯次编排、组训任务规划、机务保障工作规划等航空装备维修保障工作决策存在一定的局限性和短视性。本文通过对典型深度强化学习方法进行分析，围绕某飞行训练部队航空装备维修保障策略智能化开展可行性分析，建立基于多智能体的航空装备维修保障深度强化学习模型，为后续完成航空装备维修保障环境建模、多智能体强化学习训练等工作提供理论支持，对航空兵建立科学、高效的航空装备维修保障策略，实现智能化、科学化的维修保障方式，具有十分重要的现实意义。

关键词

航空装备维修保障策略，智能化，多智能体，深度强化学习，可行性分析

Feasibility Analysis on the Application of Deep Reinforcement Learning Method in Aviation Equipment Maintenance

Qingxiao Meng, Yuansuo Wang, Minglei Ge

Unit 91331 of PLA, Xingcheng Liaoning

Received: January 26, 2026; accepted: March 17, 2026; published: March 24, 2026

Abstract

Currently, aviation equipment maintenance strategies primarily rely on manual decision-making, which exhibits certain limitations and short-sightedness in tasks such as lifecycle scheduling, training mission planning, and maintenance support planning. This paper analyzes typical deep reinforcement

文章引用: 孟庆骁, 王远锁, 葛明磊. 关于深度强化学习方法在航空装备维修保障中应用的可行性分析[J]. 国际航空航天科学, 2026, 14(1): 21-29. DOI: 10.12677/jast.2026.141004

learning methods, conducts a feasibility analysis on the intelligent development of aviation equipment maintenance and support strategies for a flight training unit, and establishes a multi-agent-based deep reinforcement learning model for aviation equipment maintenance and support. It provides theoretical support for subsequent tasks such as environmental modeling for aviation equipment maintenance and multi-agent reinforcement learning training. This study holds significant practical importance for establishing scientific and efficient maintenance and support strategies for aviation forces, enabling intelligent and scientifically optimized maintenance approaches.

Keywords

Aviation Equipment Maintenance Strategy, Intelligent, Multi-Agent, Deep Reinforcement Learning, Feasibility Analysis

1. 引言

近年，随着人工智能技术的兴起，特别是 DeepMind 的 AlphaGo 击败人类围棋冠军，其背后的核心算法——深度强化学习(Deep Reinforcement Learning, DRL)，也成为继深度卷积神经网络后，又一个广泛受人关注的前沿热点。强化学习也与监督学习(Supervised Learning)和无监督学习(Unsupervised Learning)一起，并称为三大机器学习技术[1]。

2. 深度强化学习

强化学习(Reinforcement Learning, RL)是机器学习的一个重要分支，根据智能体的数量不同，可以将其分为单智能体强化学习(Single-Agent Reinforcement Learning, SARL)和多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)两类。其本质是描述和解决智能体在与环境的交互过程中学习策略以最大化累计奖励(Reward)或实现特定目标的问题[2]。

2.1. 单智能体强化学习(SARL)

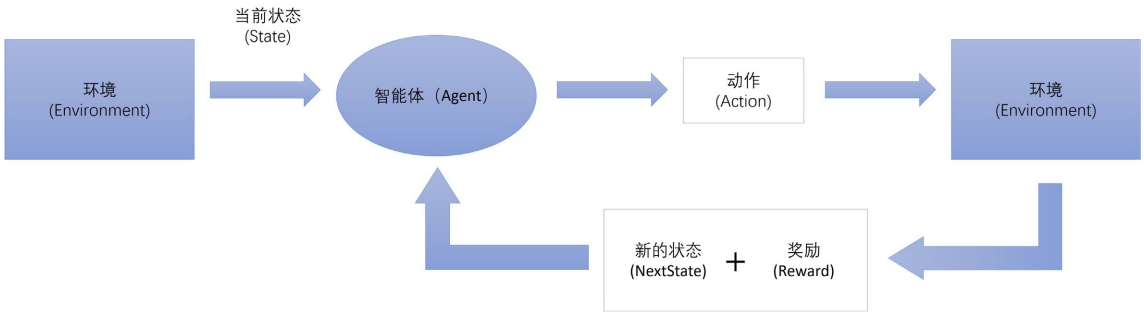

Figure 1. Basic framework of single-agent reinforcement learning
图 1. 单智能体强化学习基本框架

从 1956 年 Bellman 提出动态规划(Dynamic Programming, DP)方法[3]，到上世纪末 Sutton、Watkins、Rummery 等人提出时间差分(Temporal Difference, TD)算法、Q-Learning、SARSA 学习算法等[4]-[6]，再

到 2015 年，随着深度学习(Deep Learning, DL)的成功，人们将深度学习的方法与传统的强化学习算法相结合，由 Google DeepMind 公司提出的深度动作价值学习网络(Deep Q-Network, DQN)算法[7] [8]和蒙特卡洛树搜索(MonteCarlo Tree Search, MCTS)算法[9]，在雅达利(Atari)游戏和围棋博弈系统 AlphaGo 上获得了巨大的成功，使得完全可观测的单智能体强化学习的研究和应用得到迅速发展。

单智能体强化学习有四个重要的组成部分：动作、状态、奖励和环境。通过智能体与环境进行不断的交互来进行学习，每当智能体采取某一个动作与环境发生一次交互，就会从环境获得一个奖励(Reward)，而整个系统的状态会在智能体的动作影响下改变，其基本框架如图 1 所示。

2.1.1. 智能体的目标

智能体的任务是通过不断地与环境进行交互来获得奖励，从而学习到一种最优的策略来最大化累计奖励。当智能体以某一策略与环境进行交互时，假设其与环境的交互轨迹为 $\boldsymbol{S}_0$，$\boldsymbol{A}_0$，$\boldsymbol{R}_1$，$\boldsymbol{S}_1$，$\boldsymbol{A}_1$，$\boldsymbol{R}_2$，…(设系统的初始状态为 $\boldsymbol{S}_0$，当智能体选择动作 $\boldsymbol{A}_0$ 时，智能体与环境发生交互，得到一个奖励 $\boldsymbol{R}_1$，同时系统进入下一个状态 $\boldsymbol{S}_1$)，将智能体在 t 时刻之后能够获得的累计奖励设为 G_t，则可以得到累计奖励 G_t 的定义为：

$$G_t = R_1 + \gamma R_2 + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

其中：γ 为折扣因子($0 < \gamma < 1$)，其作用是为了计算累计奖励时，根据未来的奖励与当前时刻的时间距离进行折扣，γ 值越大说明智能体越重视遥远未来的奖励，γ 值越小说明智能体越重视当前的奖励，当 $\gamma = 0$ 时，说明智能体只关心当前的奖励，即目光短浅。

2.1.2. 价值函数

当智能体以某一策略与环境进行交互时，每遇到一个状态或做出一个动作，都会根据某一评价标准来判断当前状态或者当前采取的动作是“好”还是“坏”，而最直接的标准就是在当前状态或者在当前状态采取某一动作后，在未来能够获得多少累计奖励，这个评价标准就是价值函数。智能体的状态价值函数定义为：

$$V_\pi(s) = \mathrm{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] \tag{2}$$

即，当前状态 s 下未来累计奖励的期望。

同样，我们定义智能体的动作价值函数为：

$$q_\pi(s,a) = \mathrm{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right] \tag{3}$$

即，在当前状态 s 下采取动作 a 后，未来所有可能的交互轨迹所产生奖励的期望。

值函数的数值越高，说明智能体在当前状态 s 或在当前状态 s 执行动作 a 后遵循策略 $\pi\left(a_t \middle| s_t\right)$ 所获得的累计奖励值越高。

2.1.3. 典型单智能体强化学习算法——DQN [7] [8]

DQN 是一种基于价值(Value-Based)的强化学习算法，其通过获取最优动作价值函数，选取最大化动作价值函数所对应的动作，来构建最优策略。在基于异策略(off-policy)的 Q-Learning 算法[5]基础上结合深度神经网络对动作价值函数进行估计，采用时间差分的方法对深度神经网络参数进行单步或多步更新。假设系统全部可观测，即观测量等于状态量，将观测数据输入深度 Q 网络，输出为最大化动作值函数的动作，并通过 ε-*greedy* 策略平衡智能体对未知策略和已有策略的探索与利用。 ε-*greedy* 策略：

$$\text{Action}=\begin{cases}\arg\max_a Q(a) & \text{概率为}1-\varepsilon \\ \text{随机动作} & \text{概率为}\varepsilon\end{cases} \tag{4}$$

DQN 算法的学习框架如图 2 所示。

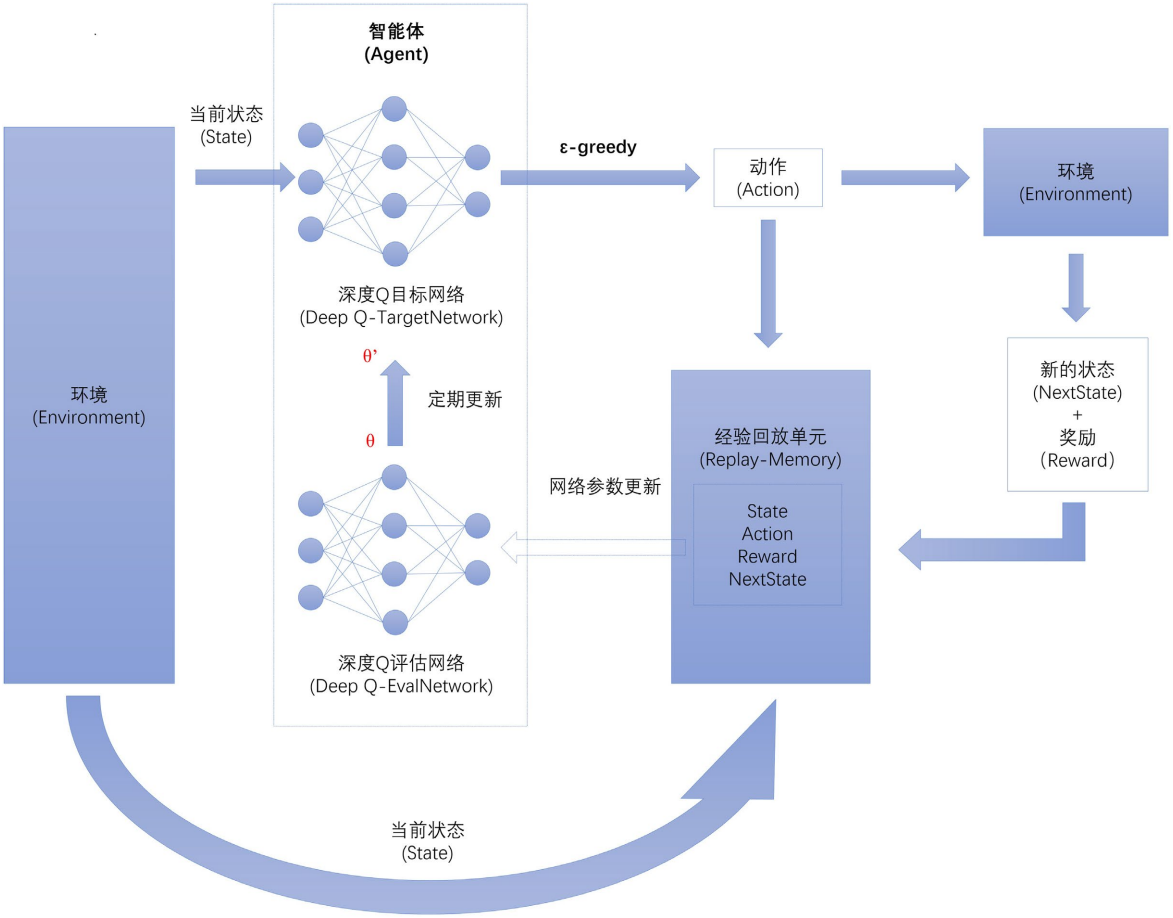

Figure 2. Basic framework of the DQN algorithm
图 2. DQN 算法基本框架

2.2. 多智能体强化学习(MARL)

多智能体强化学习算法按照其任务类型可以分为完全协作型(Fully Cooperative)、完全竞争型(Fully Competitive)和混合型(Mixed)三种算法类型[2] [10]。完全协作型任务中智能体的奖励函数是相同的，即所有的智能体都是为了实现同一个目标而努力；完全竞争型任务中，智能体的奖励函数则是相反的，环境中一般存在有两个完全敌对的智能体，智能体的目标是最大化自身的奖励，同时尽可能的最小化对方的奖励；混合型任务中智能体的奖励函数并没有确定的关系，该模型适合于自利型(Self-interested)的智能体，即最大化自身利益。

因为本文讨论的是如何通过多智能体强化学习算法，提高航空装备维修保障决策的科学性和准确性，因此可以认为是一个完全协作型的任务。

完全协作型强化学习算法的目的是，在多智能体系统中让所有智能体通过互相合作来完成一个共同的任务。与单智能体算法不同，在多智能体系统中，智能体的奖励回报不仅与环境有关，也与系统中其

他智能体的动作紧密相连，智能体不仅仅需要考虑最大化自身奖励，还需要考虑其他智能体所获得的奖励，即多智能体强化学习的目标是最大化系统全局奖励，这就导致其任务的求解相对于单智能体系统来说更加复杂。

目前，为了解决这个问题，多智能体强化学习研究主要有两个方向：一是中心式的多智能体强化学习，另一个是分布式的多智能体强化学习。

2.2.1. 中心式多智能体强化学习

中心式的多智能体强化学习，通过构造一个中心化的联合动作学习器，基于所有智能体的局部可观测信息 $O_{tot}=\{o_1,o_2,o_3,\cdots\}$ 和所有智能体的联合执行动作 $A_{tot}=\{a_1,a_2,a_3,\cdots\}$ 来学习针对所有智能体的联合动作值函数 $O_{tot}=(O_{tot},A_{tot})$，然后基于该联合动作值函数 $O_{tot}=(O_{tot},A_{tot})$ 直接给出所有智能体的行为决策 A_{tot}。这种方法相当于将整个系统看成一个统一化的单智能体，从而使用单智能体强化学习领域的方法来解决该多智能体的问题。但这种方法忽视了智能体之间的共性策略，肆意放大策略空间维度，随着系统中智能体数量的增多，使其动作空间的数量呈现指数型增长，最终导致“维度灾难”而使策略收敛缓慢甚至不可解。

2.2.2. 分布式多智能体强化学习

分布式的多智能体强化学习，其各智能体之间对于状态和动作评估的值函数之间是互相独立的，每个智能体都将其他智能体视为环境的一部分，其基于自身的局部观测信息 o_i 来求解自身的动作值函数 $Q_i(o_i,a_i)$，而后通过 $Q_i(o_i,a_i)$ 来进行动作决策，从而不用再面临“维度灾难”的问题。但这种方法通过将其他智能体视作环境的一部分，直接将环境返回的全局奖励作为自身的奖励，从而忽略了其他智能体策略的变化而导致其环境具有非平稳性，在较为复杂的协作任务中收敛较为缓慢。

2.2.3. 基于值函数分解的多智能体强化学习

基于值函数分解的多智能体强化学习方法在执行过程中与分布式强化学习方法相似，即每个智能体基于自身的局部观测信息 o_i 来求解自身的动作值函数 $Q_i(o_i,a_i)$。不同之处在于，与分布式强化学习方法中直接将环境的全局奖励视作单个智能体的奖励不同，值函数分解方法通过将独立智能体的动作值函数进行组合形成全局动作值函数用于全局奖励的估计。这种方法的代表有基于线性分解的 VDN (Value Decomposition Network)算法[11] [12]和基于单调性分解的 QMIX 算法[13] [14]。

1) 典型多智能体强化学习算法——VDN

VDN 算法通过将每个智能体的局部动作值函数 $Q_i(o_i,a_i)$ 经过简单的线性求和相加得到联合动作值函数 $O_{tot}=(O_{tot},A_{tot})$。

$$Q_{tot}(O_{tot},A_{tot})=\sum_i Q_i(o_i,a_i) \tag{5}$$

其中，$O_{tot}=\{o_1,o_2,o_3,\cdots\}$ 表示智能体的联合观测，$A_{tot}=\{a_1,a_2,a_3,\cdots\}$ 表示智能体的联合动作，$O_{tot}=(O_{tot},A_{tot})$ 表示智能体的联合动作值函数，$Q_i(o_i,a_i)$ 表示每个智能体的动作值函数。

因为全局动作值函数分解到各个局部动作值函数的方式是多种多样的，所以当真实的分解方式恰好为线性分解的时候，以 VDN 的值函数分解方式就可以保证全局动作值函数与分解后的局部动作值函数在取得最大值的时候保持一致，即局部动作值函数取最大值的时候全局动作值函数恰好也取最大值。但要满足上述分解方式在实际的问题中很难实现，因此提出了基于 QMIX 的值函数分解方式，即单调性分解。

2) 典型多智能体强化学习算法——QMIX

QMIX 算法是在 VDN 算法基础上的一种拓展。当值函数的分解方式满足线性分解方式时，VDN 分

解能够使当局部动作值函数取最大值时，全局动作值函数也取最大值。但反过来，当局部动作值函数取最大值时，全局动作值函数也取最大值的分解方式不一定满足线性分解，即 VDN 线性分解是满足局部最大整体最大的充分非必要条件。因此 QMIX 算法应运而生，它强制性地要求 Q_{tot} 与每个 Q_i 之间满足单调递增的关系，即

$$\frac{\partial Q_{tot}}{\partial Q_i} \geq 0, \quad i = 1, \cdots, n \tag{5}$$

因篇幅所限，QMIX 的具体分解方法在本文不做赘述，文献[13]给出了 QMIX 算法的具体实现方法，并在星际争霸多智能体挑战仿真环境(SMAC)中取得了很好的效果。

3. 基于多智能体的航空装备维修保障模型

3.1. 基于 QMIX 强化学习算法的航空装备保障策略可行性分析

当前，航空装备维修保障策略主要依靠于人工决策，对于航空装备寿命梯次编排、组训任务规划、机务保障工作规划等航空装备维修保障工作决策存在一定的局限性和短视性。

以某型航空器为例，对于独立的航空器其维修保障工作主要分为周期性工作和临时性工作两大类：周期性工作是指按照航空器或航空器的相关部附件因其寿命或技术要求，每经过一段时间(日历寿命)或每飞行一段时间(使用寿命)就需要进行的工作；临时性工作则是指航空器因其他原因临时需要进行的工作，包括更换故障部附件、落实技术通报、开展特定检查等工作。

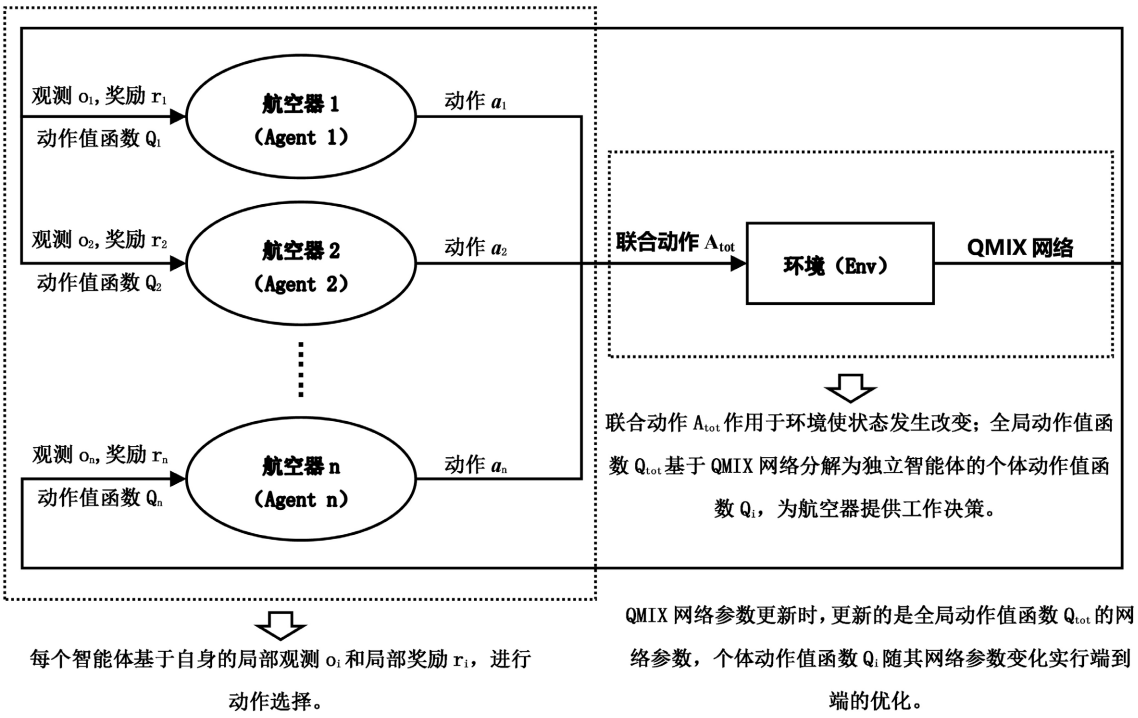

Figure 3. Multi-agent maintenance model based on QMIX
图 3. 基于 QMIX 的多智能体维修保障模型

由于当前保障人员数量和工作时间的限制，难以同时快速完成多种维修保障工作，同时，为防止因装备寿命消耗过快导致的周期性工作集中进行(例如集中进厂大修、定检等)而造成工作积压，因此，需要

通过质量控制来对航空装备的工作任务和使用计划进行规划，以满足组训任务要求。

对于飞行团等单位均可以视为多航空器系统，其航空装备维修保障仅仅依靠人工决策进行质量控制存在一定的局限性和短视性，难以进行科学合理的装备维修保障工作规划。而基于强化学习算法可以通过大量的仿真建模训练，基于式(1)的奖励设置，考虑更加长远的任务规划需求，使计算机根据年度训练任务指标自行收敛于最佳的组训、工作计划编排，并基于实时任务指标完成情况，自行调整出勤、工作任务规划。

因此，本文建立了基于多智能体的维修保障深度强化学习模型，根据年度训练任务和多航空器技术状态，考虑人员、备件等其他限制因素，为航空兵建立智能化、科学化、精细化维修保障策略提供了理论支撑。

3.2. 模型总体方案设计

对于的飞行团等多航空器系统，每个航空器作为一个独立的智能体，以完成年度训练任务为同一个目标，因此，可以将规划任务视作多智能体完全协作型的任务，基于多智能体深度强化学习框架，建立维修保障模型如图 3 所示。

3.3. 航空装备维修保障策略建模

3.3.1. 状态空间

基于 QMIX 算法，考虑人－机－环境三个方面的状态作为全局环境状态(即状态空间)：

(1) 保障人员的状态包括机组数量、机组编制、有效人数以及每个人对应的专业、累计工作时间、连续工作时间等；

(2) 航空装备的状态包括每架航空装备的完好情况、各个机件的缺件情况、发动机时间、机体时间、机上各有寿件时间、所有机件的备件情况、当前进行工作等；

(3) 自然环境状态包括当前的温度、湿度、风力、天气(晴、阴、雨、雪等)等。

状态空间结构建模如表 1 所示。

Table 1. State space
表 1. 状态空间

序号		状态空间		
1	机组 AgentCrew	编号 ID	…	
		机组数量 CrewNum	机组编制 PeopleNum	
			专业编配 SpecialtyNum	
			当前人数 NowPeopleNum	
			…	…
		工作 Work	编号 ID	…
			名称 Name	定检 RegularCheckup
				大修 HeavyRepair
				…
			耗时 ElapseTime	…
		…	…	…

续表

		编号 ID	…	
		名称 Name	…	
2	航空器 AgentAircraft	机件 Component	编号 ID	…
			名称 Name	…
			型号 Type	…
			寿命 UsablePeriod	…
			已用寿命 ElapsedTime	…
			剩余寿命 LeftTime	…
			故障概率 MalfunctionPro	…
			排故工作时间 TroubleshootingTime	…
			备件数量 SpareComponent	…
			…	…
		工作 Work	定检 RegularCheckup	编号 ID
				名称 Name
				定检周期 UsablePeriod
				…
			大修 HeavyRepair	编号 ID
				名称 Name
				大修次数 TimesofHeavyRepair
				…
		…	…	…
3	自然环境 NatureFactor	场地容量 Capacity	…	
		气象日概率 WeatherPro	…	
		停飞整顿 Reorganize	整顿时间 ReorganizeTime	…
			整顿概率 ReorganizePro	…
		停飞时间 Rest	停飞时间 RestStart	…
			开飞时间 RestStop	…
		…	…	

3.3.2. 动作空间

智能体的动作空间包括两个方面，一是保障人员工作，包括出勤、值班、待命等；二是航空装备工作，包括出勤、替补、待命、维修等。

3.3.3. 奖励函数

奖励函数设置为一段时间内的航空装备累计工作时长、发动机/机体剩余寿命梯度合理性、保障人员

的连续工作时长、航空装备完好率等。例如：当航空装备累计工作时长越短、寿命梯度合理、保障人员连续工作时间越短、航空装备完好率越高时给智能体一个正向奖励，反之则给智能体一个负奖励(即惩罚)。

3.3.4. 其他约束条件

本文模型约束条件可以分为确定性约束条件和随机因素约束条件两类。

保障人员模型中的确定性约束有定检用时、排故用时等，随机性因素有工作效率、工作失误率等。

航空装备模型中的确定性约束有航空装备定检周期、有寿件更换周期等，随机性因素有设备机件的故障率等。

自然环境模型中的确定性约束有雨雪大风天气时的保障人员和航空装备只能处于待命状态，随机性因素有天气和风力的骤变概率等。

4. 结束语

深度强化学习近年来在人工智能领域得到了广泛的关注，通过对已知环境建模，使智能体不断与环境进行交互、试错，逐步学习到适合于当前环境的智能体动作策略。因此，本文通过分析单智能体强化学习方法 DQN 和多智能体学习方法 VDN、QMIX，对基于多智能体强化学习的航空装备维修保障策略的智能化进行可行性分析，建立了航空装备维修保障策略模型，为后续完成航空装备维修保障环境建模、多智能体强化学习训练等工作提供了理论支持，对航空兵实现智能化、科学化的维修保障，具有十分重要的现实意义。

参考文献

[1] 李茹杨, 彭慧民, 李仁刚, 赵坤. 强化学习算法与应用综述[J]. 计算机系统应用. 2020, 29(12): 13-25.

[2] 孙彧, 曹雷, 陈希亮, 徐志雄, 等. 多智能体深度强化学习研究综述[J]. 计算机工程与应用. 2020, 56(5): 13-24.

[3] Bellman, R. (1956) Dynamic Programming and Lagrange Multipliers. *Proceedings of the National Academy of Sciences*, **42**, 767-769. https://doi.org/10.1073/pnas.42.10.767

[4] Sutton, R.S. (1988) Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, **3**, 9-44. https://doi.org/10.1023/a:1022633531479

[5] Watkins, C.J.C.H. and Dayan, P. (1992) Technical Note: Q-Learning. *Machine Learning*, **8**, 279-292. https://doi.org/10.1023/a:1022676722315

[6] Rummery, G.A. and Niranjan, M. (1994) On-Line Q-Learning Using Connectionist Systems. Technical Report, 1-7.

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., *et al*. (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529-533. https://doi.org/10.1038/nature14236

[8] Mnih, V., Kavukcuoglu, K., Silver, D., *et al*. (2013) Playing Atari with Deep Reinforcement Learning.

[9] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., *et al*. (2016) Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, **529**, 484-489. https://doi.org/10.1038/nature16961

[10] 杜威, 丁世飞. 多智能体强化学习综述[J]. 计算机科学. 2019, 8(46): 1-8.

[11] Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., *et al*. (2018). Value-Decomposition Networks for Cooperative Multi-Agent Learning Based on Team Reward. *International Joint Conference on Autonomous Agents and Multiagent Systems*, Stockholm, 10-15 July 2018, 2085-2087. https://doi.org/10.65109/jsrc7365

[12] 李盛祥. 基于强化学习的多智能体协同关键技术及应用研究[D]: [博士学位论文]. 郑州: 战略支援部队信息工程大学, 2021.

[13] Rashid, T., Samvelyan, M., Witt, C.S., *et al*. (2018) Qmix: Monotonic Value Function Factorization for Deep Multi-Agent Reinforcement Learning. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 4292-4301.

[14] 吴昊霖. 基于协作多智能体强化学习的飞行冲突解脱策略研究[D]: [博士学位论文]. 四川: 四川大学, 2021.