# 基于不确定场景的多决策风格智能任务分配方法

刘家义[1]，王 刚[2]，贾晨星[1]，付 强[2]，明月伟[1]
（1.国防大学联合作战学院，石家庄，050084；2.空军工程大学防空反导学院，西安，710051）

**摘要** 现代信息化战争中，战场环境复杂多变，具有高动态、不完全信息和不确定性等特点，深度强化学习为其中的任务分配问题提供了新思路。针对智能体在不确定场景中泛化能力不足的问题，提出了面向不确定场景的多决策风格智能体架构，增强了智能体面对不确定复杂环境的适应能力；针对深度强化学习方法中单一奖励函数很难训练出符合人类决策逻辑的智能体问题，提出了基于事件的奖励机制，合理引导智能体学习；最后在数字战场仿真环境中验证了所提方法的可行性和优越性。

**关键词** 深度强化学习；任务分配；多智能体系统；决策风格
**DOI** 10.3969/j.issn.2097-1915.2025.01.013
**中图分类号** TP391.9 **文献标志码** A **文章编号** 2097-1915(2024)01-0104-07

# An Intelligent Task Assignment Method for Different Decision Styles Based on Uncertain Scenario

LIU Jiayi[1], WANG Gang[2], JIA Chenxing[1], FU Qiang[2], MING Yuewei[1]
(1. Joint Operations College, National Defense University, Shijiazhuang 050084, China;
2. Air Defense and Antimissile School, Air Force Engineering University, Xi'an 710051, China)

**Abstract** The battlefield environments being complex, dynamic, characterized by high dynamics, incomplete information, and uncertainty, the deep reinforcement learning (DRL) is enabled to provide a new way of thinking about task assignment in modern information warfare. Aimed at the problem that the agent system is inadequate in generalization ability under condition of uncertain scenario, this paper proposes an event-based reward mechanism to reasonably guide the learning of the agent, and the problem that in deep reinforcement learning, a single reward function is difficult to train an agent of being in keeping with human decision logic, this paper proposes an event-based reward mechanism to reasonably guide the learning of the agent. And this paper proposes a multi-agent architecture for different decision styles, enhancing the ability of the agent to adapt to complex environments. Finally, the feasibility and superiority of the proposed method are verified on a digital battlefield.

**Key words** deep reinforcement learning; task assignment; multi-agent systems; decision styles

武器目标分配(weapon target assignment, WTA)是现代防空作战、指挥决策中的关键问题，其核心是如何把不同的来袭目标分配到具有不同杀伤力和经济价值的拦截武器，以构成整体优化的火力

**收稿日期**：2024-01-25
**基金项目**：国家自然科学基金(62106283)
**作者简介**：刘家义(1996—)，男，福建福州人，工程师，博士，研究方向为智能辅助决策。E-mail：sixandone1@163.com

*引用格式：刘家义，王刚，贾晨星，等.基于不确定场景的多决策风格智能任务分配方法[J].空军工程大学学报，2025，26(1)：104-110. LIU Jiayi, WANG Gang, JIA Chenxing, et al. An Intelligent Task Assignment Method for Different Decision Styles Based on Uncertain Scenario[J]. Journal of Air Force Engineering University, 2025, 26(1): 104-110.*

打击体系[1]。任务分配是在目标分配基础上提出的概念，当作战任务被分解为不同类型的元任务后，目标分配将转化为任务分配。任务分配以不同类型作战要素的武器装备为完成任务的基本单元，将作战任务分解为作战要素可执行的元任务[2]。比如可将防空反导作战任务分解为跟踪、拦截两大子任务，子任务又分为多个可执行的元任务。结合包以德循环理论（observation orientation decision action，OODA）以及杀伤链和杀伤网中信息流转的概念[3]，任务分配可以充分利用防空反导武器系统中的传感器和拦截器，构造一个严密的杀伤网，灵活性高，抗毁性强，更适合于分布式作战。

当前战场环境动态多变、约束众多且复杂，在分布式任务分配问题中，传统数学建模无法逼真反映真实的战争过程，且现有求解方法也存在求解速度不足的问题。而多智能体系统（multi-agent system，MAS）对复杂系统的描述能力和深度强化学习（deep reinforcement learning，DRL）对动态环境的行为建模能力恰好能解决这一问题[4]。

针对信息化作战中的分布式任务分配问题，本研究结合 MAS 和 DRL 提出一种面向多决策风格（multiple decision style，MDS）的智能体架构，以解决完全分布式多智能体结构全局协调性差和 DRL 奖励函数难以准确构建的问题，提出基于关键事件的奖励机制，合理引导智能体学习到类似人类的决策风格，在对抗环境中对该方法的有效性和优越性进行了验证，结果表明，智能体能根据态势变化在不同风格间转换，让体系内功能不同的武器系统能够进行有效的组织协调和协同作战，有效应对大规模不确定场景。

## 1　智能体设计

强化学习（reinforcement learning，RL）是利用试错法和奖励来训练智能体学习的方法，广泛应用于求解大规模复杂化问题，其基本环境是一个马尔科夫决策过程。智能体（agent）从环境感知当前状态（state），然后做出相应的行为（action），得到对应的奖励（reward）。然而在实际的大规模复杂问题中，RL 常常会遇到维数灾难的问题。学者们利用深度学习的深度神经网络作为函数拟合器来解决这一问题，与 RL 结合，诞生了 DRL[5]。

指挥决策是通过将对抗过程在思维中进行提前勾画，预测对手可能的行动并有针对性地指导我方行动。将 DRL 应用于指挥决策时，面对复杂多变的战场环境，使用单一奖励函数训练的智能体难以兼顾多种情况进行合理决策，原因是在大规模不确定场景中往往很难使用一个简洁而准确的奖励值函数去评价当前动作的好坏[6]。因此，在此类复杂场景的深度强化学习训练中，如何合理定义奖励值函数是一个较大的难题。前期研究使用的奖励函数相对简单，一定程度上导致智能体需要较长训练回合才能得到相对较为合理的策略，且智能体应对不确定场景泛化能力较弱。针对以上问题，本研究提出面向不确定场景的 MDS 智能体架构，在一主多从多智能体（one-general agent with multiple narrow agents，OGMN）架构[7]的基础上增加了一个风格选择智能体，根据当前态势信息选择具有不同决策风格的智能体进行决策，增强场景泛化能力。

### 1.1　严格不确定型决策问题

决策问题可以分为确定型决策问题、严格不确定型问题和风险型决策，其中的严格不确定型问题是指由于存在不确定因素，同一个决策可能会对应多个不同的状态，此时已知可能出现的所有状态，但每个状态发生的概率却是未知的。面对不确定决策，需要确定一些决策准则，根据具体情况和决策准则来选择方案。常见的决策准则有追求风险与利益并存的乐观准则、总是持保守态度的悲观准则以及既不过于冒险也不过于保守的折中准则等[8]。

### 1.2　MDS 智能体架构研究

当人类作为决策主体时，会利用经验和模型识别，针对不同的态势匹配不同的决策风格，快速高效地做出决策[9]。面对多变的场景，不同决策风格的科学搭配能够提高决策效果。因此，本研究提出了 MDS 多智能体架构，模仿人类指挥员的决策理念，基于不确定型问题的决策准则设计了乐观、悲观和折中 3 种不同风格的调度智能体，分别用于应对不同的场景，并在调度智能体之上增加了一个风格选择智能体，根据当前态势有针对性地选择决策风格。该方法本质上是用多个不同的奖励函数来增强奖励值的合理性，提升红方智能体决策的博弈对抗能力，具体架构如图 1 所示。

在此架构中，3 种不同风格的智能体代表了 3 套不同的神经网络参数。风格选择智能体根据作战规则进行决策，其输入为当前的态势信息，输出为使用 3 种风格中的哪一种智能体进行决策，风格切换的决策频率为 3 s/次。针对智能体的训练，3 种不同风格的智能体根据各自的奖励函数单独训练，训练算法为近策略优化（proximal policy optimization，PPO）[10]。

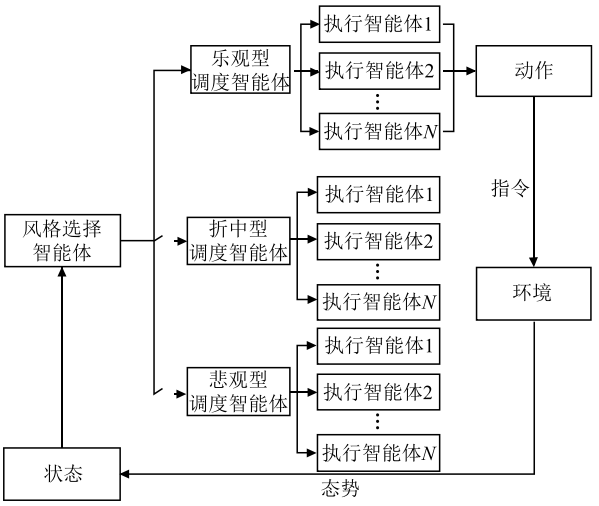



**图 1 MDS 多智能体架构**

**Fig. 1 Multi-agent architecture for different decision styles**

### 1.3 风格选择智能体设计

大规模博弈对抗过程极其迅速和多变，风格选择智能体的判断是在时间极其有限的情况下进行的，其判断内容包括空情、我方单位的状态和能力、射击条件等。

定义空情复杂度系数 $\alpha$ 为：

$$\alpha=\frac{m_{\mathrm{t}}}{R_{\mathrm{F}}} \tag{1}$$

式中：$m_{\mathrm{t}}$ 为当前时刻识别到的蓝方目标总数；$R_{\mathrm{F}}$ 为红方当前时刻可用的火力通道总数。

定义目标威胁系数 $\beta$：

$$\beta=\frac{m_{\mathrm{h}}}{m_{\mathrm{t}}} \tag{2}$$

式中：$m_{\mathrm{h}}$ 为蓝方目标中的有人目标数量；$m_{\mathrm{t}}$ 为当前时刻识别到的蓝方目标总数。

定义蓝方目标的可杀伤距离为蓝方目标到达该距离时，可以攻击红方单位且被红方单位消灭的概率很低，或蓝方目标可以在被消灭时损伤红方单位。可杀伤距离取决于蓝方目标的高度、速度以及红方使用的武器类型。

根据相关领域知识可确定风格选择智能体的切换规则，部分规则如下：

1）空情复杂度系数 $\alpha<0.2$ 时，优先使用乐观型。

2）空情复杂度系数 $0.2\leqslant\alpha<1$ 时，优先使用折中型。

3）空情复杂度系数 $\alpha\geqslant1$ 时，优先使用悲观型。

4）若高威胁目标进入可杀伤距离范围，则切换为乐观型优先拦截。

5）若发现的目标都尚未进入可杀伤距离范围，且目标威胁系数 $\beta<0.1$ 时，切换为悲观型拦截。

6）目标威胁系数 $\beta\geqslant0.5$ 时，切换为乐观型。

7）当红方已暴露的单位都在进行杀伤任务，且目标威胁系数 $0.1\leqslant\beta<0.5$ 时，切换为折中型拦截。

### 1.4 基于关键事件的奖励机制

本研究基于不确定型决策问题的决策准则设计了乐观型、折中型和悲观型 3 种决策风格，区分不同决策风格的主要方法是在训练中使用不同的奖励函数，同时使用基于关键事件的奖励机制，通过对抗场景和作战规则提炼能够引导智能体学习相应决策风格的关键事件，通过奖励值鼓励智能体学习特定的行为。不同决策风格智能体对应奖励函数的设计思路是基于不确定型决策问题的决策准则和各个决策风格的特点，通过多次仿真测试优化对各事件的奖励权重进行增减，最终确定每个决策风格对应的具体奖励设计，以训练出固定决策风格的智能体。乐观型、折中型和悲观型 3 种决策风格，区分使用奖励分数。奖励设计方面，由于红蓝双方单位数量较多，因此状态空间和动作空间都较大，若只在每轮对战结束后，一次性给予胜负的奖励值，则奖励非常稀疏，因此，除了给出最终的奖励即 episodic reward，还添加了较密集的基于事件的奖励机制，以训练出固定决策风格的智能体。每种风格的奖励函数设计都是为了鼓励特定的行为，各个决策风格的具体奖励分数如下：

1）乐观型主智能体的奖励值如表 1 所示。

**表 1 乐观型主智能体的奖励值**

**Tab. 1 Reward values for optimistic main agents**

| 结果 | 获胜 | 指挥所遭攻击 1 次 | 机场遭攻击 1 次 | 中远程雷达毁伤 |
|---|---|---|---|---|
| 分数 | 50 | −20 | −10 | −7 |
| 结果 | 近程雷达毁伤 | 消耗中远程导弹 | 消耗近程导弹 | 拦截无人机 |
| 分数 | −3 | −0.2 | −0.1 | 5 |
| 结果 | 拦截战斗机 | 拦截轰炸机 | 拦截敌方导弹 | 射击距离过近 |
| 分数 | 15 | 15 | 1 | −2 |

乐观型智能体主要用于需要快速拦截蓝方高威胁度目标的情况，为了引导智能体形成基于乐观型决策准则的决策风格，在奖励值的直观体现是增加打击蓝方有人目标的正向奖励权重；为了让智能体尽快打击蓝方单位，当射击距离太近时智能体会得到相应的惩罚；为了让智能体积极进攻，减少了消耗弹药的负反馈权重。

2）折中型主智能体的奖励值如表 2 所示。

**表 2　折中型主智能体的奖励值**

**Tab. 2　Reward values for mediumsized main agents**

| 结果 | 获胜 | 指挥所遭攻击 1 次 | 机场遭攻击 1 次 | 中远程雷达毁伤 |
|---|---|---|---|---|
| 分数 | 50 | −20 | −10 | −7 |
| 结果 | 近程雷达毁伤 | 消耗中远程导弹 | 消耗近程导弹 | 拦截无人机 |
| 分数 | −3 | −0.5 | −0.3 | 2 |
| 结果 | 拦截战斗机 | 拦截轰炸机 | 拦截巡航导弹 | 拦截飞行器的导弹 |
| 分数 | 8 | 8 | 1 | 2 |

折中型智能体主要用于使用一些战术战法的情况，有意识引导对方决策，占据主动。因此，折中型智能体作为对战中使用频率最高的决策风格，需要平衡各种资源的使用，把握机会在不同风格间转换，最终取得胜利。

3）悲观型主智能体的奖励值如表 3 所示。

**表 3　悲观型主智能体的奖励值**

**Tab. 3　Reward values for pessimistic main agents**

| 结果 | 获胜 | 指挥所遭攻击 1 次 | 机场遭攻击 1 次 | 中远程雷达毁伤 |
|---|---|---|---|---|
| 分数 | 50 | −25 | −15 | −12 |
| 结果 | 近程雷达毁伤 | 消耗中远程导弹 | 消耗近程导弹 | 拦截无人机 |
| 分数 | −7 | −0.9 | −0.5 | 3 |
| 结果 | 拦截战斗机 | 拦截轰炸机 | 拦截巡航导弹 | 拦截飞行器的导弹 |
| 分数 | 6 | 6 | 1 | 5 |

悲观型智能体主要用于不确定因素较多的情况，进行保守防御观察态势。主要是在保护红方单位的前提下，尽可能多节省资源，暴露较少单位。因此增加了红方单位损失的惩罚，提高拦截蓝方导弹的正向奖励，降低拦截蓝方单位的正向奖励。

## 2　对抗场景设置

本研究的对抗环境为面向基于 DRL 防空作战模拟框架（DRL-oriented air defence combat simulation framework）的数字战场[11]。对抗场景为在想定作战区域内，针对一定数量的蓝方进攻兵力，红方利用有限的资源保卫重要的单位，根据蓝方的威胁程度等因素进行任务分配，在使用最少资源的前提下保护重要单位不被摧毁。双方的胜负条件为：①当红方重要单位受到 3 次攻击时，蓝方胜；②红方所有单位损失超过 60％时，蓝方胜；③蓝方有人单位损失超过 30％时，红方胜。考虑红方拦截单位的火力衔接和重叠，同时保证一定的杀伤纵深，红方的资源部署如图 2 所示。

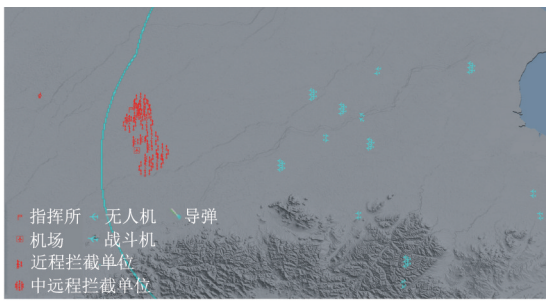



**图 2　各单位部署情况**

**Fig. 2　Deployment by units**

红方的单位包括：需要保护的指挥所和机场，1 架预警机，9 个中远程传感器、72 个中远程拦截器、7 个近程传感器、21 个近程拦截器，216 枚中远程导弹、372 枚近程导弹。导弹对巡航弹高杀伤概率为 45％，低杀伤概率为 35％；对其他单位高杀伤概率为 75％，低杀伤概率为 55％。场景中传感器探测范围将受到地图中地物遮蔽影响，其极限视距为：

$$R_{\max}=4.12(\sqrt{H_{\mathrm{T}}}+\sqrt{H_{\mathrm{R}}}) \tag{3}$$

式中：$H_{\mathrm{T}}$ 为蓝方来袭目标的海拔高度；$H_{\mathrm{R}}$ 为红方传感器天线的海拔高度，在本实验中 $H_{\mathrm{R}}=4\ m$。

蓝方的单位包括：18 枚巡航导弹，30 架无人机，16 架战斗机，4 架轰炸机，2 架干扰机，156 枚反辐射导弹、62 枚空对地导弹。反辐射导弹命中率 80％，空对地导弹命中率 80％。第 1 批次，蓝方的所有巡航导弹从 2 个方向分别攻击红方的机场和指挥所，飞行高度为 100 m，红方必须合理规划资源，在拦截的前提下让弹药资源消耗最小；第 2 批次，蓝方的 30 架无人机 2～3 km 高度突防，16 架战斗机飞行高度 100 m 超低空突防；最后一个批次，蓝方所有轰炸机突防轰炸红方要地。

实验定义蓝方的突防路线、到达时间、分队编成随机变换，来袭方向整体不变的场景为随机场景。本实验中 PPO 算法的超参数 $\varepsilon=0.2$，学习率为 $10^{-4}$，批尺寸为 5 120，神经网络中隐藏层单元数分别为 128 和 256，并行运行 72 个 Actor，在每轮中更新时迭代 3 次。

## 3　实验与结果

### 3.1　决策风格比较

训练硬件配置为：CPU，型号 Intel Xeon E5-2678V3，88 核，256 G 内存；GPU * 2，型号 Nvidia GeForce 2080Ti，72 核，11 GB 显存。

3.1.1 行为分析

智能体在数字战场环境中训练 50 000 步，不同决策风格的智能体都产生了相应的有效战术行为。这些行为是在训练过程中，通过基于事件的奖励机制引导，由智能体自主探索的。

3.1.1.1 乐观型智能体

重点目标，优先摧毁。蓝方战斗机是高威胁目标，应迅速拦截，避免空情更加复杂后难以拦截。1 架战机通常携带多枚导弹，在发射导弹前进行拦截也能有效节约拦截资源。经过训练后，乐观型智能体能在保证自身单位安全的前提下，有效锁定战斗机并拦截。优先摧毁重点目标行为如图 3 所示。

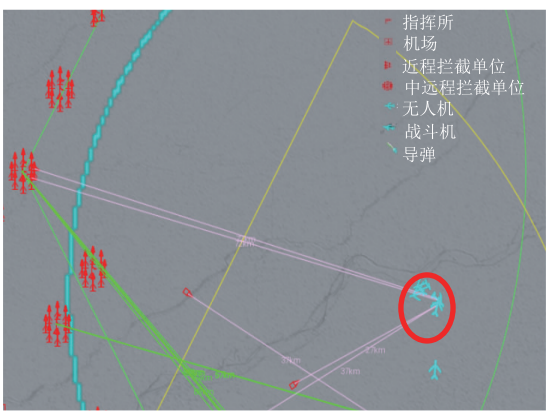



**图 3 优先摧毁重点目标行为示意图**

**Fig. 3 Diagram of priority target destruction behaviour**

3.1.1.2 折中型智能体

火力协同，合理拦截。训练后，面对更加复杂的空情，智能体能够调度拦截资源协同保卫要地。对图中低空突防的几枚 ARM，智能体不仅调度下方的火力单元进行拦截，而且调度上方的火力单元进行拦击，以免贻误战机。对于进入防区内的战斗机，调度远程防空资源进行拦截，对于锁定其的蓝方导弹则调度周围火力单元协同拦截，避免其被毁伤。通过训练，折中型智能体在调度拦截资源时表现出了更好的火力协同能力。火力协同行为如图 4 所示。

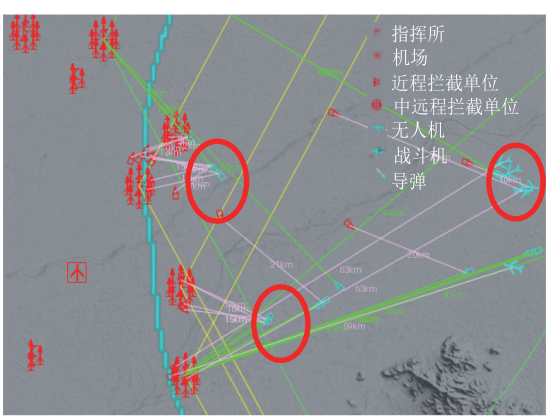



**图 4 火力协同行为示意图**

**Fig. 4 Diagram of fire synergy behaviour**

3.1.1.3 悲观型智能体

稳扎稳打，伺机而动。对低价值、易拦截目标，如巡航导弹，应该尽量节约远程拦截资源，采用近程拦截资源进行拦截。既可以避免暴露远程雷达位置，又能节约远程拦截资源攻击高价值目标。训练后，智能体能在蓝方巡航弹来袭初期保持雷达静默，避免暴露红方位置，当蓝方目标进入红方近程火力单元拦截范围时，红方近程雷达开机，由近程火力单元进行拦截，既节约了火力资源，又能避免红方阵地暴露。防御行为如图 5 所示。

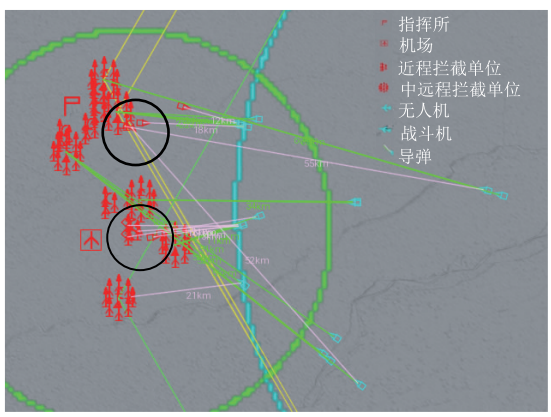



**图 5 防御行为示意图**

**Fig. 5 Diagram of defensive behaviour**

3.1.2 战斗数据对比

虽然单一风格的智能体在行为上能够涌现出相应的战术，但在整局对战中，面对复杂多变的态势，表现还是略显被动。根据环境在智能体训练过程实时记录的战斗数据，分析 3 种不同决策风格智能体的表现，选取了不同风格智能体在训练 50 000 步后，红蓝双方对抗 100 局的平均对抗情况，如表 4 所示。

**表 4 训练过程中智能体表现情况**

**Tab. 4 Performance of the intelligent agent during the training process**

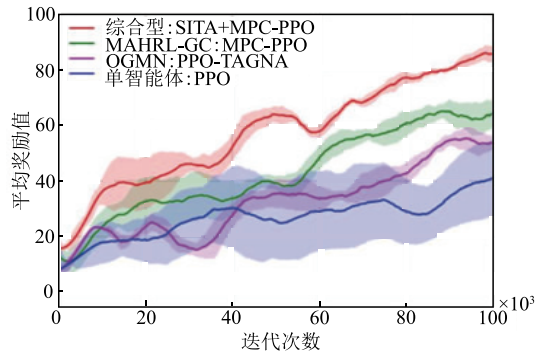
| 指标 | 乐观型 | 折中型 | 悲观型 |
|---|---|---|---|
| 中远程传感器损失 | 6.787 | 5.315 | 3.298 |
| 近程传感器损失 | 4.593 | 3.584 | 1.782 |
| 受保卫对象被攻击次数 | 4.129 | 2.543 | 1.059 |
| 中远程导弹发射次数 | 201.650 | 185.350 | 182.140 |
| 近程导弹发射次数 | 182.650 | 289.610 | 319.060 |
| 无人机毁伤平均数 | 10.369 | 11.635 | 8.783 |
| 战斗机毁伤平均数 | 6.861 | 4.483 | 2.134 |
| 轰炸机毁伤平均数 | 1.387 | 1.137 | 0.639 |
| 巡航弹拦截平均数 | 17.365 | 17.640 | 17.620 |
| 反辐射弹拦截平均数 | 53.240 | 75.410 | 92.630 |
| 空对地导弹拦截平均数 | 19.190 | 26.410 | 38.150 |

可以看出，经过不同的奖励函数引导进行训练后，几个智能体都形成了各自的决策风格：乐观型智

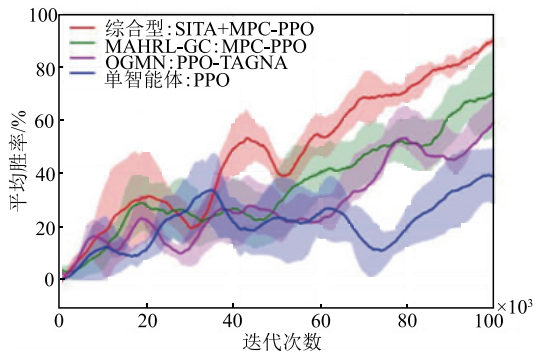
能体远程导弹发射次数最多，击毁蓝方飞机数量最多，但也由于雷达过早开机而暴露，中远程雷达损失最多，且后期弹药匮乏导致要地被攻击；折中型智能体在资源使用和拦截蓝方的单位类型上都比较平均，但往往容易被蓝方低价值目标所误导，错失拦截高威胁目标的时机；悲观型智能体近程资源使用率最高且拦截蓝方导弹最多，同时雷达损失少，但由于前期一味防御错失进攻时机，导致后期无法拦截轰炸机而失败。因此，各种风格都各有利弊，能够针对特定场景做出对应的决策，但面对复杂多变的态势，还需集众家之所长，训练一个面向多种决策风格的综合型智能体，灵活运用多种风格掌控局面。

### 3.2　面向不同决策风格的综合型智能体优越性验证

为分析当训练场景和应用场景均随机时 4 种方法的决策水平差异，分别将 4 种方法在随机场景中迭代训练 100 000 次，统计 4 种方法所获得的平均奖励值和平均胜率，对比结果如图 6 所示（各个阴影区域为 3 次实验的置信区间）。




(a)随机场景的平均奖励值



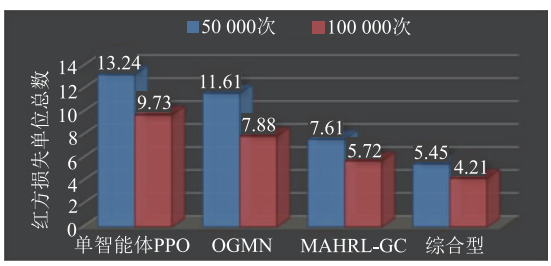

(b)随机场景的平均胜率

**图 6　训练数据曲线**

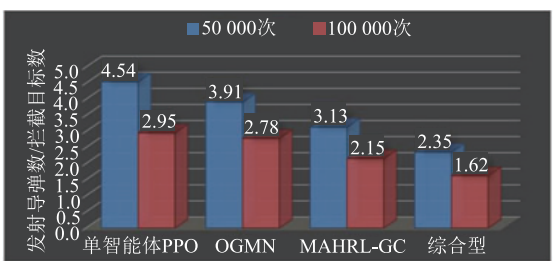
**Fig. 6　Training data graphs**

从图中可以看出，随着训练次数的增加，4 种方法所获得的奖励值和胜率都有所提高，说明各方法都能在训练过程中提升智能体的决策水平，但面对随机场景，单智能体 PPO 方法[10]和 OGMN 方法[7]的训练均不太稳定，曲线震荡较大，上升趋势不明显；MAHRL-GC 方法[12]中包含基于领域知识的规则，故具有一定的泛化能力，在随机的场景中表现较好，平均奖励值提升较为稳定，最终胜率可提升至 70%左右；综合型智能体包含了 3 种决策风格，面对随机的场景训练收敛速度最快，曲线整体的上升趋势明显，获得的平均奖励值最高可达到 82 左右，最终对抗胜率约为 88%。实验表明，本研究提出的面向不同决策风格的分配策略是有效的，能够提升智能体适应不确定环境的场景泛化能力。

实验选取 4 种方法在随机场景中训练 50 000 次和训练 100 000 次时的模型，分别在随机场景中与蓝方对抗推演 100 局，战斗情况统计结果如图 7 所示。其中，图 7(a)为损失数量统计结果，横轴是迭代训练次数，纵轴是红方损失单位数量之和；图 7(b)为发射的导弹与拦截的目标之比，值越高表明拦截同样多目标所需资源数量越多，拦截效率越低，值越低则拦截效率越高。




(a)损失单位数量统计




(b)拦截效率统计

**图 7　对抗数据统计**

**Fig. 7　Matching statistics**

总结发现，随着训练的进行，4 个方法的损失单位总数、发射导弹数与拦截目标数之比都在逐渐下降，说明 4 种方法的决策水平都在逐渐提升；综合型智能体训练 50 000 次和 100 000 次的模型，均好于其他 3 个方法同时期的模型，且综合型智能体的 2 个模型差异最小，说明该方法的训练效率最高，智能体在 50 000 次时已经初步学习到了有效的策略，表明本研究提出的切换风格的机制能够使智能体的训练速度和决策水平都有所提升。

因此，实验表明本研究的多智能体分层架构结合不同决策风格的切换机制，形成的综合型智能体能更好地适应复杂的不确定场景，在不确定场景中的训练效率更高，智能体能够更有效优化和分配拦截资源。

## 4 结语

本研究将防空作战任务分配问题用 DRL 方法求解，结合 MAS 提出了面向不同决策风格的智能任务分配策略。提出了基于事件的奖励策略，合理引导智能体在训练中尽快学习到类似人类的决策风格。提出了面向不同决策风格的智能体架构，其中包含风格选择智能体和 3 个具有不同决策风格的智能体。风格选择智能体根据当前态势选择合适的风格进行决策，以期最大化决策过程中的作战效能。在训练过程中发现，不同的奖励函数能够引导智能体逐渐形成相应的决策风格。从结果看，所提出的方法与单一风格智能体相比有明显优势。下一步，将基于态势认知来训练智能体，通过当前态势信息预测对方未来时刻的策略，缩短风格切换的时间。

### 参考文献


[1] KONG L R, WANG J Z, ZHAO P. Solving the Dynamic Weapon Target Assignment Problem by an Improved Multiobjective Particle Swarm Optimization Algorithm[J]. Applied Sciences, 2021, 11(19): 9254.

[2] 刘建，陈桂明，李新宇. 基于遗传算法的作战任务分配和资源调度问题研究[J]. 兵工自动化，2023，42(7)：59-63，73.
LIU J, CHEN G M, LI X Y. Research on Combat Task Assignment and Resource Scheduling Based on Genetic Algorithm[J]. Ordnance Industry Automation, 2023, 42(7): 59-63, 73. (in Chinese)

[3] 陈登，陈楚湘，周春华. 基于 OODA 环的杀伤网节点重要性评估[J]. 兵工学报，2024，45(2)：363-372.
CHEN D, CHEN C X, ZHOU C H. Importance Evaluation of Kill Network Nodes Based on OODA Loop[J]. Acta Armamentarii, 2024, 45(2): 363-372. (in Chinese)

[4] AN B Z, HUANG B M, ZOU Y, et al. Distributed Optimization for Uncertain Nonlinear Interconnected Multi-Agent Systems[J]. Systems & Control Letters, 2022, 168: 105364.

[5] NGUYEN T T, NGUYEN N D, NAHAVANDI S. Deep Reinforcement Learning for Multiagent Systems: a Review of Challenges, Solutions, and Applications[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839.

[6] 殷宇维，王凡，吴奎，等. 基于改进 DDPG 的空战行为决策方法[J]. 指挥控制与仿真，2022，44(1)：97-102.
YIN Y W, WANG F, WU K, et al. Research on Air Combat Behavior Decision-Making Method Based on Improved DDPG[J]. Command Control & Simulation, 2022, 44(1): 97-102. (in Chinese)

[7] LIU J Y, WANG G, FU Q, et al. Task Assignment in Ground-to-Air Confrontation Based on Multiagent Deep Reinforcement Learning[J]. Defence Technology, 2023, 19: 210-219.

[8] 岳超源. 决策理论与方法[M]. 北京：科学出版社，2003.
YUE C Y. Decision Theory and Method[M]. Beijing: Science Press, 2003. (in Chinese)

[9] WALSH S E, FEIGH K M. Understanding Human Decision Processes: Inferring Decision Strategies from Behavioral Data[J]. Journal of Cognitive Engineering and Decision Making, 2022, 16(4): 301-325.

[10] GUAN W, CUI Z W, ZHANG X K. Intelligent Smart Marine Autonomous Surface Ship Decision System Based on Improved PPO Algorithm[J]. Sensors, 2022, 22(15): 5732.

[11] FU Q, FAN C L, SONG Y F, et al. Alpha C2-An Intelligent Air Defense Commander Independent of Human Decision-Making[J]. IEEE Access, 2020, 8: 87504-87516.

[12] LIU J Y, WANG G, GUO X K, et al. Intelligent Air Defense Task Assignment Based on Hierarchical Reinforcement Learning[J]. Frontiers in Neurorobotics, 2022, 16: 1072887.


（编辑：杜娟）