多智能体风险决策理论与方法研究综述

李 鹏[1)] 陈少飞[1)] 易楚舒[1)] 李 顺[1)] 兴军亮[2)] 陈 璟[1)]

[1)](国防科技大学智能科学学院 长沙 410073)

[2)](清华大学计算机科学与技术系 北京 100084)

摘 要 目前,学术界已有众多关于多智能体决策的研究,形成了一系列理论与方法,能够有效表达多个决策主体在合作、竞争等环境下的交互关系并求解得到合理的行为策略,在策略游戏、交通控制等诸多方面取得了成功应用。然而,在现实世界中,智能体在决策时可能会面临环境状态变化、自身方法误差等风险因素,使得智能体获得的损益值往往偏离预期值,而且其他智能体策略带来的非平稳性、应对风险的不同态度也会给该智能体带来进一步的决策挑战。因此,许多学者致力于研究多智能体风险决策理论与方法,为多智能体系统面临风险决策时提供合理的策略选择方案。本文针对当前关于多智能体风险决策理论和方法的研究工作进行系统性综述,首先形式化描述环境风险、智能体自身风险、其他智能体风险等三个方面风险来源,然后综述了基于最优控制理论、强化学习理论和博弈均衡理论等多智能体风险决策的理论与方法,最后总结了多智能体风险决策方法在人机协作、自动驾驶、交通控制和智能电网领域中的应用,并展望了多智能体风险决策研究中可能需要重点关注的五个开放性问题。

关键词 多智能体;风险决策;最优控制;强化学习;博弈均衡

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02338

Theories and Methods of Multi-Agent Decision Under Risk: A Survey

LI Peng[1)] CHEN Shao-Fei[1)] YI Chu-Shu[1)] LI Shun[1)] XING Jun-Liang[2)] CHEN Jing[1)]

[1)](College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073)

[2)](Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract Currently, the academic community has studied and designed a series of decision theories and methods for multi-agent systems. These theories and methods can effectively express the interactive relationships among multiple decision agents in cooperative, competitive, and other environments, and can determine reasonable behavioral policies. They have been successfully applied in areas such as strategic games and traffic control. However, when deployed in future real-world scenarios, agents may encounter numerous risk factors, such as fluctuations in environmental states and method error, which can cause the rewards obtained by the agents to frequently deviate from their expected values. Furthermore, the diverse risk management policies of other agents will also pose additional decision challenges. In recent years, numerous studies have focused on developing theories and methods for multi-agent decision making under risk, providing reasonable policy options for multi-agent systems facing risk. Therefore, this paper presents a systematic review of existing research on the theories and methods of multi-agent decision making under risk, including the theoretical foundations of individual and multi-agent decision making under risk, as well as research on multi-agent decision making methods under risk

收稿日期:2024-09-19;在线发布日期:2025-05-09。本课题得到国家自然科学基金(No.62376280)资助。**李 鹏**,博士研究生,主要研究领域为风险敏感强化学习、多智能体风险决策。E-mail:lipeng@nudt.edu.cn。**陈少飞**,博士,副教授,主要研究领域为计算机博弈、智能决策。E-mail:chensf005@163.com。**易楚舒**,博士研究生,主要研究领域为最优控制。**李 顺**,博士研究生,主要研究领域为多智能体强化学习。**兴军亮**,博士,研究员,主要研究领域为计算机博弈、智能博弈交互。**陈 璟**(通信作者),博士,教授,主要研究领域为智能决策、计算机博弈。E-mail:15576652503@163.com。

based on optimal control, reinforcement learning, and game theory. Finally, this paper summarizes the applications of multi-agent decision making methods under risk in the fields of human-machine collaboration, autonomous driving, traffic control, and smart grids, and also outlines five key open issues that may need to be addressed in future research on multi-agent decision making under risk.

Keywords　multi-agent; decision making under risk; optimal control; reinforcement learning; game equilibrium

1　引　言

多智能体系统由多个具有自主感知和决策能力的智能体组成，且智能体之间根据系统任务目标的不同往往产生合作、竞争等不同的交互关系。通过分散部署的方式，多智能体系统不仅能够有效弥补单个智能体在全局观测能力、复杂任务执行能力等方面的不足，而且具有更好的鲁棒性、容错性和自主性。当前，多智能体系统已经广泛应用在机器人集群[1-2]、电力系统优化[3-4]、传感网络[5-6]等复杂的现实世界场景中，有力推动了社会经济发展。

然而，多智能体系统在真实环境中部署时往往会面临各种风险，将直接影响系统任务的完成效果。“风险”在其最初的含义中表示在面临自然的、客观的危险时可能遭受的损失。在多智能体决策问题中，智能体面临的风险来源更广、范围更大，不仅包括智能体与环境[7-8]、智能体与智能体交互时的不确定性[9]，还包含决策方法本身的局限性和参数扰动等造成的不确定性[10-11]。此外，多智能体系统面临的风险决策问题相较于单个智能体风险决策问题更为复杂，因为系统中的智能体不仅需要权衡自身利益和环境风险，还要考虑风险存在时受其他智能体影响的任务目标达成。特别是近些年来，具身智能、大模型技术的发展和应用，更是突出了多智能体在开放环境下的风险决策需求，例如大模型在赋能多机器人进行路径规划时，需要更多关注如何应对不确定性带来的风险[12-13]。为此，本文将多智能体风险决策中风险的来源总结为三类：

(1)环境风险。环境风险主要是指智能体在执行各自策略后导致的结果会受到环境变化随机性和智能体获得奖励随机性的影响，且影响时机、影响程度及持续时间往往与智能体本身无关。为了应对这种结果的不确定性，目前产生了以整合式理论[14-16]、启发式理论[17-18]和统计学理论[19]为基础的风险应对模型和方法，为多智能体风险决策中智能体个体的风险决策提供理论支撑。

(2)智能体自身风险。智能体自身风险表现在智能体的策略安全上，由于决策方法误差、计算能力有限、传感器测量误差等原因造成智能体输出策略和期望策略间存在偏差，例如智能体自身对策略价值的估计不准[20]。针对智能体自身风险，研究者们从智能体个体决策方法出发试图探索一种能够控制智能体自身风险的方案，进而提出了针对神经网络的不可解释性[21]、大模型幻觉[22]、鲁棒策略学习[8]或安全策略学习[23]等方面的风险决策方法。

(3)其他智能体风险。其他智能体风险来源于其他智能体策略的不确定性对当前智能体策略执行结果的影响，即多智能体策略具有非平稳性。其他智能体风险是多智能体风险决策复杂于个体风险决策的最显著特征。无论合作还是对抗，智能体间的策略都会相互影响，如智能体在考虑自身利益、风险偏好的同时还要考虑集体利益和其他智能体可能采取的风险偏好。为了应对其他智能体风险，不同领域的研究者们从控制、学习和博弈三个层面形成了以稳定状态[24]、收益最大化[25]、策略均衡[26]为风险决策目标的三类多智能体风险决策方法，主要包括基于最优控制理论、基于强化学习理论和基于博弈均衡理论的方法。

尽管当前关于多智能体决策的研究涉及领域广、解决方案多，但缺乏对多智能体风险决策问题的系统综述。目前，关于不确定性的研究综述较多，例如神经网络不确定性[27]、深度学习不确定性[28-29]，但它们更关心不确定性的分类以及如何从方法上更好地规避或应对[30]，而忽略了不确定性导致的智能体在决策层面面临的风险困境。为此，本文围绕多智能体风险决策这一核心问题展开全面、系统的综述，首先围绕风险来源，总结了当前关于风险决策的基本理论、常见应对方法和风险度量；其次从控制、学习和博弈三个层面综述了多智能体在面临风险时遵

循的典型风险决策理论：最优控制理论、强化学习理论和博弈均衡理论，并分析了当前研究中对应三种理论的三类多智能体风险决策方法；最后概括了多智能体风险决策方法在人机协作等现实领域中的应用，并对可能存在的问题进行了展望和讨论。各部分之间关系如图1所示。

本文主要贡献为：(1)系统梳理并总结了不同领域中多智能体风险决策面临的三类共同风险来源，并以此为基础提出了多智能体风险决策的具体概念和统一形式化描述；(2)从控制、学习和博弈三个层次出发，提出了将当前不同领域内的研究分为以稳定状态、收益最大化和均衡策略为风险决策目标的三类多智能体风险决策理论和对应方法；(3)提出了未来在多智能体风险决策研究过程中需要重点关注的五个开放性问题，为后续研究提供参考。需要说明的是，由于多智能体风险决策涉及计算机科学、人工智能、控制科学、管理科学、心理学、经济学等众多领域和学科，具有极为显著的学科交叉特点，因此本文难以全面覆盖多智能体风险决策的相关技术，仅致力于为多智能体风险决策问题的研究进行系统性的讨论，并为各领域中面临的多智能体风险决策难题提供必要借鉴。

2 问题描述

本节首先介绍风险决策的基本概念与形式化表达，然后详细描述了多智能体风险决策的内涵并给出了多智能体风险决策的概念，最后分析了多智能体风险决策问题的挑战与相关特性。

2.1 风险决策基本概念

在学术界，目前并没有关于风险的统一定义和公认描述，一般认为风险就是事件发生的可能性及其后果的组合。美国经济学家弗兰克·奈特认为：风险是可测定的不确定性[31]。也有相关学者将风险定义为某一事件出现的实际状况与预期状况之间因不确定性导致了背离从而产生的一种损失[32]。然而，无论风险定义如何，人们普遍认同风险具有两个典型特征：客观性和不确定性。即风险是一种客观存在，根本原因在于影响决策的全要素之间存在复杂关系、不可控因素等带来的不确定性。

基于此，本文将风险定义为描述智能体决策过程中不确定性的可度量随机变量。换而言之，通过这个可度量随机变量，可以将风险表达成从概率空间到实数集的可测映射，具体可采用智能体获得损益值的均值、条件风险值[33]等对其进行度量及概率统计分析(详见第3.3节)。根据智能体概念模型中环境与智能体的关系，可以分别考虑环境不确定性和智能体决策方法误差两个方面的风险因素。

首先，可以将环境风险看成是智能体在环境中执行某个策略所获得的随机性结果。例如，我们可以用某个概率分布P来描述执行策略υ后所导致结果X的随机性：

$$P\left(X(\upsilon)=x_k\right)=p_k, k=1,2,\cdots,d \tag{1}$$

其中，d表示可能产生的结果个数，x_k和p_k分别表示可能出现的第k个结果以及对应的发生概率。例如，在图2所示的一次性决策问题中，智能体为了达到终点，可以选择最近的路线(即策略A)移动，有80%的概率能够顺利快速通过(比如耗时5分钟)，但是有

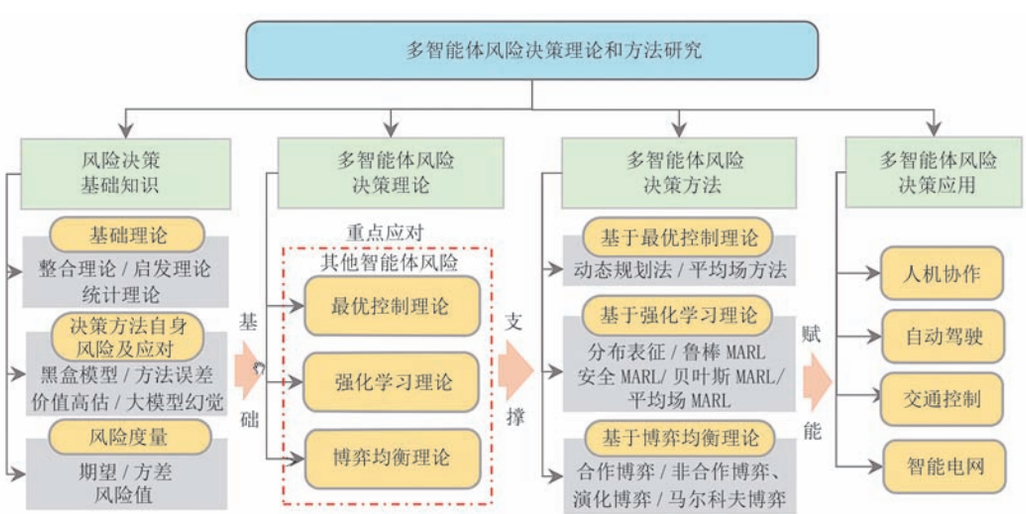

图1 本文组织架构

20%的概率因道路维修等原因延迟很久后通过（比如耗时60分钟）。即策略A的结果可以表示为$x_1=5$分钟，$p_1=80\%$，$x_2=60$分钟，$p_2=20\%$。与这种一次性决策问题所不同，序贯决策问题（如马尔科夫决策过程问题）中的环境风险可能是由状态转移函数或者是奖励函数的随机性引起的，使得策略执行结果X所服从的概率分布P变得更为复杂。此外，有时概率分布P也存在动态变化的情况，如环境中存在一些干扰因素，且出现的时机和幅度都是动态的。

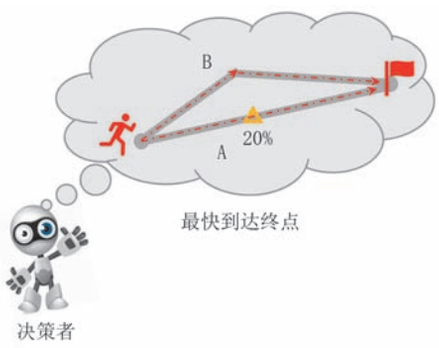

图2　环境风险下的智能体策略选择问题

然后，可以将智能体自身风险看成是由于决策方法误差等原因造成智能体输出策略$\tilde{\upsilon}$和预期策略υ间存在的偏差（即输出策略υ具有一定随机性）。与环境风险类似，同样可以将这种智能体自身风险表示成执行策略υ时智能体获得如公式（1）表示的结果X的随机性，只是造成这种随机性的原因变成了决策方法等智能体自身原因。因此，本文将智能体在环境风险或决策方法风险下进行风险度量或策略制定的问题统称为风险决策问题。

2.2　多智能体风险决策问题描述

当决策主体扩展至2个及以上时，风险决策问题成为了具有多个参与者的风险决策问题，决策者在决策时需要同时考虑其他决策者的策略选择。而且，决策者既可以是人也可以是机器，例如多人协商决策应对突发事件、人机协作诊断治疗病患恶疾、多无人机合作打击敌对目标等。从近几年的研究中我们发现，在具有多个参与者的风险决策中无论决策者是以人为主还是以机器为主，风险决策的核心依旧聚集于决策者在多个策略中如何选择的过程。因此，本文将多人、人机、多机面临的风险决策问题统称为多智能体风险决策问题。

具体地，本文将其他智能体风险看成是在环境风险基础上，当多个智能体采用合作/竞争/混合等方式进行策略选择时，其他智能体策略的不确定性带来的执行结果的不确定性。此时的结果不确定性可以由式（1）扩展表示为多个智能体执行联合策略$(\upsilon_i,\upsilon_{-i})$后所导致的结果$X$服从概率分布$P$：

$$P\left(X(\upsilon_i,\upsilon_{-i})=x_k\right)=p_k,k=1,\cdots,d \tag{2}$$

其中，υ_i表示智能体i的策略，υ_{-i}表示除智能体i外其他智能体的联合策略。

与个体风险决策相比，多智能体风险决策需要考虑其他智能体策略υ_{-i}对未来结果的影响。如图3所示，智能体为了最快到达目标就需要考虑对方是如何选择的，并在此基础上选择一条能够尽可能先于对方到达的路线，然而最短路径总是风险最大。例如，当智能体1选择距离最短但风险较大的道路a1时，智能体2可能会采取规避风险的态度选择道路b2，尽管b2距离较长但遭遇道路维修的概率较低。当距离a2＜b2时且智能体1选择了道路a2，那么智能体2为了最快到达终点可能采取风险寻求的态度选择道路b1。因此为了最快到达终点，智能体就需要考虑对方如何进行风险决策。

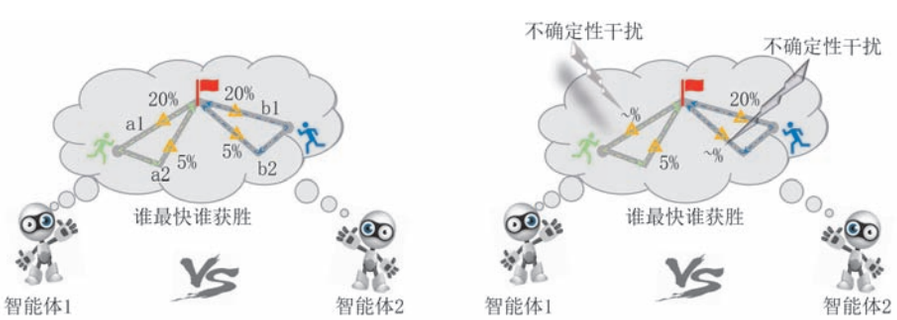

图3　多智能体风险决策问题（既包含多个智能体在面对结果概率分布P确定情况下的策略选择问题（左图），也包含P受到干扰而变化时的策略选择问题（右图）。其中两个智能体的行进速度相同，道路长度相同（a1＝b1，a2＝b2））

多智能体风险决策继承了个体风险决策的概念，既包含智能体在考虑其他智能体如何决策后对P已知的策略的选择，也包含在不确定性影响下，在考虑其他智能体如何风险决策后对P变化的策略的

选择。其他智能体风险表现为其他智能体策略υ_{-i}对智能体i的策略υ_i的影响，进而导致联合策略$(\upsilon_i,\upsilon_{-i})$作用下的结果$X$具有不确定性。

2.3 多智能体风险决策问题挑战及特性分析

尽管不同领域内多智能体面临的风险决策问题各有不同，但总的来说当前面临的主要挑战包括：

(1)风险来源广。智能体决策时需要同时考虑环境和其他智能体的风险，并且风险来源、介入时间均难以确定；

(2)利益权衡难。多智能体场景下每个智能体需要综合考虑个体利益、集体利益和风险，如何在风险环境下权衡个体利益和集体利益的冲突是多智能体风险决策的一个重要挑战；

(3)演进方向不可控。风险的动态变化和不可预测加剧了智能体间策略的非平稳性，使得多智能体风险决策时智能体对当前自身策略和环境演进方向间的关系很难把握，从而对可能造成的后果和损失难以估计。

基于上述分析，本文总结出三点多智能体风险决策不同于个体风险决策的主要特性

(1)异质性：多智能体系统中智能体可能具有相同或不同的属性或结构，但通常具有不同的功能使命、观测信息和风险偏好。同时，由于智能体硬件区别、功能限制、立场不同，其能够获取的信息往往存在差异，因此拥有不同观测信息的智能体往往会根据其目标做出具有不同风险偏好的决策[34]，从而导致多智能体风险决策中普遍存在的异质性特征。

(2)不确定性：决策问题受智能体、环境等诸多因素影响，而其中每一种因素都有可能被干扰，从而使决策面临风险。多智能体风险决策中不确定性来源更加广泛，包括智能体自身状态、情绪的不确定性[35]，其他智能体的态度、目标不确定性[36]，环境变化的不确定性等都有可能对决策产生负面影响，这也是多智能体风险决策问题中最典型的特征。

(3)演化性：由于决策的连续性和长期性，多智能体在进行风险决策时并非按照固定模式在静态环境下决策，而是随着风险的变化动态调整自身决策目标、联盟对象、风险偏好等[37-38]，进而表现出多智能体风险决策的演化特性。例如在群体投资决策中，随着健康损害的增加，可能导致智能体在财务选择和投资组合上的风险偏好发生变化[39]。

3 风险决策基础知识

3.1 风险决策基础理论

在当前的研究中，具有代表性的个体风险决策理论包括整合式理论、启发式理论和统计学理论，如表1所示。

表1 个体风险决策理论

类别	代表性理论	方法	文献
整合式理论	期望效用	效用值最大化原则 边际效用递减	[40]
	前景理论	决策权重代替概率 “框定-编辑-评估”过程	[15]
	累积前景	决策权重与概率、结果大小的顺序有关	[41]
启发式理论	占优启发式	比较两个选项的结果并根据情况选择合适选项	[17]
	齐当别	在差异明显的维度上选择优势方案	[18]
统计学理论	贝叶斯	基于概率计算风险期望	[19]

3.1.1 整合式理论

整合式理论假设智能体在对各种信息进行整合后，选择一个总体价值或效用最大的策略[42-44]，主要包括期望价值理论、期望效用理论[40]、前景理论[15]以及累积前景理论[41]等一系列理论模型。

期望价值理论认为决策者最关注的是不同方案的期望价值，例如对于不同的赌博游戏决策者会先计算每种游戏对应收益的数学期望，然后选择期望值最大的一种游戏参与。另一些研究者主张采用主观效用代替客观价值，提出边际效用递减的概念。Neumann等[44]在主观效用的基础上提出了期望效用理论，该理论的决策目标定义为：$EU(\cdot)=p_1u(x_1)+\cdots+p_nu(x_n)$，其中$EU(\cdot)$表示期望效用函数，$u(x_i)$表示结果$x_i$对应的效用函数，且概率满足$p_1+\cdots+p_n=1$。期望效用理论中决策者在风险决策时遵循效用值最大化的原则。随后，研究者们在期望效用理论的基础上又发展出了多种变式，例如Savage等[45]主张用主观概率代替客观概率，提

出了主观期望效用理论，即决策者在风险决策时考虑的是收益的效用和收益的主观概率的乘积。预期效用理论[46]和心理效用理论[47]等都是从不同角度对期望效用理论的进一步完善。

期望效用理论提供了一种策略选择框架，但并不能有效解释人类的某些决策行为，如阿莱悖论。为此，研究者提出了前景理论[15]，主张用决策权重代替概率，并假设在对决策问题正式“估计”前有一个“框定”和“编辑”的过程。前景理论的决策目标函数定义为：$V\left(x_1,p_1;\cdots;x_n,p_n\right)=\sum_{i=1}^{n}\omega(p_i)v(x_i)$，其中$V(\cdot)$表示前景的整体价值。$\omega(p_i)$为概率权重函数，表示决策者对概率的感知，$v(x_i)$表示结果$x_i$对应的主观价值函数。为了在多目标决策及具有不确定性的环境下应用，Tversky等[41]在前景理论的基础上进一步发展出了累积前景理论，即采用了更为复杂的权重函数，认为决策权重不仅与概率有关而且与结果大小的顺序相关。当前，前景理论和累积前景理论已逐渐取代期望效用理论成为了主流的风险决策理论。

3.1.2　启发式理论

启发式理论假设决策者缺乏对各种信息进行整合的能力，而是采用各种启发式或经验法则来进行决策[48-49]。启发式理论主要包括占优启发式[17]和“齐当别”模型[18]。

Brandstatter等[17]提出占优启发式模型，该模型并不假设策略价值和概率间存在复杂的转换关系，而是利用一套更简洁的启发式原则来指导决策者进行风险决策。在有限理性的假设下，李纾等[18]构建了“齐当别”风险决策模型，认为在进行风险决策时决策者会故意忽略那些在某几个维度上差异不大的备选方案，即进行“齐同”，而在那些差异明显的维度上则会选择具有优势的方案。然而，该模型需要决策者在决策时能够获取大量的信息，使其能够在行为上实现“齐同”，而在决策上实现“区别”。

3.1.3　统计学理论

统计学理论则主要采用贝叶斯方法帮助决策者理解和量化风险，通过结合先验知识、观测数据和后验概率进行推断。风险决策中贝叶斯方法可以有效综合模型信息、数据信息和先验信息，以获得最优决策方案，在处理不确定性高、信息不完全的风险决策问题上具有天然的优势。

贝叶斯风险决策方法[19]常通过随机试验获得观测值$\hat{\theta}$，$p(\theta)$为自然状态θ的先验概率分布，然后根据条件分布函数，利用贝叶斯公式计算后验概率分布$p(\theta\left|\hat{\theta}\right.)$。贝叶斯风险决策的目标函数为：$R(\pi)=\mathbb{E}\left(J(\theta,\pi(\hat{\theta}))\right)$，其中$\pi(\hat{\theta})$为决策函数，$J(\theta,\pi(\hat{\theta}))$为损失函数。贝叶斯风险决策方法从概率的角度给出了风险的期望值，常常通过结合样本信息和先验信息推断后验分布，从而帮助决策者做出风险敏感策略。

此外，除了上述三种典型的风险决策理论外，还有一部分学者从生物学、心理学等领域进行研究，以更深入地探索风险决策理论的机理。例如，从生物学角度出发，风险决策的眼动研究重点考察眼动过程与选择偏好之间的关系[50]。此外，还有研究揭示注视转换在预测决策者选择策略方面的优势[51]。Van等[52]发现当选项中存在互相冲突的属性信息时，采用非补偿性策略进行决策的决策者会表现出更大的背内侧前额叶激活状态，这为风险决策的策略选择过程提供了神经层面的证据。

3.2　决策方法自身风险及应对

决策方法由于原理上存在的固有问题，智能体在使用这些方法进行决策时就存在风险，比如神经网络黑盒模型、强化学习价值高估等都会带来决策风险。此外，随着大模型的涌现，采用大模型指导智能体进行决策的方法也同样使智能体面临着决策方法本身的风险。这四类风险虽然主要在单智能体场景下进行了探讨，但并不局限于单智能体场景，作为最基础的决策方法风险在大部分多智能体风险决策方法中均存在。由于这四类风险均属于方法在原理层面存在的问题，并不涉及多智能体间的交互特性，因此本节主要对不同的风险决策方法自身面临的共性风险进行总结，如表2所示。

3.2.1　神经网络黑盒模型

随着神经网络的快速普及，使用神经网络进行决策已成为大多数方法的基础。神经网络以其卓越的函数逼近能力拟合感知输入和策略输出间的关系。然而，神经网络的可解释性不足给实际决策活动带来了潜在风险。这种不可解释性主要源于神经网络作为“黑盒”模型的本质，即其内部机制和决策过程对决策者而言并不透明[21]。

为了缓解不可解释性带来的决策风险，研究者们提出了多种解决方案。Gal等[53]通过集成统计模型来估计神经网络的不确定性，这对于需要在决策中量化不确定性的安全应用至关重要。Lundberg

表2 决策方法自身风险及应对

方法风险	风险危害	应对策略	具体方案
神经网络黑盒模型	决策过程缺乏透明性	提升神经网络可解释性	集成统计模型[53] 特征预测[54-55] 全局解释转化[56]
机器学习方法误差	难以逼近真实价值	风险最小化	SVM方法[57] 径向风险最小化[58] 经验风险函数[59-60] 倾斜经验风险最小化[61]
强化学习价值高估	策略趋向次优解或有害解	优化Q值学习过程	高估限制[62-63] 集成[64-66] 策略优化[67-68] 算子替换[20,69-71]
大模型幻觉	依赖不可靠决策信息	评估、预防、减少大模型幻觉	幻觉评估[72] 外部知识注入[73-74] 检索增强[75] 幻觉奖励[76]

等[54]采用SHAP方法，为每个特征分配一个预测重要性值，提供了一个统一框架来解释复杂模型的预测，并具有更优的计算性能和一致性。Pedreschi等[55]提出了一种从局部到全局的框架，使用逻辑规则的语言来表达，通过审计目标实例附近的黑盒来推断局部解释，进而将多个局部解释优化为简单的全局解释。Ghorbani等[56]通过提供特征重要性映射来解释神经网络的预测，有助于理解哪些特征对模型决策的影响最大。这些方法各具特点，但共同目标是提升神经网络的透明度和可解释性，从而降低决策方法带来的风险。

3.2.2 机器学习方法误差

机器学习方法中大部分的学习模型遵循概率近似正确（Probably Approximate Correct，PAC）的框架。PAC框架中通常假设学习样本是独立同分布的，且该分布是固定和未知的。在此假设基础上，也逐渐演化出了学习器的误差表示方法：泛化误差、经验误差和结构误差，也被称为风险、经验风险和结构风险。泛化误差（风险）衡量的是学习器在整个数据分布上的期望损失，但由于获取完整数据分布的难度，实践中的学习器通常在有标签的样本集（实际可用的数据集）上优化经验误差（经验风险）。为了防止学习器过拟合，当前研究中通常在经验风险的基础上加上一个正则化项，并引入了结构误差（结构风险）。因此，PAC假设样本的独立同分布造成的学习器风险也是当前大多数方法面临的问题。

当前为了降低机器学习方法误差带来的风险，有研究从风险识别与评估、模型鲁棒性等方面展开，其中最直接的思路就是尽可能降低结构风险和经验风险。目前最流行的机器学习算法SVM就是一种基于统计学习理论和结构风险最小化原理的鲁棒监督技术[57]。针对经验风险最小化问题，Matthew等[58]针对损失函数具有较大的Lipschitz模量与伪尖锐极小值问题，提出了径向风险最小化方法，它考虑了参数空间中邻域内最坏情况的经验风险。Samir等[59]在假设有限测度条件下提出了具有相对熵正则化的经验风险最小化方法。在固定数据集和特定条件下，当模型从该方法的解中采样时，经验风险被证明是一个亚高斯随机变量。Ibrahim[60]证明了在构建电容模式分类和电容式词袋过程中使用电容经验风险函数能减少过拟合，提高了所考虑模型的泛化能力。Li等[61]将统计学、概率论、信息论领域中常见的指数倾斜技术融入经验风险最小化问题中，使用指数倾斜来灵活调整损失造成的影响，并且能够增加或减少异常值的影响，具有方差减少特性，以实现公平性或鲁棒性。

3.2.3 强化学习价值高估

强化学习通常采用时间差分方法估计策略价值，但在训练中可能面临着对动作价值严重高估的威胁。在学习过程中，强化学习方法若对某些低价值动作产生高估，则会在后期学习中不断强化算法在面临相似状态下对该动作的选择，从而导致策略趋向次优解。同时，由于高估误差的存在，算法在训练过程中往往存在学习曲线严重振荡、突降等不稳定现象。例如基于值的Q学习算法[62]通过迭代优化动作值来学习策略，但由于在更新目标动作价值

是采用了自举更新的方法，并引入了最大化算子，所以产生了高估风险。基于此，研究者们做了大量卓有成效的工作，试图解决这高估风险造成的算法学习不稳定的问题。

Thrun等[63]在理论上分析了Q学习因函数近似导致的高估误差，并给出了Q学习高估误差期望的上界。当前研究中通常采用平均或集成多个Q值的方法降低高估误差。例如平均DQN算法[64]认为目标近似误差的存在导致了学习过程中稳态策略的偏移，因此利用多个历史目标Q网络的输出的平均值代替TD目标值，从而降低目标近似误差的方差。Double DQN算法[65]通过对目标网络进行微调，将动作选择和价值估计解耦，从而降低高估误差。TD3算法[66]结合双重网络、策略平滑和延迟更新等多种方法降低高估误差。

在多智能体决策中，MATD3算法[67]关注策略梯度方法，将TD3算法扩展至MADDPG算法[68]中，采用集中的双重评论家减少高估。Pan等[20]则通过添加正则化项惩罚偏离基线的Q值，并采用Softmax算子计算目标Q值。另一种方式是直接替换价值估计时的Max算子。例如DeepMellow算法[69]采用可微的Mellowmax算子代替Max算子，它在任何温度参数设置下都是凸的，能有效缓解Max算子导致的高估。Soft Mellowmax算子[70]解决了Mellowmax对参数敏感且容易出现过渡平滑的问题。Zhao等[71]在DQN中将Max算法替换为Softmax算子，在Arcade环境中取得了很好的性能。

3.2.4　大模型幻觉

近些年来，大模型逐渐用于指导或协助智能体决策。大模型决策带来的风险往往源于大模型幻觉。大模型幻觉指模型生成的内容不是基于事实或准确信息的现象，即智能体的感知输入与模型的策略输出之间的矛盾，这一现象通常出现在大语言模型和大视觉-语言模型中。幻觉通常表现为模型无意产生的似乎合乎逻辑的内容，可能由多种因素造成，如暴露偏差、缺乏实时信息访问通道、对上下文准确理解的限制等[77-78]。在利用大模型辅助决策的场景中，大模型幻觉下的错误断言将以智能体自身风险的形式干扰决策过程，使得智能体过度依赖不可靠信息，对决策的可靠性和安全性构成威胁。

在临床医学、金融等高风险的决策领域中，如何检测幻觉以及采取措施预防和减少幻觉现象至关重要。在幻觉评估方面，Guan等[72]提出了首个的幻觉评估基准HallusionBench，系统地剖析大视觉-语言模型的多种幻觉类型。Sebastian等[79]提出了采用语义熵检测大语言模型的幻觉，其中语义熵建立在不确定性估计的概率模型基础上，用来克服混淆策略带来的影响。在缓解幻觉方面，从智能体层面出发，可以通过诱导或注入外部知识减轻模型幻觉。Martino等[73]设计了一种知识注入的策略，将与文本生成任务相关的上下文数据从知识图谱映射到文本空间，减少了错误断言输出的数量，提升信息正确性。Zhang等[74]提出了知识对齐问题，通过引入与人类用户和知识库交互的框架MixAlign，将生成的文本与相关的事实知识对齐，使得智能体能够使用户能够交互式地指导模型的响应。此外，Bhaskarjit等[75]通过结合检索增强生成技术和元数据的方式，提升从收益报告转录文本中提取信息的效率。从决策过程而言，Zhang等[76]提出从智能体反馈阶段的强化学习中设计一个针对幻觉的特定奖励评分，并通过强化学习优化缓解幻觉现象。

3.3　风险度量

风险度量是风险决策中评估和量化风险的重要工具。在投资决策、金融风控等领域对风险的精准评估和度量是制定高质量策略的关键一环，而在智能决策中进行风险度量是学习风险敏感策略的前提。在不同领域内由于决策任务的差异对风险的度量方式存在较大差异，但大多都是从表示风险的概率分布P出发，基于概率特性或风险价值来度量风险，最常见的方式包括采用期望/方差、风险值等对风险进行度量。

3.3.1　期望/方差

风险决策中我们采用概率分布P来表示智能体执行策略s后带来的结果X的随机性，因此智能体在选择策略s时就面临着一定风险。而度量该风险最直接的方法就是评价其分布P，因此大多数方法采用统计学中的概率特性来度量风险，例如期望$E(X)$或方差$D(X)$来度量风险。最优控制中常常将风险表示为一种随机过程并在系统动力学中给予考虑，然后采用期望的上下界来度量风险，而博弈论中的均值-方差均衡则采用方差度量风险，当策略的期望收益相同时，方差越大则风险越大。特别的，机器学习方法中常常将风险作为优化要素予以考虑，并不直接度量风险，而是采用误差期望值的大小表示风险带来的损失，因此本文也认为其属于期望/方差的风险度量方式，这在鲁棒强化学习和安全强化学习中最为常见。

3.3.2 风险值

风险值是一种统计风险的度量，常用于在特定的时间范围和给定的置信度水平上估计策略的潜在损失。风险值主要包括风险价值（VaR）和条件风险值$CVaR$。VaR和$CVaR$属于金融领域中两种重要的风险度量工具，VaR_α主要用于衡量一定置信水平α下的最大可能损失，是一个点估计，而$CVaR_\alpha$主要衡量损失超过VaR_α阈值时平均损失，是一个区间估计，能够更全面地反映尾部风险。具体而言，给定一个置信度变量平α，VaR_α的计算公式为：$VaR_\alpha(P)=\inf\{p\in R:F_P(p)\geqslant\alpha\}$，$CVaR_\alpha$的计算公式表示为：$CVaR_\alpha(P)=\mathbb{E}\{p|p\leqslant VaR_\alpha(P)\}$，其中$F_P$表示分布$P$的累积分布函数。除此之外，风险值还可以从不同角度进行度量，例如采用模糊程度[80]或场强大小[81]。

4 多智能体风险决策理论

个体风险决策更加强调个体行为，主要关注环境风险或智能体自身风险对决策的影响，而多智能体风险决策问题在环境风险的基础上更要考虑其他智能体的风险，这并不是多个个体风险决策问题的简单叠加。在研究多智能体交互关系的相关文献中，整体上围绕控制、学习和博弈三个层面展开，分别探索基于动力学建模、基于策略学习优化和基于耦合约束的多智能体系统决策模式。尽管所属领域不同，但在当前的研究中早已呈现出“你中有我，我中有你”的交叉态势。因此为了更清晰地理解多智能体风险决策问题，本文从智能体的控制、学习和博弈三个层面出发，结合多智能体风险决策时的目标导向，将当前的多智能体风险决策理论分为以稳定状态、最大化收益、均衡策略为决策目标的三类风险决策理论：(1)基于最优控制理论；(2)基于强化学习理论；(3)基于博弈均衡理论，如图4所示。其中，最优控制理论用于在风险条件下制定多智能体系统的控制策略，重点研究基于动力学建模的多智能体系统的动态演化过程，并利用动态规划、平均场等方法求解多智能体风险决策时的最优控制策略。强化学习理论主要描述多个智能体间的合作关系，通过以最大化集体利益为目标学习最优风险敏感策略。博弈均衡理论则强调每个智能体为了个体利益或集体利益在风险决策过程中的博弈关系，目标是获得多智能体风险决策时的均衡策略。侧重点上，控制理论提供了基础的决策和执行框架，重点关注多智能体系统的精确性和稳定性，而强化学习理论为智能体提供了从经验中学习和适应环境的能力，博弈论则重点处理智能体之间的互动和策略选择带来的冲突与合作关系。

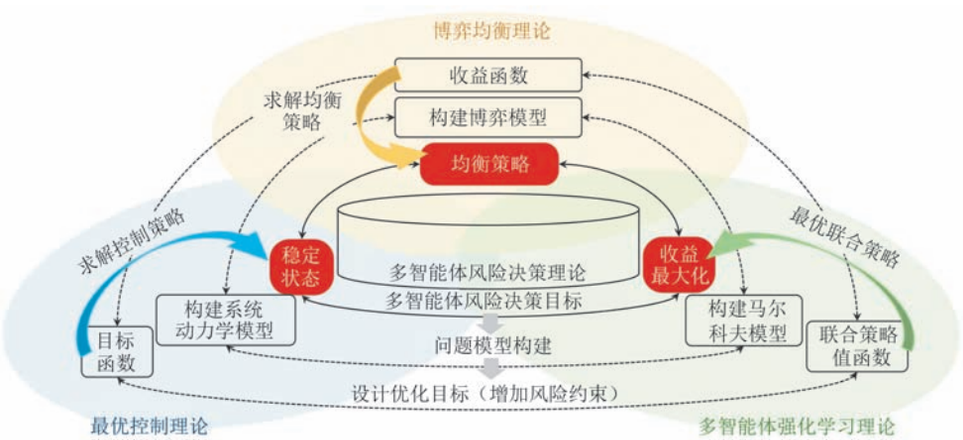

图4 本文围绕控制、学习和博弈三个层面总结的三类多智能体风险决策理论

4.1 基于最优控制理论

最优控制理论将多智能体风险决策过程建模为多智能体动力学方程，进而求解使多智能体系统在风险条件下能够保持稳定状态的最佳响应策略。最优控制常采用随机动力学方程建模并求解多智能体风险决策问题，其中风险通常以随机项（例如高斯分布）的形式被考虑。实践中，风险通常体现在系统的模型参数中[82-85]，通过对参数的干扰，模拟多智能体决策时面临的环境或其他智能体风险。

不失一般性地，本文提出一个存在合作和竞争关系的普适性多智能体最优控制模型，包括智能个体和N个与其构成合作或竞争关系的其他智能体。

对于$N+1$个智能体建立系统动力学方程：

$$\begin{cases} \mathrm{d}x_t^i = f^i\left(x_t^i, \boldsymbol{X}_t, u_t^i, t\right)\mathrm{d}t + \theta_i\sigma_i(t)\mathrm{d}\omega_t^i \\ s_0^i = s_{i,0} \\ i = 0,1,2,\cdots,N, t\in[0,\mathcal{T}] \end{cases} \tag{3}$$

其中，$f^i(\cdot)$是第i个智能体的激活函数，$\boldsymbol{X}_t=\left(x_t^0,\cdots,x_t^N\right)$表示状态向量，$u_t^i$表示智能体的控制输入，$\{\omega_t^i\in\mathbb{R}^r\}$表示定义在滤波概率空间$\left(\Omega,\mathcal{F},\left(\mathcal{F}_t^{[N]}\right)_{t\in\mathcal{T}},\mathbb{P}\right)$上的标准维纳过程（物理上表述为布朗运动），其中$\Omega$表示样本空间，$\mathcal{F}$表示$\sigma$-代数，$\left(\mathcal{F}_t^{[N]}\right)_{t\in\mathcal{T}}$表示滤波，$\mathbb{P}$则表示概率测度，$\sigma_i(t)$是关于$\mathcal{F}$可测的随机变量。$\theta_i$为0/1变量，$\theta_i=1$表示智能体$i$面临风险，反之$\theta_i=0$表示无风险。

在涉及到大量智能体甚至趋于无限多智能体的最优控制问题中，单个智能体对系统的影响往往可以忽略不计，起到决定性作用的是个体与群体决策行为之间的交互。平均场理论是解决这一类问题的基础。值得注意的是，此时多智能体系统动力学方程发生变化。数量趋于无限的多智能体系统中同时存在同质智能体和异质智能体，且由于异质智能体具有不同的特性、策略，在面对相同态势时决策模式有所差异。因此本文将多智能体分为$\Lambda<\infty$个不同的类型，第λ个类型中有N_λ个智能体，每一组类型的动力学方程的激活函数都对应不同的模型参数。记$\mathcal{I}_\lambda$为第λ个类型中的智能体的指标集，即基数满足$|\mathcal{I}_\lambda|=N_\lambda$，此时平均场动力学方程表示为

$$\begin{cases} \mathrm{d}x_t^i = f^i\left(x_t^i, \bar{\boldsymbol{X}}_t, u_t^i, t\right)\mathrm{d}t + \theta_i\sigma_i(t)\mathrm{d}\omega_t^i \\ s_0^i = s_{i,0} \\ i = 0,1,2,\cdots,N, t\in[0,\mathcal{T}] \end{cases} \tag{4}$$

其中，$\bar{\boldsymbol{X}}_t=\left(\bar{x}_t^1,\cdots,\bar{x}_t^\Lambda\right)$表示在时刻$t$的种群平均场，其中$\bar{x}_t^\lambda,\lambda=1,\cdots,\Lambda$代表第$\lambda$个类型在时刻$t$的平均场，定义为

$$\bar{x}_t^\lambda = \lim_{N_\lambda\to\infty}\frac{1}{N_\lambda}\sum_{i\in\mathcal{I}_\lambda}x_t^i \tag{5}$$

4.2　基于强化学习理论

多智能体强化学习（Multi-agent reinforcement learning，MARL）将多个智能体的决策过程建模为马尔科夫模型，其中多个智能体通过与环境的不断交互学习最优策略，以获得最大累积奖励[86]。以元组形式定义涉及N个智能体的马尔科夫决策过程：$M=\left(N,\hat{\mathcal{S}},\mathcal{A},R,P,\mathcal{Z},\mathcal{O},\gamma\right)$。其中$\hat{\mathcal{S}}$表示状态空间，$\mathcal{A}=A_1\times\cdots\times A_N$为联合动作空间。具体地，对于每一个智能体，在当前观测下都有c个动作可以选择，即第i个智能体的动作空间为$a_i\in\left\{a_i^1,\cdots,a_i^c\right\}$。联合动作表示为$\boldsymbol{a}=(a_i,\cdots,a_N)$。$R:\hat{\mathcal{S}}\times\mathcal{A}\to\mathbb{R}^N$是联合奖励函数，定义第$i$个智能体在时间$t$返回的奖励为$r_i^t$，联合奖励为$r^t$。$P\left(\cdot|\hat{s},\boldsymbol{a}\right)$是马尔科夫状态转移概率函数，$\gamma$是折扣因子。在部分可观设置中，智能体观测为$o\in\mathcal{O}$，该变量由观察函数$\mathcal{Z}(\hat{s},i):\hat{\mathcal{S}}\times N\mapsto\mathcal{O}$确定。以$\tau_i\in\mathcal{T}\equiv(\mathcal{O}\times\mathcal{A})^*$表示第$i$个智能体历史的观测-动作数据。随机策略$\pi_i(a_i|\tau_i):\mathcal{T}\times\mathcal{A}\mapsto[0,1]$表示智能体$i$在$\tau_i$下选择动作$a_i$的概率。

风险条件下多智能体强化学习的目标同样是以最大化收益为目标来学习联合策略$\pi=(\pi_i,\pi_{-i})$，但最重要的差异就是在策略学习过程中将风险作为策略优化的关键项，其中π_{-i}表示除去智能体i的其他智能体策略。驱动策略搜索的关键目标函数通常有两类，其一为联合价值函数：

$$V(\pi,\hat{s}) = \mathbb{E}\left(\sum_{t=0}^{\infty}\gamma^t r^t|(\pi_i,\pi_{-i}),\hat{s}\right) \tag{6}$$

其二为联合动作价值函数：

$$Q(\pi,\hat{s},\boldsymbol{a}) = R(\hat{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\hat{s}'}\left(V(\pi,\hat{s}')\right) \tag{7}$$

此外，与估计价值函数的期望值不同，通过估计价值函数的分布来捕捉多智能体决策时面临的风险具有天然的优势。其中价值分布MARL算法通过建立新的随机变量$\boldsymbol{Z}(\pi,\hat{s},\boldsymbol{a})$来描述随机回报，其期望值为动作价值函数$Q(\pi,\hat{s},\boldsymbol{a})$，即$Q(\pi,\hat{s},\boldsymbol{a})=\mathbb{E}\left(\boldsymbol{Z}(\pi,\hat{s},\boldsymbol{a})\right)$。$\boldsymbol{Z}(\pi,\hat{s},\boldsymbol{a})$在形式上用递归方程描述，但与$Q(\pi,\hat{s},\boldsymbol{a})$不同的是$\boldsymbol{Z}$具备分布性质[87]。该特性为智能体根据回报分布做决策提供了理论基础，使得决策模型能够灵活地调整智能体风险偏好。

为应对不同来源的风险，通过在标准MARL模型基础上引入风险度量或风险偏好能够极大提高风险决策的稳健性和安全性。例如，对于第i个智能体在联合策略π下的值函数V_i，考虑风险度量CVaR_α时目标就是学习一种在最大化全局收益下控制风险的策略：

$$\pi_i^* = \arg\max_{\pi_i}\mathrm{CVaR}_\alpha(V_i) \tag{8}$$

4.3　基于博弈均衡理论

均衡是博弈论中的一个重要概念，通常在多智能体系统中表示一种平衡状态，即在均衡的联合策略下没有智能体有动机改变自身策略。当前常用均衡来刻画单个智能体对其他智能体策略的最佳响

应，是一种分析多智能体决策问题的通用框架。为此，本节从博弈均衡角度出发，对当前多智能体风险决策研究中的博弈均衡相关理论进行概述。

具体地，设有N个智能体参与博弈，$[N]\backslash i$表示除第i个智能体以外的其他智能体。定义第i个智能体可能的纯策略集为Ψ_i，$\psi_i\in\Psi_i$是第i个智能体的纯策略。记ψ_{-i}为智能体$[N]\backslash i$的纯策略，则联合纯策略为$\psi=(\psi_1,\cdots,\psi_N)=(\psi_i,\psi_{-i})$。对于任意集合$\Psi_i$，定义$\Sigma_i$为所有在$\Psi_i$上的概率分布的集合，$\Sigma_i$为第$i$个智能体的混合策略集。第$i$个智能体的混合策略为$\rho\in\Sigma_i$，$\rho_{-i}$为智能体$[N]\backslash i$的混合联合策略。$\rho(\psi_i)$表示第$i$个智能体使用策略$\psi_i$的概率。给定混合联合策略$\rho_{-i}$，智能体$i$在使用策略$\psi_i$时的目标函数为$\overline{obj}_i(\psi_i,\rho_{-i})$。利用全概率定理，均值函数为：

$$\overline{\text{mean}}_i(\psi_i,\rho_{-i})=\mathbb{E}\left(\overline{obj}_i(\psi_i,\rho_{-i})\right)=\sum_{\psi_{-i}\in\Psi_{-i}}\left(\text{mean}_i(\psi_i,\psi_{-i})\cdot\rho(\psi_{-i})\right) \tag{9}$$

基于博弈均衡的多智能体风险决策理论按照目标函数可以分为两类：一是当目标函数为效用、收益或随机回报函数时，记$\overline{obj}_i(\psi_i,\rho_{-i})=U_i(\psi_i,\rho_{-i})$，则对于给定的混合联合策略$\rho_{-i}$，第$i$个智能体的最佳响应纯策略集合为：$\arg\max_{\psi_i\in\Psi_i}P\left(U_i(\psi_i,\rho_{-i})\geqslant U_i(\Psi_i\backslash\psi_i,\rho_{-i})\right)$；二是当目标函数为成本或延迟函数时，记$\overline{obj}_i(\psi_i,\rho_{-i})=J_i(\psi_i,\rho_{-i})$，则对于给定的混合联合策略$\rho_{-i}$，第$i$个智能体的最佳响应纯策略集合为：$\arg\max_{\psi_i\in\Psi_i}P\left(J_i(\psi_i,\rho_{-i})\leqslant J_i(\Psi_i\backslash\psi_i,\rho_{-i})\right)$，其中最佳响应混合策略集为该纯策略集上所有概率分布构成的集合。而按照均衡类型则可以划分为三类常见的风险决策均衡[88-89]。

(1)纳什均衡中每个智能体在考虑其他智能体的策略时，都会选择使自己期望效用最大化的策略。以成本作为目标函数时，对于给定的混合联合策略ρ_{-i}，第i个智能体对ρ_{-i}的最佳响应：

$$\psi_i^*=\arg\min_{\psi_i\in\Psi_i}\overline{\text{mean}}_i(\psi_i,\rho_{-i}) \tag{10}$$

在随机博弈中，考虑预期收益的纳什均衡可能会产生方差较大的风险。尽管如此，由于纳什均衡的普适性及均衡解存在性定理的理论支撑，目前仍然是最常见一类均衡。

(2)均值-方差均衡通过同时考虑决策结果的期望收益(均值)和风险(方差)帮助智能体在不同策略间权衡，可以适应不同类型智能体的风险偏好。以成本作为目标函数时，对于给定的混合联合策略ρ_{-i}，第i个智能体对ρ_{-i}的最佳响应：

$$\psi_i^*=\arg\min_{\psi_i\in\Psi_i}Var(\overline{obj}_i(\psi_i,\rho_{-i}))+\rho\cdot\overline{\text{mean}}_i(\psi_i,\rho_{-i}) \tag{11}$$

其中，ρ为超参数。仅采用期望模型可能导致较大的方差，因此通常使用均值-方差均衡来实现低延迟和低方差间的平衡。

(3)风险度量均衡将风险度量CVaR_α与决策理论结合来实现决策。在给定风险水平的置信度$\alpha\in(0,1]$时，以成本作为目标函数，混合联合策略为ρ_{-i}，则第i个智能体对ρ_{-i}的最佳响应：

$$\psi_i^*=\arg\min_{\psi_i\in\Psi_i}\text{CVaR}_\alpha(\overline{obj}_i(\psi_i,\rho_{-i})) \tag{12}$$

5 多智能体风险决策方法

多智能体风险决策方法是多智能体在现实世界中面临风险时进行风险决策的关键。本节基于多智能体风险决策理论综述当前研究中多智能体在面对环境风险、其他智能体风险以及智能体自身风险时的风险决策方法。基于最优控制理论的风险决策方法主要以控制理论为基础，重点关注系统的动态模型以及如何通过设计控制器使系统在风险条件下达到最优性能，通常以系统的动态模型为核心，通过对系统状态方程、输出方程以及各种约束条件的数学描述来构建控制策略。这类方法往往依赖于精确的系统模型，并重在求解满足一定最优性准则(如最小化成本函数)的控制律。而基于MARL的风险决策方法则大大降低了对于系统模型的要求，主要基于强化学习理论，关注智能体和环境交互时的互动关系，通过试错和经验积累来学习风险敏感策略，目的是通过智能体不断地探索和利用环境来最大化长期累积奖励。基于博弈理论的风险决策方法侧重于研究多个智能体之间的利益冲突和合作关系，主要通过分析智能体的策略集、收益函数等寻找智能体间具有耦合关系下的博弈均衡。

其中，基于最优控制与基于强化学习的风险决策方法具有天然的互补性，最优控制为多智能体系统提供了基本的理论框架和稳定性保证，而强化学习可以在此基础上通过不断学习来适应复杂多变的外部环境。基于最优控制与基于博弈均衡理论的风险决策方法可以协同用于决策，其中最优控制为博

弈均衡理论中的策略实施提供物理基础，而博弈均衡理论中的策略选择和均衡状态往往又会受到系统控制能力的约束，所以最优控制中的控制策略也需要考虑博弈参与者的利益和行为特点，以避免因不合理的控制策略导致博弈参与者采取不利于系统稳定的行为。基于强化学习与基于博弈均衡理论的风险决策方法可用于系统的联合优化，例如在一些复杂的系统中需要同时考虑智能体的学习能力（强化学习）和相互之间的利益博弈（博弈均衡理论），从而更有效地优化系统的整体性能。

5.1　基于最优控制理论的多智能体风险决策方法

基于最优控制理论的多智能体风险决策方法以寻求多智能体在面对环境、其他智能体风险时的稳定状态为决策目标，通过动力学建模多智能体系统来求解风险策略，主要应对环境风险、其他智能体风险影响下的控制策略生成问题。基于最优控制的方法具有很好的精确性和稳定性，能够提供精确的数学模型来描述系统的动态行为，从而实现精确控制，能有效地应对各种干扰和不确定性因素，保证系统的稳定性，适用于系统动态特性已知且可以用数学模型准确描述的情况。本节将基于最优控制理论的多智能体风险决策方法分为两类：(1)智能体数量有限时的动态规划方法；(2)智能体数量趋于无穷时的平均场方法，如表3所示。

表3　基于最优控制理论的多智能体风险决策方法

方法类型	作者	环境风险	智能体自身风险	其他智能体风险	风险度量	风险应对方案
动态规划	Hamadene等[90]	√		√	期望/方差	采用布朗运动表示风险，利用耦合HJB方程求解纳什均衡
	Chen等[91]	√		√	期望/方差	采用布朗运动表示风险，利用耦合HJB方程求解再保险策略
	Ghosh等[92]	√		√	期望/方差	采用维纳过程表示风险，利用多参数特征值方法分析平稳马尔科夫策略空间中的纳什均衡
	Biswas等[93]	√		√	期望/方差	采用维纳过程表示风险，利用HJI方程求解基于风险敏感成本准则的零和博弈
	Patil等[94]	√		√	期望/方差	采用维纳过程表示风险，通过建立路径积分框架对风险最小化的零和博弈进行数值求解
平均场[95-98]	Moon等[99]	√		√	期望/方差	采用高斯分布表示风险，通过求解局部风险敏感最优控制问题得到鲁棒分散控制器
	Liu等[100]	√		√	期望/方差	采用维纳过程表示风险，风险敏感性以指数积分形式建模
	Djehiche等[101]	√		√	期望/方差	采用布朗运动表示风险，将部分可观问题转化为完全可观问题求解
	Saldi等[102]	√		√	期望/方差	通过指数效用函数引入每个agent的风险敏感性，将部分观测随机控制问题转化为完全观测问题
	Moon等[103]	√		√	期望/方差	通过指数成本函数和最坏情况下风险中性成本函数结合来考虑风险
	Barreiro等[104]	√		√	期望/方差	通过布朗运动、跳跃过程和状态切换表示风险，构造并求解矩阵值线性二次平均场型博弈
	Ren等[105]	√		√	期望/方差	引入指数成本和共同的布朗运动，揭示平均场博弈下的风险敏感行为
	Huang等[80]	√		√	风险值	通过增加一个未知的扰动关注漂移不确定性带来的风险
	Tchuendom等[106]	√		√	期望/方差	采用布朗运动表示风险，利用随机极大值原理描述问题的最优解
	Tembine等[107]	√		√	期望/方差	采用布朗运动表示风险，利用随机极大值原理描述问题的最优解
	Moon等[108]	√		√	期望/方差	采用布朗运动表示风险，利用风险敏感极大值原理求解风险敏感平均场博弈

5.1.1　动态规划方法

动态规划是最优控制中设计多智能体系统控制器的通用方法。当系统面临噪声、干扰等风险因素时，通过调整控制输入改变系统状态以实现性能指标的最小化。性能指标通常是与系统状态量及控制输入相关的均值-方差函数，因此在最优控制中常采用期望/方差度量环境风险或其他智能体风险。HJB（Hamilton-Jacobi-Bellman）方程及HJI（Hamilton-Jacobi-Isaacs）方程是动态规划方法中的核心工具，两者共同构成了最优控制策略求解的重要数学框架。

(1)基于HJB方程的动态规划方法

HJB方程是动态规划方法在连续时间上的具体应用，是最优性原理及递归关系的数学表达。基于最优性原理，HJB方程通过值函数的偏导数来预测未来成本，因此通过求解HJB方程能够确定最优控

制策略。事实上，通过将智能体的策略视为控制变量就可以利用HJB方程来求解随机微分对策问题，其中微分对策将对策作为控制器，用最优控制的数学原理求解策略。基于耦合HJB方程我们可以得到主特征函数，也就是马尔科夫决策过程的最优值函数，从而找到马尔科夫决策空间的纳什均衡[90]。

动态规划过程往往涉及多个智能体之间的相互影响，有时需要建立耦合HJB方程组将多个智能体的最优化问题统一到同一系统中。在某些多维度、多变量的金融衍生品定价和风险管理问题中，通常采用HJB方程组同时考虑多个相关资产或变量的相互作用和影响[91]，即利用随机动态规划方法分析不确定性带来的风险决策问题。

(2)基于HJI方程的动态规划方法

微分对策理论将微分对策的解与HJB方程的解联系起来，通过极大极小原理描述智能体间的相互作用，从而得到HJI方程。HJI方程为值函数的扩展，它将最优性原理表示为每个时间点都要考虑其他智能体的反应，从而利用值函数分析智能体自身以及其他智能体的最优成本。

HJI方程与微分动态规划联系紧密，通过数值方法近似求解HJI方程以获得最优控制策略和系统性能指标。特别地，多智能体风险决策中动态规划的形式化描述的变化会引起最优控制问题的形式发生改变。因此，通过考虑风险得到类HJI方程的后向偏微分方程，以更好地描述最优控制问题[92]。正向和后向随机微分方程理论提供了适合分析这类问题的数学工具。将鞍点理解为HJI方程的极小极大选择器，通过分析鞍点的存在性揭示最优联合策略的存在性[93-94]。例如，针对零和随机微分博弈问题，Biswas等[93]研究受控扩散过程的风险敏感模型，建立了鞍点平衡点；Patil等[94]研究了连续时间风险最小化的双人零和情形的博弈问题，借助路径积分控制进行数值计算，给出基于Dirichlet边界条件的HJI方程的鞍点策略显式表达。

5.1.2 平均场方法

平均场方法主要用于智能体数量较大时的场景，用于近似求解大量智能体造成的高维优化问题。平均场博弈(Mean field game，MFG)将分散的博弈问题转化为集中的随机控制问题[95]，提供了近似纳什均衡，其中风险敏感MFG则用于求解风险敏感策略[96]。耦合HJB方程和FPK(Fokker-Planck-Kolmogorov)方程[97]用来描述平均场动力学系统，其中前者描述最优控制问题的动态规划形式，后者描述系统的概率特性。在HJB方程中，扰动以随机噪声的形式出现；FPK方程中扰动则通常以扩散项的形式出现。前者干扰智能体的控制策略选择，后者影响系统的状态分布，使得系统在决策过程中面临着随机性带来的风险。

MFG首先利用基于HJB方程的动态规划求解给定平均场分布下的每个智能体的最优策略，而后将最优策略代入FPK方程更新整个系统的状态分布。这一新的状态分布将影响下一步的HJB方程求解，持续这一迭代反馈过程直至收敛到纳什均衡。如图5所示，该平均场系统的交互过程将HJB方程分析过程同FPK方程分析过程联系起来，通过反映智能体之间在随机扰动风险下的相互影响和动态变化描述多智能体系统的决策行为。后向HJB方程用于确定智能体对平均场的最佳响应，前向CK(Chapman-Kolmogorov)方程计算智能体最佳响应下的平均场，基于这两个耦合方程产生平均场均衡。此外，还有一种基于动态规划方程与

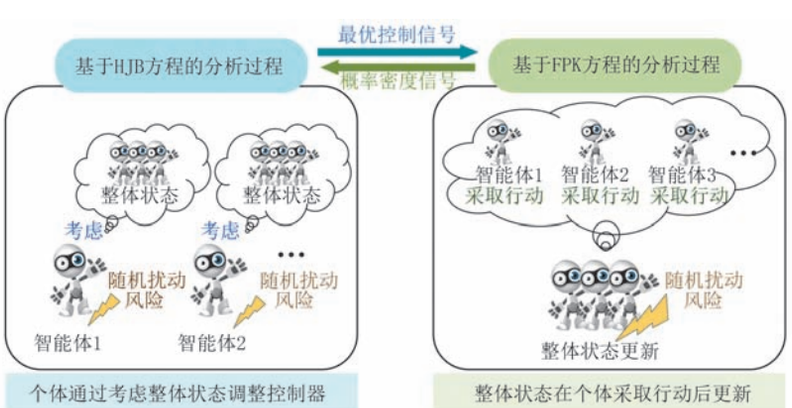

图5 平均场系统交互过程(在随机扰动下，通过HJB方程分析过程求解每个智能体最优策略，通过FPK方程更新整个系统状态分布)

Kolmogorov方程的耦合系统模型，通过综合风险度量对智能体的风险厌恶进行建模[98]。

(1)基于线性模型的平均场方法

平均场方法常用于分析多智能体线性二次博弈模型[99-100]。具体地，利用平均场风险敏感控制理论，Moon等[99]研究了离散时间下线性二次风险敏感平均场对策。而对于连续时间通过构建具有指数形式的成本函数考虑风险，Djehiche等[101]基于随机极大值原理，提供了部分可观的风险敏感平均场博弈的研究思路。Saldi等[102]则基于信念空间将部分可观随机控制问题转化为完全可观问题。在数学上信念是指博弈的不同时间步上所有智能体的状态和控制的概率分布。同样是在指数成本函数的基础上，Moon等[103]考虑了具有线性个体动态和智能体局部状态信息的连续时间风险敏感平均场博弈，通过求解不确定性最优控制问题，确定产生平均场均衡的控制策略。

平均场型LQG(Linear-quadratic-gaussian)博弈描述的是状态动力学或成本函数涉及平均场项的LQG博弈模型，通常采用随机最大原理或HJB方程、FPK方程与动态规划的结合来解决问题。然而，对于条件平均场型LQG博弈，则需要新的求解方法。Barreiro等[104]针对一类跳跃-扩散过程的条件平均场型LQG建立了风险中立均衡方法和鲁棒对抗平均场型鞍点。Ren等[105]研究了具有常见噪声的LQG风险敏感平均场对策，引入了指数成本和共同的布朗运动，目标是最小化反映其风险敏感性的指数成本函数，然后利用凸分析方法推导出当智能体数量趋于无穷时的最优策略。此外与风险敏感MFG密切相关的其他研究包括稳健MFG[80]以及量化MFG[106]，前者将模型模糊性纳入策略优化中，利用模糊度度量风险，而后者则建立了针对种群分布的特定分位数，利用期望度量风险。

(2)基于非线性模型的平均场方法

非线性控制中状态变量和控制变量之间存在非线性关系，这使得系统的描述和分析更加复杂和困难。涉及到随机性因素的非线性问题通常较为复杂，往往都需要引入随机极大值原理来求解[107-108]。随机极大值原理核心思想是在随机环境下寻找一个控制策略，使得随机过程的期望值最大化或最小化。具体地，Tembine等[107]设计了一个具有Lp范数结构的状态动力学的连续时间平均场博弈，利用随机极大值原理来描述最优解。Moon等[108]构造了固定概率测度下的最优控制问题，利用随机极大值原理求解风险敏感平均场博弈的近似纳什均衡。

当前研究中，基于最优控制原理的多智能体风险决策方法在实验设置上主要围绕构建多智能体系统的动态模型展开。在这些实验中，数值仿真占据主导地位。其中，风险往往以随机过程的形式呈现，被添加到智能体的状态方程或环境模型中，用来模拟现实中存在的各种不确定性因素。通过数值仿真实验，探索在理论上能够优化风险条件下智能体策略的方法，从而提升系统的稳定性和可靠性。

5.2　基于强化学习的多智能体风险决策方法

MARL通常以最大化奖励或收益来求解多智能体策略，作为一种端到端的智能决策方法已经在游戏领域展现出了巨大的潜力，具有操作简洁、便于求解的优点。本节我们综述以最大化奖励为风险决策目标的强化学习方法，如表4所示，其中，值分布MARL从风险表征出发捕捉多智能体决策时面临的风险；鲁棒MARL、安全MARL则关注决策中受各种风险影响时的策略鲁棒性和安全性；而贝叶斯MARL则基于先验推理来有效应对风险；平均场MARL则基于平均场理论应对大量智能体存在时的风险。基于MARL的风险决策方法具有很强的适应性和学习能力，能够在不断的交互过程中学习和适应复杂环境，适用于系统动态特性难以用精确数学模型描述或者环境动态变化频繁的情况。

5.2.1　分布表征

值分布强化学习是一种基于价值分布的强化学习方法，它通过概率分布来表征环境中可能存在的风险信息，包括环境风险、智能体自身风险和其他智能体风险，并利用分布的变化描述智能体在面对风险时的风险态度变化，如图6所示。风险敏感强化学习在传统强化学习中考虑了风险度量，来改善智能体决策时选择高风险动作时可能带来的恶劣影响，进而降低智能体自身风险。

值分布强化学习通过对随机回报的分布进行建模，充分挖掘分布中包含的附加信息，例如更丰富的环境特性，比非分布式强化学习算法能更加充分利用反映环境随机性的完整信息[109-110]。相关生物学上的研究也表明，人类大脑也采用了类似价值分布的机制进行学习[111]。PD-FAC算法[112]针对多机器人的可靠搜索问题，将多机器人系统的整体价值分布分解为单个机器人价值分布的线性组合，在两个典型的多机器人搜索仿真环境(即OFFICE和MUSEUM)上获得了最大的综合目标捕获概率。而后在自建的室内环境中进行了搜索运动目标的实

表4 基于强化学习的多智能体风险决策方法

方法类型	作者	环境风险	智能体自身风险	其他智能体风险	风险度量	风险应对方案
分布表征[109-111]	Sheng等[112]	√		√	期望/方差	值分布表示风险
	Qiu等[113]	√	√	√	风险值	值分布表示风险，并考虑风险值
	Sun等[114]	√		√	期望/方差	值分布表示风险，学习风险敏感策略[115-116]
	Shen等[25]	√		√	风险值	效用值分布表示风险，并考虑风险值
	Son等[9]	√		√	期望/方差	风险源分离，值分布表示风险
	Kumar等[117]	√	√		期望/方差	结合期望值理论和前景理论应对风险
	Ghafarian等[118]	√	√		期望/方差	采用期望表示机器学习方法误差带来的风险
	Fei等[119]	√	√		风险值	指数效用函数表示风险，通过参数调整风险偏好
鲁棒MARL	Lin等[8]	√	√		期望/方差	利用强化学习训练攻击者策略，然后使用有针对性的敌对示例迫使受害智能体行动
	He等[120] Han等[121]	√	√		期望/方差	将一组状态扰动对手引入到马尔科夫博弈中，并采用鲁棒多智能体Q算法求解鲁棒均衡
	Zhou等[122-123]	√	√		期望/方差	采用对抗状态下的动作损失函数表示风险
	Zhang等[124]	√	√		期望/方差	通过安全屏蔽来处理扰动或不确定状态输入
	Li等[125]	√		√	期望/方差	采用对抗学习应对对手策略扰动
	Li等[126]	√		√	期望/方差	利用智能体对其他智能体策略的后验信念促进合作，最大限度地减少对抗性动作的干扰
	Zhang等[127]	√			期望/方差	采用隐性参与者生成对抗策略
	Xue等[128]	√		√	期望/方差	采用消息重构的防御方法，在消息攻击下保持多智能体的协调
	Xiao等[129]	√	√		期望/方差	每个智能体使用局部批评家计算一个基线来减少其策略梯度估计的方差
	Alexander等[130]	√		√	期望/方差	通过对抗性正则化应对多种随机干扰的叠加
安全MARL[131-136]	Gu等[137]		√	√	期望/方差	最大化其奖励的同时满足自身安全约束并考虑其他机器人的安全约束以保证安全的团队行为
	Zheng等[138]			√	期望/方差	基于领导者-追随者双层优化模型提高智能体的奖励和安全性能
	Vinod等[139]		√		期望/方差	利用凸优化方式过滤不安全策略以确保系统安全
	Ding等[140]		√		期望/方差	通过对抗奖励函数和随机效用函数表示风险，采用上置信度强化学习求解拉格朗日优化问题
	Ingy等[141]		√		期望/方差	通过屏蔽来监控所有智能体的联合动作
	Zhang等[142]		√		期望/方差	采用基于控制屏障功能的安全屏蔽模块检查智能体的不安全操作
	Cai等[143]		√		期望/方差	采用分散的多个控制屏障函数屏蔽不安全动作
	Daniel等[144-145]		√		期望/方差	使用指定的屏蔽智能体纠正不安全动作
贝叶斯MARL[146]	Chalkiadakis等[147]	√			期望/方差	采用贝叶斯方法推理环境和策略的不确定性
	Foerster等[148]	√		√	期望/方差	通过近似贝叶斯更新来获得智能体对环境中其他智能体所采取策略的信念
	Wang等[149]	√		√	期望/方差	采用动态贝叶斯网络捕捉环境变量、状态、行动和奖励之间的不确定性
	Chen等[150]	√		√	期望/方差	采用可微有向无环图方法学习具有部分可观察性和各种风险场景下的上下文感知贝叶斯网络策略
	Du等[151]	√	√		期望/方差	利用高斯先验知识在估计值函数时整合了更多的先验知识，提高了学习效率
	Eriksson等[152]	√		√	期望/方差	采用多目标梯度法应对类型不确定性带来的风险
平均场MARL[153-155]	Ding等[156-157]		√		期望/方差	利用决斗网络降低值分数逼近过程的经验风险
	Wang等[158]	√			期望/方差	利用多分辨率观测信息和基于平均场的聚合全局信息训练
	Dey等[159]	√		√	期望/方差	采用分布式强化学习算法学习策略

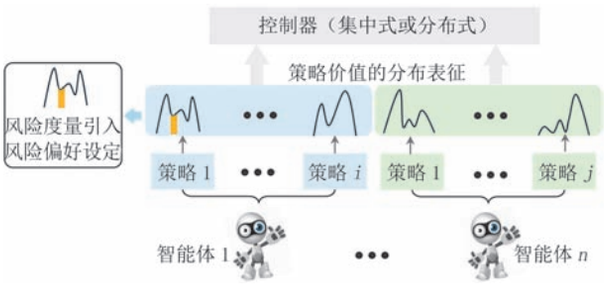

图6　策略价值的分布表示

验，采用了DM3008差速驱动机器人，使用单光束LiDAR进行地图构建，运动目标为C30差速驱动机器人，捕获概率在时间预算为7.5时达到100%。目前已有相关研究将值分布扩展到多智能体融合中，但研究工作较少，尚处于起步阶段。RMIX算法[113]在智能体动作值分布基础上引入条件风险值进行决策。DFAC算法[114]提出均值形状分解，在策略网络中采用隐式分位数网络对策略值进行分布化表示。RiskQ算法[25]则引入了风险敏感型个体-全局最大化原则，提出将分位数建模为每个智能体收益分布的效用，并通过加权来建模联合收益的分布表征。DRIMA算法[9]将风险来源分离为环境风险和合作风险，通过在策略探索阶段进行风险态度设置进行策略学习。值得注意的是，DRIMA算法需要通过手工设置智能体选择策略和应对环境风险的态度。RMIX、DFAC、RiskQ和DRIMA算法属于值分解MARL算法，采用集中训练分布执行的架构实现，均在基准环境星际争霸SMAC上进行实验测试，如图7所示。其中SMAC环境中每个行动单元由独立的智能体进行控制。这四种算法的参数配置均采用基准框架PYMARL的默认参数。性能方面，以困难设置下的地图3s_vs_5z为例，RiskQ算法报告了在探索（考虑其他智能体大量探索）、噪声（考虑其他智能体随机策略）和困境（考虑环境负奖励）等风险条件下的最佳性能。

在多智能体合作中，智能体只能根据自己的观测选择策略，因此其他的智能体和环境的影响都可纳入到环境风险中考虑。但除了环境风险之外，还有一点不容忽视，即是智能体选择的策略可能导致的风险，即涉及到风险敏感策略的概念。风险敏感策略是智能体在现实世界中应对风险所必须考虑的[115-116]。风险敏感强化学习使用风险度量优化策略，将具有较小预期总回报的风险降至最低，例如RMIX算法通过将条件风险值作为TD目标来优化策略。Kumar等[117]利用强化学习研究在具有共同

图7　星际争霸SMAC环境。SMAC是具有部分可观特性的合作控制环境，目标是利用我方AI控制的智能体对战内置AI智能体，以获取最大胜率。

风险的社会困境中人类的决策模式。Ghafarian等[118]在分类任务中也考虑了风险因素，提出使预期风险最小化的主动鲁棒学习方法。风险敏感值迭代和风险敏感Q学习方法[119]考虑使用指数效用函数选择策略，在面对不确定性时用于寻求风险和规避风险的策略探索方式。

5.2.2　鲁棒MARL

鲁棒MARL旨在学习一种考虑模型不确定性的鲁棒策略，以系统应对方法误差（自身风险）、环境（环境风险）、对手干扰时（其他智能体风险）的不确定性，已经在自动驾驶、工业机器人领域发挥着重要的作用。由于智能体在学习策略时，学习过程的每个要素（状态观测、动作、过渡动力学和奖励）都可能受到干扰（如传感器噪声、感知误差和对手攻击，下文称为过程干扰），所以鲁棒的策略就是要在面对这些干扰时依然能够保持稳定，而这种干扰实际影响了智能体选择策略后结果发生的概率。鲁棒MARL并没有显式的直接度量风险，通常采用优化误差的方式来训练风险敏感策略，因此在度量风险时本质上采用了期望/方差的形式。

智能体通过观测环境状态进行决策，但由于状态的不确定性导致信息并非完全准确。因此针对状态的不确定性，Lin等[8]研究了合作MARL中的状态扰动，提出了一种通过攻击智能体状态以降低团队奖励的方法，通过在SMAC环境中的2s3z地图上对单个智能体观察攻击将团队胜率从98.9%降至0。相反，He等[120]、Han等[121]考虑了智能体的状态可能被对手干扰的情况，引入了鲁棒均衡的概念，并提出学习一种对抗策略来干扰每个智能体的观测。他们基于多智能体粒子环境（MPE）测试状态不确定性下合作策略的鲁棒性，并采用了与MADDPG

算法相同的实验配置，通过设置不同类型的扰动形式模拟风险，用于在训练时提升智能体在最坏情况下的性能，实验结果表明通过考虑状态风险有效提高了训练策略对随机和对抗状态扰动下的鲁棒性。MPE环境如图8所示。Zhou等[122]提出通过最小化智能体在非摄动状态和摄动状态下行为之间的交叉熵损失来学习鲁棒策略，并在MAgent环境[123]中的战斗和追击场景下证明了所提出的动作损失函数能够提高训练模型的鲁棒性。Zhang等[124]考虑了状态不确定情况下的无人机安全控制问题，提出了一种具有安全防护能力的鲁棒多智能体近端策略优化方法来处理扰动或不确定状态输入。

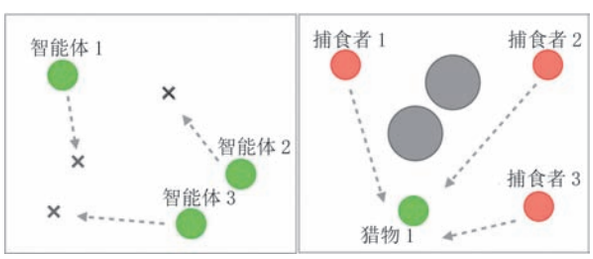

图8 多智能体粒子环境(MPE)(合作导航任务中智能体必须学会在避免碰撞的同时覆盖所有的地标(左图)。捕食者猎物任务中捕食者需要通过合作追击猎物以获取奖励(右图))

由于智能体间存在广泛的合作或对抗行为，使得智能体的动作不确定性一直是鲁棒MARL关心的焦点问题，特别是在多机器人控制中。M3DDPG算法[125]通过对智能体施加小扰动来训练连续动作空间中的协调策略稳定性，这种扰动形式被称为对抗性策略或非遗忘对手。M3DDPG算法同样在MPE环境下测试，实验配置与MADDPG算法相同，通过考虑对手智能体的对抗性策略，测试最坏情况下算法的收敛性，结果表明M3DDPG算法在不同难度的对抗性策略下都获得了更高的奖励。在此基础上，其他工作利用凸松弛等方法进一步增强M3DDPG算法能力，或通过假设其他智能体都是对抗性智能体来应对智能体动作的不确定性。

然而，这些方法并没有考虑多智能体决策目标和鲁棒性之间的权衡问题，为此，Li等[126]提出了一个贝叶斯对抗鲁棒框架，允许智能体基于它们对其他智能体类型的后验信念来学习策略，从而最大限度地减少对抗性动作带来的干扰。进一步，Li等将鲁棒MARL构建为一个推理问题，在所有威胁场景下通过评估策略来优化最坏情况下的策略鲁棒性。而后，Li等在矩阵博弈和SMAC环境下进行了大量实验，结果显示在所有环境中贝叶斯对抗鲁棒框架始终能够保持强大的性能，并在最坏情况下能够自适应地与队友结盟，显示出在非遗忘对手、随机盟友、攻击观察和攻击状态转移时的鲁棒性。

此外，实际应用中智能体很难获得完全准确的模型知识，即智能体的奖励函数和转移概率存在不确定性。Zhang等[127]使用鲁棒马尔科夫博弈来系统表征MARL中的模型不确定性，并将模型不确定性视为由隐式参与者做出的决策，进而引入自然智能体来建模不确定性，其中自然智能体是通过在每个状态下选择最坏情况的数据来对抗智能体，而后在MRE环境下进行了大量模拟，通过函数逼近和小批量更新来验证所提算法的性能，结果表明通过引入自然智能体的方法优于几种不考虑模型不确定性的基线MARL方法。

通信是多智能体间共享信息、实现高效合作的重要因素。然而在存在噪声或攻击的环境中通信往往最容易被干扰或操纵。因此，针对通信鲁棒性问题，一些研究通过考虑恶意攻击者的存在来扰乱正常消息或发送虚假消息，破坏多智能体系统的正常运行。这些方法中智能体通常会平等地接收所有消息来进行稳健决策，属于被动防御思想。而R-MACRL[128]采取了一种主动防御策略，即直接纠正受干扰信息，通过设置包含异常检测器和消息重构器两个组件的消息过滤器来防御消息攻击，并在网格世界和SMAC环境下进行了实验测试，验证了R-MACRL算法能够持续地恢复多智能体合作并提高算法在消息攻击下的鲁棒性。

除此之外，一些研究从方差缩减的角度出发提升策略鲁棒性，例如，ROLA算法[129]允许每个智能体学习局部评论家的单个动作值函数来减少策略梯度估计时的方差，通过在网格世界和MPE环境中的实验证明了ROLA算法相比于基线算法获得了更快的学习速度、更高的回报和更低的方差。实际上，外界干扰往往是多种随机因素叠加的结果，因此一些研究试图在多种干扰下提升策略鲁棒性。例如，Alexander等[130]提出了基于对抗正则化的鲁棒MARL框架ERNIE，并通过对抗性正则器实现了对观测、状态转移和恶意智能体的鲁棒性。

5.2.3 安全约束

与风险紧密相关的便是安全。在某些对安全要求严格的领域中，研究人员不仅要关注智能体长期回报最大化，还要关注如何避免损害智能体或危险事件的发生，例如在机器人操作、自动驾驶等应用中

对智能体自身风险的控制都提出了很高的要求[131-132]。安全RL指模型在学习或使用过程中，在满足一定安全约束的同时最大化长期回报的马尔科夫决策过程[133]。安全的策略应明确考虑训练期间对智能体动作探索的限制，以防止智能体处于危险状态，同时也要在优化目标中适当考虑环境风险和其他智能体风险，以提升策略学习的安全性能。

安全MARL通常采用三种范式实现，第一种是传统的无约束方法，通过对奖励的变换实现对智能体策略的安全控制；第二种利用约束准则，将最坏情况、风险敏感等安全因素考虑到优化过程中；第三种通过整合外部知识来影响智能体的探索过程。最早的安全MARL常用无约束方式，在多机器人协作导航和避障任务中应用[134-135]，其中安全往往采用碰撞惩罚或其他在奖励上的设计来实现[136]。与鲁棒MARL方法类似，安全MARL方法同样采用期望/方差的形式间接度量风险。

在优化方程中考虑安全因素是一种常用做法。在约束MARL中智能体在控制成本的条件下最大化其收益[23]。MACPO和MAPPO-L方法[137]针对合作中的多机器人安全控制问题，在策略优化过程中采用安全约束，通过在MuJoCop等仿真环境下与前沿算法相比，表明了其能够在提升收益的同时显著增强策略的安全性。MAPPO-L是在MAPPO算法考虑安全时的变体，其中MAPPO算法是PPO算法在多智能体中的扩展，基于集中式训练分布式执行架构实现，通过引入策略优化机制实现了更加高效的合作策略学习。然而，这些方法并没有考虑策略非平稳性的影响，缺少对其他智能体的有效推理和建模。为此，Zheng等[138]提出了一种基于斯塔伯格模型的安全MARL方法在联合策略学习过程中考虑对其他智能体进行建模。另一个研究方向是基于参数共享假设。Vinod等[139]针对多智能体路径规划问题开发了一个基于二次规划的安全滤波器，利用历史数据的同时将安全性作为规划中的硬约束，通过过滤不安全策略来确保系统安全，实验中使用了6架Crazyflie四旋翼无人机进行测试，与传统的模型预测控制方法相比，融合安全滤波器的方法安全性更好且运动规划成功率更高。Ding等[140]提出了一种零和博弈框架下的安全MARL方法，利用拉格朗日优化来维持系统安全。

通过影响智能体的探索过程学习安全策略的方法最为直接，例如屏蔽不安全动作，但这类方法往往需要预训练或先验知识来为过滤动作创建屏蔽。模型预测屏蔽算法的基本思想是使用备份控制器来检查策略是否安全。Ingy等[141]将单智能体RL中的屏蔽技术扩展到MARL中，开发了集中屏蔽和因子屏蔽，保证了学习过程中策略表示的安全。然而，这种方法需要结合外部信息进行修正，在实际操作中不太友好。因此，Zhang等[142]提出了一种采用控制屏障函数的MARL安全验证技术，其中安全屏蔽模块具有安全检查能力，可以防止智能体采取不安全动作。针对屏蔽方法对观测信息具有较强假设的限制，Cai等[143]提出将MARL与基于可用局部信息的分散控制屏障函数相结合，建立了具有分散屏障函数的安全MARL框架。Daniel等[144-145]考虑无通信及部分可观环境下使用屏蔽算法，使用指定的屏蔽智能体来管理和纠正系统内其他智能体可能导致的不安全动作。上述屏蔽算法主要采用网格世界进行验证，均能有效实现零碰撞，其中因子屏蔽在处理大量智能体时更为有效，而去中心化屏蔽尽管在性能和计算复杂性上有所折中，但提供了在通信受限和部分可观察环境中的有效解决方案。采用屏障函数的方法在MPE环境中进行了验证，仿真结果表明在避碰率和平均回报上均优于基线算法。

5.2.4　贝叶斯MARL

贝叶斯方法基于当前观测和先验知识求解后验概率分布，能够对不确定性带来的风险进行解释[146]。贝叶斯MARL利用贝叶斯推理将风险纳入到策略学习过程中，假设智能体可以在概率分布中表达先验信息并使用贝叶斯推理来合并新信息。

Chalkiadakis等[147]提出了一个解决MARL中最优探索的贝叶斯模型，该模型允许智能体推理环境风险，通过在马尔科夫链上的实验结果也表明与试图诱导算法收敛到最优平衡的启发式探索相比，采用贝叶斯方法来建模MARL可以提高多智能体在风险决策中的性能。Foerster等[148]提出了一种新的多智能体策略学习方法-贝叶斯动作解码器，它通过近似贝叶斯更新来获得智能体对环境中其他智能体所采取策略的信念。在双人博弈的卡牌游戏Hanabi环境中，贝叶斯动作解码器在近60%的游戏中达到了完美得分，并且在跟踪推断信息方面表现出远超人类的能力。Wang等[149]针对离线学习中的信用分配问题，提出了一种将学习过程建模为动态贝叶斯网络来捕捉强化学习各要素间关系的MACCA算法。该算法通过分析每个智能体奖励的因果关系来衡量它们的贡献，从而确保生成准确和可解释的信用分配策略，并与QMIX-CQL等基准

算法在SMAC环境上的对比结果展现出了更好的鲁棒性。Chen等[150]将现有的MARL算法与可微有向无环图相结合，引入了一个贝叶斯网络来建立智能体在策略选择时的相关性，从而学习在具有部分可观和各种风险场景中的应对策略。Du等[151]采用贝叶斯线性回归来估计关于值函数参数的后验分布，而不是通过点估计来近似Q值的期望值，从而得到具有表示能力更强的值函数。Eriksson等[152]关注与类型不确定性相关的风险，研究了由于其他智能体策略不确定性导致的风险敏感性，提出了基于风险中性随机博弈的MARL算法的风险敏感版本，如迭代最佳响应、虚拟博弈方法，并在随机玩具问题中进行了验证，表现出了更好的风险应对能力。

5.2.5 平均场MARL

平均场理论不仅在最优控制中备受关注，在MARL中也得到了广泛应用。随着智能体数量的增加，状态空间、动作空间的维度指数级增长，使得学习过程变得难以处理，而平均场MARL就是在MARL中应用平均场理论求解多智能体策略的一种方法[153-154]。Yang等[155]第一个提出了平均场MARL方法，利用平均场近似将复杂的多智能体系统简化为单个智能体与其邻近智能体平均效应之间的交互，从而通过智能体与群体动力学之间的相互关系来学习最优策略。

当前大量的文献侧重于研究利用平均场理论建模多智能体系统，而旨在学习风险敏感策略的平均场MARL研究相对较少。Ding等[156]结合了平均场理论和价值分解来处理多智能体决策模型的可扩展性问题，提出了一种多智能体决斗Q学习算法，利用决斗网络架构降低值分数逼近过程的经验风险，在MAgent环境以及更复杂的Google Research Football环境[157]中均表现出色，具有很好的泛化能力。Wang等[158]针对无人机编队在风险环境中的路径规划问题，提出了用于避碰的3M-RL算法，通过结合更详细的局部信息等多分辨率观测信息和基于平均场的聚合全局信息来训练，解决了MARL中的维度灾难问题，通过在具有二维和三维连续空间的仿真环境下测试了3M-RL，仿真结果表明3M-RL产生了良好的路由策略并有效提升了无人机在具有障碍物等环境风险下的适应能力。为了在风险环境下学习针对大规模多智能体系统的分布式策略，Dey等[159]开发了一种非平衡的平均场博弈理论并设计了自适应概率密度函数分解算法和分布式强化学习算法，具体而言，通过归纳式的概率密度函数参数估计方法将密度约束分解为多个不平衡的范数分布，采用约束K均值聚类算法将多智能体分解成多组以获得期望的最终概率密度函数约束，而后利用多因子临界质量算法求多组耦合HJB和FPK方程的解。

5.3 基于博弈均衡理论的多智能体风险决策方法

多智能体风险决策方法中最为广泛应用的当属基于博弈均衡理论的风险决策方法，该类方法在进行风险决策时重点考虑其他智能体影响下的多个智能体间的博弈关系。这类方法通常在多智能体系统的均衡策略指导下进行风险决策，从而得到了各种不同类型博弈模型下的风险决策方法。基于博弈均衡理论的多智能体风险决策方法专门用于研究多个参与者之间的利益冲突和合作关系，用于平衡各方利益，在决策层为参与者提供了一种风险决策分析的工具，适用于存在多个智能体且智能体之间存在利益冲突或合作需求的场景。本节中我们主要介绍了合作博弈、非合作博弈、演化博弈和马尔科夫博弈等四种常见博弈中涉及的风险决策方法，如表5所示，其中：合作博弈和非合作博弈通常在理性人假设下进行建模求解；演化博弈放松了对于理性人假设的要求，只需要假设参与人有限理性即可；而马尔科夫博弈则是一种最为常见的随机博弈模型，同时包括了静态博弈和动态博弈。

5.3.1 合作博弈

合作博弈中多个智能体以合作或协商的形式寻求能使集体利益最大化的联合策略。合作博弈中面临的风险决策问题常见于无人驾驶场景。在该场景中，风险条件下驾驶员和车辆间的控制权切换、车辆和车辆间的规划决策等均可以被建模为基于合作博弈的多智能体风险决策问题。

合作博弈中更多关注合作主体的选择和风险度量的考虑。在合作主体的选择中，经典方法是将人和机器视作具有一定差异的主体进而考虑其他合作主体的风险。例如在车辆轨迹规划研究中将存在耦合关系的路径和速度作为两个独立的合作主体[160]、在车辆变道决策问题中将多个车辆视为合作主体[81]。在合作博弈中引入风险度量能将智能体的风险偏好体现在策略选择过程中。例如在无人驾驶中由于视野角度的限制，使得无人车辆的每次决策均面临着策略选择的风险，Zheng等[161]建立了车辆碰撞模型以提高对环境风险表达的准确性，Fu等[81]则提出静电场理论的风险感知方法来量化风险，通过在开源交通仿真环境SUMO平台下设计了四种阻

表5　基于博弈均衡理论的多智能体风险决策方法

方法类型	作者	环境风险	智能体自身风险	其他智能体风险	风险度量	风险应对方案
合作博弈[160]	Fu等[81]		√	√	风险值	采用静电场理论的风险感知方法来量化风险，用于确定智能体运动状态的不确定性
	Zheng等[161]		√	√	风险值	进行驾驶风险评估促进无人车决策
	Vimala等[162]	√		√	期望/方差	采用极大极小原理定义特征函数，并使用均值-方差奖励函数的随机博弈模型
	Shi等[163]	√			期望/方差	利用随机对偶方法，求解不确定性条件下联盟内农户的公平成本分配策略
非合作博弈[164-166]	Yang等[167] Cui等[168] Fan等[169]	√		√	期望/方差	采用基于贝叶斯网络的因果推理分析风险
	Tatarenko等[170]	√		√	期望/方差	将风险建模为损失函数，利用凸优化理论分析
	Shuvasree等[171]	√		√	风险值	采用二阶模糊集表示不确定性带来的风险
	Dong等[172]	√		√	风险值	采用二阶区间值直觉模糊集表示不确定性
	Li等[173]	√		√	期望/方差	对风险策略进行约束
	Eisenstadt等[174]	√		√	期望/方差	采用向量收益函数表示风险条件下的成本
	Jakobik等[175]	√		√	期望/方差	在防御策略的上下文中自动选择一组可能的风险应对策略
	Tan等[26]	√		√	风险值	利用条件风险值建立风险评估程序应对结果不确定性
演化博弈	An等[176]	√		√	期望/方差	通过演化博弈求解鲁棒策略均衡应对风险
	Li等[177]	√		√	期望/方差	通过代理理论构建风险条件下的演化博弈模型
马尔科夫博弈	Huang等[178]	√		√	风险值	利用极大极小风险测度分析风险敏感策略
	Mathieu等[179]	√		√	风险值	将策略回报的风险敏感度量作为学习目标
	Wang等[180]	√		√	风险值	利用在线凸博弈最小化成本函数的CVaR值
	Ghaemi等[181]	√		√	期望/方差	采用累积前景理论考虑风险，学习具有主观性的风险敏感策略
	Ma等[182]	√		√	期望/方差	定义两组不同的随机变量表示风险，利用极大极小算法求解风险条件下的最优值函数

碍无人车通过的场景对该方法进行了测试，结果表明基于静电场理论的风险感知模型在平均速度和通过时间方面都优于当前的基线模型，并且该方法可以根据道路上不同无人车的不同驾驶风格调整策略，从而在满足安全约束的同时最大限度地提高系统的整体效率。

此外，合作博弈中风险决策问题也常见于银行、保险等领域，例如投资组合、最优风险分担等。早期文献选择预期折现收益来衡量智能体在合作博弈中的收益，但只考虑期望的策略将导致智能体风险中立，很难适应环境风险的变化。因此Vimala等[162]提出具有均值-方差收益函数的合作博弈模型，并结合极大极小方法求解风险下的合作策略。此外，还可以利用随机对偶方法同线性规划相结合来求解风险条件下的合作博弈问题[163]。

5.3.2　非合作博弈

非合作博弈常用来建模多个智能体为了不同利益而产生的合作或竞争行为。基于非合作博弈的风险决策问题更多关注均衡点的求解。面对不确定性带来的决策风险，通过最大化期望效用来寻找纳什均衡最为直接，但采用单一的期望效用最大化导向可能会造成决策失误。因此，许多研究人员在决策时考虑随机变量的影响，并通过调整智能体的风险偏好来适应非合作博弈中的环境风险或其他智能体风险[164-165]，例如采用风险规避均衡[166]、均值-方差均衡、$CVaR_\alpha$均衡等，然后再利用线性规划等数值方法近似求解均衡。

非合作博弈通过结合贝叶斯理论同样能够很好地应对结果不确定性[167-168]。这类方法首先采用贝叶斯网络评估博弈中的关键因素并将风险量化为概率分布，其次将贝叶斯网络的输出作为博弈模型的输入计算均衡策略，已经在解决海盗博弈问题[169]等实际应用中取得了良好成效。此外，基于凸优化理论分析非合作博弈在理论研究中也备受关注。该类方法首先将每个智能体的损失函数建模为凸函数，然后利用凸优化理论对风险规避的凸博弈模型进行

分析和求解，通常在严格凸的势函数条件下证明策略学习算法能够收敛到纳什均衡[170]。

非合作博弈中最经典的一种博弈形式当属矩阵博弈。矩阵博弈为分析多个智能体之间的合作与冲突关系提供了一个强大的数学工具。但作为一种简化模型往往回避了现实中的信息不完全情况。因此，不完全信息可能会导致智能体对策略收益的估计错误。因此研究者们围绕模糊环境下的矩阵博弈问题展开研究。为了应对结果不确定性，模糊矩阵博弈中往往会引入多种模糊集的拓展，例如一型模糊集和二阶模糊集等通过隶属度实现对智能体对环境不确定性的应对。Shuvasree等[171]为了解决工厂预算分配问题，基于二阶模糊集构建问题求解框架，提出了基于Hausdorff度量的矩阵博弈求解方法。为了给出隶属度的近似范围，Dong等[172]提出了一种具有更强的不确定性表示能力的模糊方法，并在新能源汽车产业发展方面已被成功验证。利用模糊矩阵博弈解决不确定性环境下的风险决策可以应用到许多实际问题中，如反恐问题。

斯塔伯格安全博弈也是一种非合作博弈，通过分析攻击者可能的攻击行为及防御者潜在的防御行为来预测攻击者策略。斯塔伯格安全博弈最大的难点在于如何建立攻击者的收益函数。攻击者收益函数与攻防双方都紧密相关。例如，Li等[173]通过最大化防御者的期望平均状态估计误差的协方差来定义攻击者的收益函数，而Eisenstadt等[174]提出向量收益函数并同时考虑了攻击者同防御者的行动成本。大多数情况下斯塔伯格安全博弈假设攻击者是完全理性的，然而这种假设并不总是成立，在对手策略未知时需要定义新的收益函数来应对。因此Jakobik等[175]基于观测结果逆向建立攻击者的收益函数，从而预测攻击者的策略。此外，量子响应均衡模型在对手策略未知时也具有很好的效果，该类模型假设对手基于某种概率，采取可能不是最大化其效用的策略。应用方面，Tan等[26]利用斯塔伯格博弈对能源运营商及能源消耗者进行建模，并利用条件风险值建立风险评估程序来应对结果不确定性，进而实现基于斯塔伯格博弈的风险策略求解。

5.3.3 演化博弈

上述考虑风险的博弈方法均围绕纳什均衡展开，但纳什均衡属于静态均衡，难以体现出开放环境下博弈过程面临各种风险时的均衡变化。而演化博弈中智能体的策略可以通过自然选择过程进行学习、进化等。因此适应能力强的策略，能够很好地抵御外界环境的干扰。实际上，演化博弈从系统、宏观的角度描述了群体博弈的策略演化过程，涵盖了智能体在不同时刻面对不同风险时的微观风险决策问题。特别的，演化博弈在经济学领域的研究更多关注对结果不确定性的处理和应对，已成为演化经济学的一个主要分析手段。

在金融领域中，由于财务风险，供应链成员之间必须不断调整风险应对策略，因此参与者（供应商和制造商）对财务风险的感知是一个进化的学习过程。例如，An等[176]利用演化博弈模型对金融监管与金融创新的动态博弈过程进行建模，通过求解稳定的策略均衡实现金融机构与监管机构之间的协调工作。在企业投融资过程中，股东的博弈策略是为了在控制债务危机前提下获取投资利润，而管理者的博弈策略主要是基于对更高利益的期望，因此两者间往往涉及投融资方面的风险决策问题。为此，Li等[177]通过代理理论构建风险条件下的投融资行为演化博弈模型分析企业投融资过程中涉及的利益冲突，为股东与管理者提供了合理的风险投融资策略建议。

5.3.4 马尔科夫博弈

在多智能体马尔科夫决策过程中，如果智能体之间是拥有共同利益和目标的合作关系，那么通常可以采用MARL方法来学习风险条件下智能体之间的联合策略（见5.2节）；如果智能体之间拥有各自的利益和目标，即存在博弈关系，那么可以转化成马尔科夫博弈问题，进而从博弈的角度求解风险条件下的近似均衡策略。

Huang等[178]提出了一个考虑风险意识的非合作马尔科夫模型，其中智能体具有时间一致的风险偏好，该模型描述了环境的随机状态转换和合作者的随机混合策略带来的时间一致的动态风险，进而提出了“风险意识”马尔科夫完美均衡的概念，并设计了一种类似Q学习的算法用于计算该完美均衡。Godbout等[179]针对强化学习中的二人零和博弈问题，证明了博弈方越接近均衡点，学习策略就越接近CVaR最优策略，并提供了一个基于梯度的斯塔伯格博弈来求解。Wang等[180]针对具有风险厌恶智能体的在线随机博弈中成本CVaR值难以计算的问题，提出了一种依赖于CVaR梯度的在线风险规避学习算法，目标是风险最小化时学习最优策略。Ghaemi等[181]在网络聚合的马尔科夫博弈中考虑了累积前景理论，提出了一种分布式算法求解风险敏感策略，并在一定假设下证明了该算法收敛于一个

风险敏感的马尔科夫完美纳什均衡。Ma等[182]提出一种入侵响应决策方法，以马尔科夫博弈框架构建了入侵响应决策模型，并利用极小极大算法求解特定状态下的最优值函数。

基于博弈论的多智能体风险决策方法中合作博弈、非合作博弈和演化博弈大多面向具体应用，通过针对这些特定应用场景下的实验，探索多智能体在复杂风险环境下的决策行为，以此求解风险条件下的博弈均衡策略。而马尔科夫博弈通常以数值仿真实验为主，通过在模型中考虑风险因素测试风险敏感策略的鲁棒性。三种多智能体风险决策方法的实验验证细节如表6所示。

表6　多智能体风险决策方法实验验证环境

实验方法	类型	环境风险	智能体自身风险	其他智能体风险	风险设置	文献
数值仿真	零和/非零和	√		√	布朗运动/维纳过程	文献[90-94]
	平均场	√			布朗运动/维纳过程/高斯分布/指数成本	文献[80,99-108]
	马尔科夫博弈	√		√	扰动	文献[178-184]
游戏仿真	星际争霸SMAC	√	√	√	探索/困境/噪声	文献[8-9,25,113-114,126,128,149]
	粒子环境MPE	√	√	√	过程干扰	文献[120-121,125,129,142]
	Magent环境	√	√	√	过程干扰	文献[122,156]
	网格世界	√	√	√	过程干扰	文献[128-129,141-145]
	机器人仿真	√	√	√	策略约束	文献[112,137,158]
	SUMO交通仿真	√	√	√	交通障碍	文献[81,161]
场景测试	无人机验证	√	√	√	障碍设置	文献[139,158]
	无人车验证	√	√	√	障碍设置	文献[81,160-161]
	投资金融场景	√		√	对抗性策略	文献[176]

6　应用领域

随着人工智能技术的发展，多智能体风险决策方法在各种现实任务中都取得了超越人类水平的风险应对能力，并已经在金融、控制等领域中产生了重要影响。本节以多智能体风险决策方法在现实世界中的应用为侧重点，重点讨论人机协作、自动驾驶、交通控制和智能电网四类场景下不同风险决策方法的风险应对思路，如表7所示。在不同的应用中，基于控制理论的多智能体风险决策方法主要服务于智能体的底层控制，通过设计控制器的方式在控制层降低智能体自身决策的风险，从而追求个体的策略安全以及整体的状态稳定。基于强化学习的风险决策方法则更多关注环境动态变化时多智能体如何学习适应性更强的合作策略，从学习和优化的角度得到收益最大化的全局最优策略，重点应对环境风险和其他智能体的风险。基于博弈均衡理论的风险决策方法从智能体与智能体、智能体与环境间的交互关系出发研究不同参与者间的博弈关系对风险决策的影响，从而获得能有效应对环境风险和其他智能体的风险的均衡策略。因此，这三种理论的风险决策方法使得多智能体系统能够在充满风险的复杂环境中做出更加有效的决策。

6.1　人机协作

多智能体风险决策问题在多智能体、多人交互以及人机协作场景中广泛存在，并随着人机混合智能的发展愈发变得重要。人机协同决策通过整合人和机器的能力实现更强的环境适应能力。在人机协作场景下，风险决策不仅要考虑机器控制策略的安全性、人机协作策略学习的风险敏感性，还要重点关注风险环境下主导决策权的归属问题，即人机协作时的利益冲突博弈。

人机协作中基于控制理论的风险决策方法主要用于调节人与机器之间的交互流程，通过建立反馈控制系统监测人的操作和机器的状态。当检测到人的操作可能导致风险（如人过于靠近机器的危险工作区域）时，系统立刻采取如降低机器运动速度或暂停操作等风险应对策略，从而使协作任务能够安全完成。基于强化学习理论的风险决策方法用于学习智能体在人协作中采取最优风险敏感策略，例如Liu等[185]提出一种强化学习方法用于实现工业机器人在人机协作中的实时无碰撞运动规划，通过结合外部奖励和内在奖励使机器人通过自我探索学习学

表7 多智能体风险决策应用研究

应用场景	作者	风险决策方法	风险应对方案
人机协作	Liu等[185]	基于强化学习的方法	结合外部奖励和内在奖励学习学习无碰撞的运动规划决策
	Patel等[186]	基于博弈均衡的方法	切换控制权，算法处理较高置信度输出而人类则检查较低置信度输出
自动驾驶[187-189]	Nguyen等[190]	基于最优控制的方法	通过安全级别评估环境风险，利用极大值原理生成避碰路径
	Zhou等[191]	基于强化学习的方法	设计多目标奖励函数综合变道等环境风险
	Li等[192]	基于博弈均衡的方法	评估其他智能体的风险，并基于评估结果学习具有风格偏好的避碰策略
交通控制	Wu等[193]	基于强化学习的方法	结合MADDPG和长短期记忆网络应对环境风险
	Yekkehkhany等[194]	基于博弈均衡的方法	考虑路径延迟的不确定性，基于风险规避均衡学习风险最小化策略
	Fu等[81]	基于博弈均衡的方法	通过感知风险场理论来量化风险，基于联盟博弈模型应对变道冲突
智能电网	Xu等[196]	基于最优控制的方法	使用分布式架构应对通信负荷，利用集成共识算法和最优控制算法学习鲁棒策略
	Zhang等[197]	基于强化学习的方法	SMAS-PL算法利用信息约束学习风险可控策略
	Shi等[198]	基于强化学习的方法	利用电压主动控制方法保持电压在安全范围内
	Ni等[199]	基于博弈均衡和强化学习	采用多阶段动态博弈确定目标的最佳攻击序列

习动态地避开人类操作者手臂的策略，实现安全交互。在人机协作控制权的转移中，MARL方法可以用于学习最优策略，而基于博弈的风险决策方法则通过设计好的激励机制促使人机之间形成有效的合作关系。例如在低风险环境下机器通常主导决策，而在高风险环境下由于机器难以从不确定数据中提取可靠知识，人类则主导决策权。Patel等[186]的研究结合放射科医生和群体AI的优势，对具有较高置信度的输出由算法处理，而人类则主要检查具有较低置信度的输出，从而实现更优的组合决策。

6.2 自动驾驶

对于自动驾驶而言，行为决策层面临着多变道路、随机路人、其他车辆介入等高度不确定性的动态场景，这要求车辆必须能够在各种条件下迅速作出正确且安全的选择。因此，通过风险决策方法发现潜在危险并采取预防措施在自动驾驶中具有重要作用，特别是在车辆安全控制、轨迹规划及预测以及多车协同等方面已成为当前研究的热点[187-189]。

自动驾驶通过设计合适的控制策略就能够有效降低智能体自身策略可能带来的风险。在多车协同场景中，利用基于最优控制的风险决策方法学习多个无人车的驾驶策略以提升无人车的驾驶安全。例如文献[190]通过设计安全级别评估环境风险，并利用基于Pontryagin极大值原理的轨迹优化理论生成避免无人车碰撞的安全路径。然而在模型难以获取的场景下，运用基于强化学习的风险敏感策略学习方法可以更好地让无人车在与环境的交互中学习并优化驾驶策略，在适应变化路况等新环境时具有显著优势。Zhou等[191]将无人车的变道决策建模为MARL问题，提出了一种具有局部奖励设计和参数共享方案的多智能体优势行为者评价方法，通过综合燃油效率、驾驶舒适性和自动驾驶安全性的多目标奖励函数为每辆无人车制定安全的变道策略。考虑到与其他车辆互动时的风险，则需要研究如何协调彼此的动作，这就需要用到基于博弈均衡的风险决策方法，比如通过建立模型来预测其他驾驶员可能采取的行动，并据此规划最优路径。例如文献[192]提出了一种基于风险评估的自动驾驶多场景避碰决策算法，该算法利用基于概率模型的条件随机场态势评估模块评估无人车周围交通参与者的风险水平，并基于态势评估模块所评估的风险，提出具有攻击性或保守性的驾驶风格偏好的避碰策略，以满足不同驾驶员或乘客的需求。

6.3 交通控制

随着自动驾驶、打车业务的快速增长，地面交通车辆增多，交通拥堵及安全事故频发，这使得交通控制中的风险决策问题显得尤为重要。在交通控制领域中多智能体风险决策方法常被用于优化交通信号和车辆路径，通过算法优化交通信号灯的配时和交通监控的覆盖范围减少交通拥堵，从而提高城市交通的整体效率。

交通信号控制是控制理论的典型应用，例如通过监测交通流量等参数采用合适的控制算法调整交通信号灯的周期和相位，以优化交通流，减少交通拥堵和交通事故的风险。或者利用最优控制理论建立控制模型，根据不同区域的交通需求控制交通流量的分配，引导车辆合理地分布在道路网络中，避免局部拥堵导致的交通瘫痪风险。基于强化学习的风险决策方法适用于自适应交通控制，能够根据实时的交通数据学习最优风险敏感交通信号控制策略，特

别是在具有高度不确定性的交通状况下。Wu等[193]将MADDPG与长短期记忆网络相结合用于多路口交通灯控制，以解决由部分可观察和道路动态性状态引起的环境风险。博弈理论常用于研究不同路口或区域之间的交通资源竞争和合作关系，通过博弈分析找到一种协调机制优化区域交通以降低交通拥堵和事故风险。例如Yekkehkhany等[194]基于风险规避均衡建立随机拥堵博弈，设计了考虑交通网络中不同路径随机延迟的智能导航系统，其中考虑路径延迟的不确定性，并选择风险最小化的路径。此外，Fu等[81]提出一种基于联盟博弈的混合交通环境下协同决策模型，整合感知风险场理论来量化风险，使得获得的策略能够应用于识别多辆车辆变道引起的冲突，从而保证交通流安全。

6.4　智能电网

在电力行业中，随着新能源接入比例不断增加，传统电网正逐渐向智能化方向转变。在这个过程中，智能电网重点关注分布式发电资源的有效配置方法，特别是针对可能出现的各种突发情况，要事先制定有效的风险应对策略。

基于最优控制理论的多智能体风险决策方法在智能电网中用于电力系统的稳定控制，或实现对分布式能源资源的有效整合，以提高电网的可靠性和稳定性。Xu等[195]提出一种集成共识算法和最优控制算法的方法，只需要相邻单元间交换信息，使分布式局部控制器之间能够分担计算和通信负担，目标函数设计为同时考虑发电供应商和负荷用户的整体社会福利最优，仿真结果表明该方法对通信故障等风险具有很强的鲁棒性。随着可再生能源的集成，导致电网的不确定性和波动性不断加剧，这给电网的电压控制、安全运行带来了严峻挑战。Zhang等[196]针对配电系统中网络化微电网的最优电源管理提出了一种有监督的多智能体安全策略学习方法SMAS-PL，其中每个智能体从相邻智能体接收拉格朗日乘子变量，然后在训练中使用信息约束来保证最优控制策略函数输出风险可控的策略。Shi等[197]则提出了一种带安全层的MARL电压主动控制方法，通过集中式数据驱动的安全层将不安全动作映射到安全动作，从而保持电网母线电压在安全范围内，并通过策略网络中的动作校正惩罚损失和子网络实现高效的分散部署。与此同时，考虑到电网的安全管理，Ni等[198]基于博弈理论，针对智能电网的攻击防御问题提出了一种基于强化学习的多阶段动态博弈方案。通过确定特定目标的最佳攻击序列，例如传输线中断或发电损失，对不同的防御策略进行评估。

7　挑战与未来研究展望

尽管在多智能体风险决策的研究中产生了许多新理论和新方法，在解决不同领域内的多智能体风险决策问题上已经取得了显著的成绩，但仍面临许多开放性的问题需要进一步的探索和解决。未来可以从以下五个方面对当前多智能体风险决策研究中面临的挑战进行拓展和深化。

7.1　理论：多人风险决策理论研究

目前多智能体风险决策方法通常以控制、博弈和学习等基于优化理论的思路进行研究，目的是获得风险条件下的最优解或次优解。但在具有大量未知因素及不确定性干扰的场景下，多智能体如何制定最优合作策略往往面临着巨大挑战。这种挑战更多的是由于目前缺乏多智能体风险决策理论的支撑。人类在合作决策时面临的场景更加复杂，但并不需要大量训练就能通过自身风险偏好和与合作者的博弈关系制定策略。因此，通过研究人类在合作时的风险决策理论可能对多智能体风险决策方法的研究具有一定参考。在未来的研究中，可以借鉴群体风险决策中对决策成员风险态度、心理行为等个体水平的研究[199]，结合行为科学、心理学开展人群调查、实验室模拟等手段进一步探索多人风险决策的理论基础，从而支撑多智能体风险决策方法的研究。

7.2　技术：复杂博弈中的风险决策方法研究

基于博弈均衡理论的风险决策方法致力于在风险条件下寻找均衡策略，但这些理论和方法尚未充分探讨复杂博弈环境下的风险决策问题，例如在均衡不存在或难以求解、其他参与人不采用均衡策略、存在欺骗等情况下如何进行风险决策依然是一个悬而未决的问题。目前已有研究采用对手建模[200]、循环推理[201]等结合机器学习、认知科学的手段求解对抗条件下的策略，其中风险被视为在效用函数中考虑的一个关键因素，这对于在复杂的合作博弈条件下进行风险决策提供了方法参考。在未来的研究中，可深化对多智能体互动和策略优化的研究，发展新的算法和模型以处理更复杂博弈场景下的风险决策，并通过研究多智能体风险决策理论和方法促进人们对复杂博弈过程的理解，这将为解决复杂博弈下的风险决策提供新的视角和工具。

7.3 基准:基准环境和数据集构建

目前,在多智能体风险决策研究中普遍缺乏通用的测试基准和数据,例如在投融资、供应链、公共管理等领域中都直接建模实际问题进行风险决策研究,而在多智能体博弈、强化学习中使用的测试环境和数据集也千差万别,这就导致多智能风险决策方法很难在一个基准上进行对比,从而难以实现不同方法在性能上的提升和完善。如何构建基准环境和数据集对于体系化研究多智能体风险决策理论和方法具有重要意义。本文提供了一种思路,例如可以按风险来源分类,构建针对自身风险、环境风险和其他参与者风险的基准环境或数据集,也可以按照感知、控制、决策等不同角度设置测试基准,并在此基础上针对不同领域设置多样化的风险,从而促进研究者将关注点聚焦于多智能体风险决策方法性能的提升,以便解决当前多智能体风险决策方法分散且难以比较的困境。

7.4 评价:风险的识别与评估研究

多智能体风险决策的目标是在控制风险的前提下尽可能最大化集体收益。然而,准确地识别风险并评估其影响对于控制风险具有重要作用,这直接关乎着决策者在追求利益最大化时能承受的最大风险限度。当前的多智能体风险决策方法似乎并不关心对风险的识别和评估,而是通过参数调整等简单的方式改变对风险的关注程度,尽管这种方式在简易控制、游戏等仿真环境下能够收敛到最优策略,但在复杂场景中对风险的粗略识别和简易评估将导致决策过程难以进行或策略失效。例如在疫苗供应链中,疫苗供应商可能面临产出不确定性导致的库存积压以及其他竞争者造成的市场份额压缩,因此疫苗供应商在进行风险决策时首先需要准确识别风险类型及来源,并在准确评估其影响程度之后才能做出最有可能最大化其利润的策略[202]。因此,在未来仍然需要对准确性更高、通用性更强的风险识别方法和评估机制进行深入研究,一种思路是引入决策树、定性定量分析相结合等管理领域的方法,以更好地解决多智能体风险决策方法在真实场景应用时面临的风险控制问题。

7.5 部署:风险决策中的伦理规范制定

随着大模型、具身智能的发展,决策算法的不透明、难解释等特征更加凸显,应用这些技术的多智能体系统在部署时不可避免地会在基本人权、社会秩序、国家安全等诸多方面带来一系列伦理风险。例如战场上多无人机协同侦查进行风险决策时,其中一架无人机锁定目标位于一处有平民的建筑内,而另外一架锁定目标为医院,那么此时多无人机在选择打击目标时就存在困境。为了应对未来智能化发展到一定程度时风险决策面临的伦理困境,提前谋划制定伦理规范是必经之路。一种思路是在国家层面通过制定法律规范应对技术发展可能带来的风险,通过强调保护隐私、避免偏见和确保透明度等明确的道德指导原则,为技术开发者提供行为准则和约束。

8 结 论

针对多智能体在充满不确定性的开放环境下应用时可能面临的策略生成难题,本文创新性地从风险决策的角度综述了当前多智能体风险决策相关理论和方法。特别的,在风险决策研究中,目前少有集人工智能、控制科学、管理科学等多领域于一体的系统性综述文章。本文首先基于对当前多智能体风险决策研究现状的深入分析,总结了多智能体决策时面临的三种类型风险来源、分析了多智能体风险决策的问题特性并对风险决策的基础知识进行了全面回顾。在此基础上,根据多智能体风险决策时学习目标的差异,重点从控制、学习和博弈三个层面分析概括了多智能体在进行风险决策时普遍采用的理论框架,并根据这三种理论框架对最近的研究进展进行了系统分类,然后总结了这些方法在当前的应用现状,并展望了未来需要重点关注的几个开放性问题。随着智能决策技术的快速发展及部署应用,因风险造成的决策失误案例层出不穷,未来应更注重多智能体风险决策时的风险控制研究,通过将控制理论、强化学习以及博弈论相结合构建起更完善的风险管理体系,从而促进风险决策技术朝着更加人性化、自动化的方向发展,推动多智能体风险决策理论和方法在实际应用中发挥更大作用。

作者贡献声明 李鹏、陈少飞为共同第一作者。

参 考 文 献

[1] Gu S, Kuba J G, Chen Y P, Du Y L, Yang L, Knoll A, et al. Safe multi-agent reinforcement learning for multi-robot control. Artificial Intelligence, 2023, 55: 103905

[2] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 2022, 319: 895-943

[3] Li Y Z, Ding Y Z, He S Y, Hu F, Duan J T, Wen G H, et al.

Artificial intelligence-based methods for renewable power system operation. Nature Reviews Electrical Engineering, 2024, 1: 163-179

[4] Zhang Y, Yue M, Wang J, Yoo S. Multi-agent graph-attention deep reinforcement learning for post-contingency grid emergency voltage control. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3340-3350

[5] Liu Z L, Zhang J Y, Liu Z H, Du H Y, Wang Z, Dusit N, et al. Cell-free XL-MIMO meets multi-agent reinforcement learning: architectures, challenges, and future directions. IEEE Wireless Communications, 2024, 1-8

[6] Yang C, Wang Y S, Lan S L, Zhu L H. Multi-agent reinforcement learning based distributed channel access for industrial edge-cloud web 3.0. IEEE Transactions on Network Science and Engineering, 2024, 40(5): 1-12

[7] Ding L F, Yan G F. A survey of the security issues and defense mechanisms of multi-agent systems. CAAI Transactions on Intelligent Systems, 2020, 15(3): 425-434 (in Chinese)
(丁俐夫，颜钢锋. 多智能体系统安全性问题及防御机制综述. 智能系统学报，2020，15(3): 425-434)

[8] Lin J Y, Kristina K, Zhang S Q, Alberto L G, Nicolas P. On the robustness of cooperative multi-agent reinforcement learning//Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW). San Francisco, USA, 2020. 62-68

[9] Son K, Kim J, Yi Y, Shin J. Disentangling sources of risk for distributional multi-agent reinforcement learning//Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, USA. 2022. 20347-20368

[10] Hao J Y, Yang T P, Tang H Y, Bai C J, Liu J Y, Meng Z P, et al. Exploration in deep reinforcement learning: from single-agent to multiagent domain. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 8762-8782

[11] Ma C, Shen C, Lin C H, Li Q, Wang Q, Li Q, Guan X H. Attacks and defenses for autonomous driving intelligence models. Chinese Journal of Computers, 2024, 47(6): 1431-1452 (in Chinese)
(马晨，沈超，蔺琛皓，李前，王骞，李琦，管晓宏. 针对自动驾驶智能模型的攻击与防御. 计算机学报，2024，47(6)：1431-1452)

[12] Wang J, He G C, Kantaros Y. Safe task planning for language-instructed multi-robot systems using conformal prediction. ArXiv, 2024, DOI: 10.48550/arXiv.2402.15368

[13] Sun C N, Huang S J, Pompili D. LLM-based multi-agent reinforcement learning: current and future directions. ArXiv, 2024, DOI: 10.48550/arXiv.2405.11106

[14] Loomes G, Sugden R. Regret theory: An alternative theory of rational choice under uncertainty. The Economic Journal, 1982, 92(368): 805-824

[15] Zhao S M, Liu X T. Idiosyncratic volatility, investor preference and stock returns: an analysis based on prospect theory. Joural of Management Sciences in China, 2020, 23(3): 100-115 (in Chinese)
(赵胜民，刘笑天. 特质风险、投资者偏好与股票收益——基于前景理论视角的分析. 管理科学学报，2020，23(3)：100-115)

[16] Joshua C P, David D B, Mayank A, Daniel R, Thomas L G. Using large-scale experiments and machine learning to discover theories of human decision-making. Science, 2021, 372: 1209-1214

[17] Brandstatter E, Gigerenzer G, Hertwig R. The priority heuristic: making choices without trade-offs. Psychological Review, 2006, 113(2): 409-432

[18] Li S, Bi Y L, Liang Z Y, Sun Y, Wang Z J, Zheng R. Bounded or unbounded rationality? The implication of equate-to-differentiate theory in economic behavior. Management Review, 2009, 21(5): 103-114 (in Chinese)
(李纾，毕研玲，梁竹苑，孙彦，汪祚军，郑蕊. 无限理性还是有限理性?——齐当别抉择模型在经济行为中的应用. 管理评论，2009，21(5)：103-114)

[19] Hoff P. Bayes-optimal prediction with frequentist coverage control. Bernoulli, 2023, 29(2): 901-928

[20] Pan L, Tabish R, Peng B, Huang L, Shimon W. Robust multi-agent reinforcement learning with model uncertainty//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Online, 2021. 1365-1377

[21] Dong Y P, Su H, Zhu J. Deep neural network interpretability analysis for adversarial examples. Acta Automatica Sinica, 2022, 48(1): 75-86 (in Chinese)
(董胤蓬，苏航，朱军. 面向对抗样本的深度神经网络可解释性分析. 自动化学报，2022，48(1)：75-86)

[22] Jiang X, Tian Y, Hua F, Xu C, Wang Y, Guo J. A survey on large language model hallucination via a creativity perspective. ArXiv, 2024, DOI: 10.48550/arXiv.2402.06647

[23] Lu S, Zhang K, Chen T, Basar T, Horesh L. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Online, 2021. 8767-8775

[24] Exarchos I, Theodorou E A, Tsiotras P. Game-theoretic and risk-sensitive stochastic optimal control via forward and backward stochastic differential equations//Proceedings of the 55th IEEE Conference on Decision and Control (CDC). Las Vegas, USA, 2016

[25] Shen S Q, Ma C N, Li C, Liu W Q, Fu Y Q, Mei S Z, et al. RiskQ: risk-sensitive multi-agent reinforcement learning value factorization//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Red Hook, USA, 2023. 34791-34825

[26] Tan J J, Li Y, Zhang X P, Pan W Q, Ruan W J. Operation of a commercial district integrated energy system considering dynamic integrated demand response: A Stackelberg game approach. Energy, 2023, 274: 126888

[27] Gawlikowski J, Tassi C R N, Ali M, et al. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 2023, 56: 1513-1589

[28] Moloud A, Farhad P, Sadiq H, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 2021, 76: 243-297

[29] Ji S L, Du T Y, Deng S G, Cheng P, Shi J, Yang M, Li B. Robustness certification research on deep learning models: A

survey. Chinese Journal of Computers, 2022, 45(1): 190-206 (in Chinese)
(纪守领, 杜天宇, 邓水光, 程鹏, 时杰, 杨珉, 李博. 深度学习模型鲁棒性研究综述. 计算机学报, 2022, 45(1): 190-206)

[30] Li Y, Chen J, Feng L. Dealing with uncertainty: A survey of theories and practices. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(11): 2463-2482

[31] Knight F H. Risk, Uncertainty and Rrofit. Houghton Mifflin Company, 1921

[32] Mark S D, David A C. Introduction to risk management and insurance, 10th edition. NewYork, USA: Pearson Education Inc., 2021

[33] Keramati R, Dann C, Tamkin A, Brunskill E. Being optimistic to be conservative: Quickly learning a cvar policy//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). California, USA, 2020. 4436-4443

[34] Yao R, Guo H J. A multiattribute group decision-making method based on a new aggregation operator and the means and variances of interval-valued intuitionistic fuzzy values. Scientific Reports, 2022, 12: 22525

[35] Li W, Xi Y M, Chen S W. Review of process organizing research of group decision-making. Joural of Management Sciences in China, 2002, 5(2): 55-66 (in Chinese)
(李武, 席酉民, 成思危. 群体决策过程组织研究述评. 管理科学学报, 2002, 5(2): 55-66)

[36] Liu Y M, Wei X H, Chen X H. The effects of group emotional intelligence on group decision-making behaviors and outcomes. Joural of Management Sciences in China, 2011, 14(10): 11-27 (in Chinese)
(刘咏梅, 卫旭华, 陈晓红. 群体情绪智力对群决策行为和结果的影响研究. 管理科学学报, 2011, 14(10): 11-27)

[37] Maria K, Christian H, Stepan S, Krishnendu C, Martin A N. The effect of environmental information on evolution of cooperation in stochastic games. Nature Communications, 2023, 14: 4153

[38] Jiang Y, Liu Q, Ma X, Li C, Yang Y, Yang J, et al. Learning diverse risk preferences in population-based self-play//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada, 2024. 12910-12918

[39] Rice N, Robone S. The effects of health shocks on risk preferences: do personality traits matter? Journal of Economic Behavior & Organization, 2022, 204: 356-37

[40] Blavatskyy P R. Stochastic expected utility theory. Journal of Risk and Uncertainty, 2007, 34: 259-286

[41] Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty, 1992, 5: 297-323

[42] Gul F. A theory of disappointment aversion. Econometrica: Journal of the Econometric Society, 1991: 667-686

[43] Birnbaum M H, Chavez A. Tests of theories of decision making: violations of branch independence and distribution independence. Organizational Behavior and Human Decision Processes, 1997, 71(2): 161-194

[44] Bernoulli D. Exposition of a new theory on the measurement of risk. The Kelly Capital Growth Investment Criterion: Theory and practice, 2011, 11-24

[45] Savage L J. The foundations of statistics. Massachusetts: Courier Corporation, 1972

[46] Quiggin J. A theory of anticipated utility. Journal of Economic Behavior & Organization, 1982, 3(4): 323-343

[47] Glimcher P W, Dorris M C, Bayer H M. Physiological utility theory and the neuroeconomics of choice. Games and Economic Behavior, 2005, 52(2): 213-256

[48] Tversky A. Intransitivity of preferences. Psychological Review, 1969, 76(1): 31-48

[49] Simon H A. A behavioral model of rational choice. The Quarterly Journal of Economics, 1955, 69(1): 99-118

[50] Wang Z J, Li S. Tests of the integrative model and priority heuristic model from the point of view of choice process: evidence from an eye-tracking study. Acta Psychologica Sinica, 2012, 44(2): 179-198 (in Chinese)
(汪祚军, 李纾. 对整合模型和占优启发式模型的检验:基于信息加工过程的眼动研究证据. 心理学报, 2012, 44(2): 179-198)

[51] Brandstatter E, Korner C. Attention in risky choice. ACTA Psychologica, 2014, 152: 166-176

[52] Van D A C, Figner B, Weeda W D, Van d M M W, Jansen B R, Huizenga H M. Neural mechanisms underlying compensatory and noncompensatory strategies in risky choice. Journal of Cognitive Neuroscience, 2016, 28(9): 1358-1373

[53] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning//Proceedings of the International Conference on Machine Learning (ICML). New York, USA, 2016: 1050-1059

[54] Lundberg S M, Lee S I. A unified approach to interpreting model predictions//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017, 4768-4777

[55] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019, 3681-3688

[56] Pedreschi D, Giannotti F, Guidotti R, et al. Meaningful explanations of black box AI decision systems//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019, 9780-9784

[57] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20: 273-297

[58] Norton M D, Royset J O. Diametrical risk minimization: theory and computations. Machine Learning, 2023, 112: 2933-2951

[59] Perlaza S M, Bisson G, Esnaola I, Jean-Marie A, Rini S. Empirical risk minimization with relative entropy regularization. IEEE Transactions on Information Theory, 2024, 70(7): 5122-5161

[60] Ibrahim F G. Capacitive empirical risk function-based bag-of-words and pattern classification processes. Pattern Recognition, 2023, 139: 109482

[61] Li T, Ahmad B, Maziar S, Virginia S. On tilted losses in machine learning: theory and applications. Journal of Machine Learning Research, 2023, 24(142): 1-79

[62] Hado H. Double Q-learning//Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2010. 6382-6393

[63] Thrun S, Anton S. Issues in using function approximation for reinforcement learning//Proceedings of the 4th Connectionist Models Summer School. Erlbaum Associates, 1993. 255-263

[64] Oron A, Nir B, Nahum S. Averaged-DQN: variance reduction and stabilization for deep reinforcement learning//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017. 176-185

[65] Hasselt H v, Guez A, Silver D. Deep reinforcement learning with double Q-Learning//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI). Phoenix, Arizona, 2016. 2094-2100

[66] Scott F, Herke H, David M. Addressing function approximation error in actor-critic methods//Proceedings of the 35th International Conference on Machine Learning. Vienna, Austria, 2018. 1587-1596

[67] Zhang F, Li J, Li Z. A TD3-based multi-agent deep reinforcement learning method in mixed cooperation-competition environment. Neurocomputing, 2020, 411: 206-215

[68] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017. 6382-6393

[69] Seungchan K, Kavosh A, Michael L, George K. DeepMellow: removing the need for a target network in deep q-Learning//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao, China, 2019, 2733-2739

[70] Gan Y Z, Zhang Z, Tan X Y. Stabilizing Q learning via soft mellowmax operator//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Online, 2021, 7501-7509

[71] Zhao S, Ron P, Lawrence C. Revisiting the softmax bellman operator: new benefits and new perspective//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019, 5916-5925

[72] Guan T, Liu F, Wu X, Xian R, Li Z, Liu X, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024. 14375-14385

[73] Martino A, Iannelli M, Truong C. Knowledge injection to counter large language model (LLM) hallucination//Proceedings of the European Semantic Web Conference. Switzerland, 2023. 182-185

[74] Zhang S, Pan L, Zhao J, Wang W Y. The knowledge alignment problem: bridging human and external knowledge for large language models. ArXiv, 2023, DOI: 10.48550/arXiv.2305.13669

[75] Sarmah B, Mehta D, Pasquali S, Zhu T. Towards reducing hallucination in extracting information from financial reports using large language models//Proceedings of the 3rd International Conference on AI-ML Systems. New York, USA, 2023. 1-5

[76] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. ArXiv, 2023, DOI: 10.48550/arXiv.2309.01219

[77] Liu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, et al. A survey on hallucination in large vision-language models. ArXiv, 2024, DOI: 10.48550/arXiv.2402.00253

[78] Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. ArXiv, 2023, DOI: 10.48550/arXiv.2309.05922

[79] Sebastian F, Jannik K, Lorenz K, Yarin G. Detecting hallucinationsin large language models using semanticentropy. Nature, 2024, 630: 625-630

[80] Huang J, Huang M. Robust mean field linear-quadratic-gaussian games with unknown L-2-disturbance. SIAM Journal on Control and Optimization, 2017, 55(5): 2811-2840

[81] Fu M H, Li S W, Guo M Z, Yang Z F, Sun Y X, Qiu C X, et al. Cooperative decision-making of multiple autonomous vehicles in A connected mixed traffic environment. Transportation Research Part C: Emerging Technologies, 2023, 157: 104415

[82] Liu D R, Xue S, Zhao B, Luo B, Wei Q L. Adaptive dynamic programming for control: a survey and recent advances. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(1): 142-160

[83] Bo L, Wang S, Yu X. Mean field game of optimal relative investment with jump risk. Science China Mathematics, 2024, 67(5): 1159-1188

[84] Chang Y Y, Firoozi D, Benatia D. Large banks and systemic risk: Insights from a mean-field game model. ArXiv, 2023, DOI: 10.48550/arXiv.2305.17830

[85] Liu Q, Li J H, Wang H, Zeng J X, Chai Y. Stochastic variational bayesian learning of Wiener model in the presence of uncertainty. Acta Automatica Sinica, 2024, 50(6): 1185-1198 (in Chinese)
(刘切，李俊豪，王浩，曾建学，柴毅. 不确定性环境下维纳模型的随机变分贝叶斯学习. 自动化学报，2024，50(6)：1185-1198)

[86] Ding S F, Du W, Zhang J, Guo L L, Ding L. Research progress of multi-agent deep reinforcement learning. Chinese Journal of Computers, 2024, 47(7): 1547-1567 (in Chinese)
(丁世飞，杜威，张健，郭丽丽，丁玲. 多智能体深度强化学习研究进展. 计算机学报，2024，47(7)：1547-1567)

[87] Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017. 449-458

[88] Yekkehkhany A, Murray T, Nagi R. Risk-averse equilibrium for games. ArXiv, 2020, DOI: 10.48550/arXiv.2002.08414

[89] Yekkehkhany A, Nagi R. Risk-averse equilibrium for autonomous vehicles in stochastic congestion games. ArXiv, 2020, DOI: 10.48550/arXiv.2007.09771

[90] Hamadene S, Mu R. Risk-sensitive nonzero-sum stochastic differential game with unbounded coefficients. Dynamic Games

and Applications, 2021, 11(1): 84-108

[91] Chen S, Liu Y C, Weng C G. Dynamic risk-sharing game and reinsurance contract design. Insurance: Mathematics and Economics, 2019, 86: 216-231

[92] Ghosh M K, Kumar K S, Pal C, Pradhan S. Nonzero-sum risk-sensitive stochastic differential games: a multi-parameter eigenvalue problem approach. Systems & Control Letters, 2023, 172: 105443

[93] Biswas A, Saha S. Zero-sum stochastic differential game with risk-sensitive cost. Applied Mathematics & Optimization, 2020, 81(1): 113-140

[94] Patil A, Zhou Y J, Fridovich K D, Tanaka T. Risk-minimizing two-player zero-sum stochastic differential game via path integral control//Proceedings of the 62nd IEEE Conference on Decision and Control (CDC). Singapore, 2023. 1038-1043

[95] Shawon D, Hao X. Extended mean field game theoretical optimal distributed control for large scale multi-agent systems: An efficiency-complexity tradeoff, Information Sciences, 2025, 719: 122432

[96] Li Z, Reppen A M, Sircar R. A mean field games model for cryptocurrency mining. Management Science, 2024, 70(4): 2188-2208

[97] Tembine H, Zhu Q, Basar T. Risk-sensitive mean-field games. IEEE Transactions on Automatic Control, 2013, 59(4): 835-850

[98] Bonnans J F, Lavigne P, Pfeiffer L. Discrete-time mean field games with risk-averse agents. ESAIM: Control, Optimisation and Calculus of Variations, 2021, 27: 44

[99] Moon J, Basar T. Discrete-time decentralized control using the risk-sensitive performance criterion in the large population regime: a mean field approach//Proceedings of the American Control Conference (ACC). Chicago, USA, 2015. 4779-4784

[100] Liu H, Firoozi D, Breton M. LQG risk-sensitive mean field games with a major agent: a variational approach. ArXiv, 2023, DOI: 10.48550/arXiv.2305.15364

[101] Djehiche B, Tembine, H. Risk-sensitive mean-field type control under partial observation//Proceedings of the Stochastics of Environmental and Financial Economics. Oslo, Norway: Springer, 2014-2015. 243-263

[102] Saldi N, Basar T, Raginsky M. Partially observed discrete-time risk-sensitive mean field games. Dynamic Games and Applications, 2023, 13(3): 929-960

[103] Moon J, Basar T. Linear quadratic risk-sensitive and robust mean field games. IEEE Transactions on Automatic Control, 2016, 62(3): 1062-1077

[104] Barreiro-Gomez J, Duncan T E, Tembine H. Matrix-valued mean-field-type games: risk-sensitive, adversarial, and risk-neutral linear-quadratic case. ArXiv, 2019, DOI: 10.48550/arXiv.1904.11346

[105] Ren X Y, Firoozi D. Risk-sensitive mean field games with common noise: a theoretical study with applications to interbank markets. ArXiv, 2024, DOI: 2403.03915

[106] Tchuendom R F, Malhame R, Caines P. A quantilized mean field game approach to energy pricing with application to fleets of plug-in electric vehicles//Proceedings of the IEEE 58th Conference on Decision and Control (CDC). Nice, France, 2019. 299-304

[107] Tembine H. Risk-sensitive mean-field-type games with Lp-norm drifts. Automatica, 2015, 59: 224-23

[108] Moon J, Basar T. Risk-sensitive mean field games via the stochastic maximum principle. Dynamic Games and Applications, 2019, 9(4): 1100-1125

[109] Nguyen-Tang T, Gupta S, Venkatesh S. Distributional reinforcement learning via moment matching//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). California, USA, 2021. 9144-9152

[110] Peng Z N, Hu J P, Luo R, Ghosh B K. Distributed multi-agent temporal-difference learning with full neighbor information. Journal of Control Theory and Applications, 2020, 18(4): 379-389

[111] Lowet A S, Zheng Q, Matias S, Drugowitsch J, Uchida N. Distributional reinforcement learning in the brain. Trends in Neurosciences, 2020, 43(12): 980-997

[112] Sheng W D, Guo H L, Yau W Y, Zhou Y J. PD-FAC: probability density factorized multi-agent distributional reinforcement learning for multi-robot reliable search. IEEE Robotics and Automation Letters, 2022, 7(4): 8869-8876

[113] Qiu W, Wang X R, Yu R S, Wang R D, He X, An B, et al. Rmix: learning risk-sensitive policies for cooperative reinforcement learning agents//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Online, 2021. 23049-23062

[114] Sun W F, Lee C K, Lee C Y. DFAC framework: factorizing the value function via quantile mixture for multi-agent distributional Q-learning//Proceedings of the 38th International Conference on Machine Learning (ICML). Online, 2021. 9945-9954

[115] Zhang J Y, Bedi A S, Wang M D, Koppel A. Cautious reinforcement learning via distributional risk in the dual domain. IEEE Journal on Selected Areas in Information Theory, 2021, 2(2): 611-626

[116] Chow Y, Ghavamzadeh M, Janson L, Pavone M. Risk-constrained reinforcement learning with percentile risk criteria. The Journal of Machine Learning Research, 2017, 18(1): 6070-6120

[117] Kumar M, Dutt V. Understanding decisions in collective risk social dilemma games using reinforcement learning. IEEE Transactions on Cognitive and Developmental Systems, 2020, 12(4): 824-840

[118] Ghafarian S H, Yazdi H S. Prepare for the worst, hope for the best: active robust learning on distributions. IEEE Transactions on Cybernetics, 2022, 52(6): 5573-5586

[119] Fei Y J, Yang Z R, Chen Y D, Wang Z R, Xie Q M. Risk-sensitive reinforcement learning: near-optimal risk-sample tradeoff in regret//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 2020. 22384-22395

[120] He S H, Han S Y, Su S B, Han S, Zou S F, Miao F. Robust

multi-agent reinforcement learning with state uncertainty. Transactions on Machine Learning Research, 2023, https://openreview.net/forum?id=CqTkapZ6H9

[121] Han S Y, Su S B, He S H, Han S, Yang H Z, Miao F. What is the solution for state adversarial multi-agent reinforcement learning? ArXiv, 2022, DOI: 10.48550/arXiv.2212.02705

[122] Zhou Z Y, Liu G J, Zhou M C. A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(10):1-12

[123] Zheng L, Yang J, Cai H, Zhou M, Zhang W, Wang J, Yu Y. MAgent: A Many-Agent Reinforcement Learning Platform for Artificial Collective Intelligence//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018. 8222-8223

[124] Zhang Z L, Sun Y C, Huang F R, Miao F. Safe and robust multi-agent reinforcement learning for connected autonomous vehicles under state perturbations. ArXiv, 2023, DOI: 10.48550/arXiv.2309.11057

[125] Li S H, Wu Y, Cui X Y, Dong H H, Fang F, Russell S. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019. 4213-4220

[126] Li S M, Guo J, Xiu J Q, Yu X, Wang J K, Liu A S, et al. Byzantine robust cooperative multi-agent reinforcement learning as a bayesian game// Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria, 2024

[127] Zhang K Q, Sun T, Tao Y Z, Genc S, Mallya S, Basar T. Robust multi-agent reinforcement learning with model uncertainty//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2020. 10571-10583

[128] Xue W Q, Qiu W, An B, Rabinovich Z, Obraztsova S, Yeo C K. Mis-spoke or mis-lead: achieving robustness in multi-agent communicative reinforcement learning//Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Auckland, New Zealand, 2022. 1418-1426

[129] Xiao Y C, Lyu X G. Local advantage actor-critic for robust multi-agent deep reinforcement learning//Proceedings of the International symposium on multi-robot and multi-agent systems (MRS). Cambridge, UK, 2021. 155-163

[130] Bukharin A, Li Y, Yue Y, Zhang Q R, Chen Z H, Zuo S M, et al. Robust multi-agent reinforcement learning via adversarial regularization: theoretical foundation and stable algorithms// Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2023. 68121-68133

[131] Rokhforoz P, Montazeri M, Fink O. Safe multi-agent deep reinforcement learning for joint bidding and maintenance scheduling of generation units. Reliability Engineering & System Safety, 2023, 232: 109081

[132] Guo G D, Zhang M F, Gong Y F, Xu Q W. Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay. Applied Energy, 2023, 349: 121648

[133] Wang X S, Wang R R, Cheng Y H. Safe reinforcement learning: A survey. Acta Automatica Sinica, 2023, 49(9): 1813-1835 (in Chinese)
(王雪松，王荣荣，程玉虎. 安全强化学习综述. 自动化学报，2023，49(9)：1813-1835)

[134] Garg K, Zhang S Y, So O, Dawson C, Fan C C. Learning safe control for multi-robot systems: methods, verification, and open challenges. Annual Reviews in Control, 2024, 57: 100948

[135] Semnani S H, Liu H, Everett M, de R A, How J P. Multi-agent motion planning for dense and dynamic environments via deep reinforcement learning. IEEE Robotics and Automation Letters, 2020, 5(2): 3221-3226

[136] Long P X, Fan T X, Liao X Y, Liu W X, Zhang H, Pan J. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning//Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia, 2018. 6252-6259

[137] Gu S D, Kuba J G, Chen Y P, Du Y L, Yang L, Knoll A, et al. Safe multi-agent reinforcement learning for multi-robot control. Artificial Intelligence, 2023, 319: 103905

[138] Zheng Z, Gu S D. Safe multi-agent reinforcement learning with bilevel optimization in autonomous driving. ArXiv, 2024, DOI: 10.48550/arXiv.2405.18209

[139] Vinod A P, Safaoui S, Chakrabarty A, Quirynen R, Yoshikawa N, Di Cairano S. Safe multi-agent motion planning via filtered reinforcement learning//Proceedings of the 2022 International Conference on Robotics and Automation (ICRA). Philadelphia, USA, 2022. 7270-7276

[140] Ding D S, Wei X H, Yang Z R, Wang Z R, Jovanovic M. Provably efficient generalized lagrangian policy optimization for safe multi-agent reinforcement learning//Proceedings of the 5th Annual Conference on Learning for Dynamics and Control. Philadelphia, USA, 2023. 315-332

[141] ElSayed-Aly I, Bharadwaj S, Amato C, Ehlers R, Topcu U, Lu Fg. Safe multi-agent reinforcement learning via shielding// Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). Online, 2021. 483-491

[142] Zhang Z L, Han S Y, Wang J W, Miao F. Spatial-temporal-aware safe multi-agent reinforcement learning of connected autonomous vehicles in challenging scenarios//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK, 2023. 5574-5580

[143] Cai Z Y, Cao H H, Lu W J, Zhang L, Xiong H. Safe multi-agent reinforcement learning through decentralized multiple control barrier functions. ArXiv, 2021, DOI: 10.48550/arXiv2103.12553, 2021

[144] Melcer D, Amato C, Tripakis S. Shield decentralization for safe multi-agent reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2022. 13367-13379

[145] Melcer D, Amato C, Tripakis S. Shield decentralization for safe reinforcement learning in general partially observable multi-agent environments//Proceedings of the 23rd International Conference on Autonomous Agents and Multi-agent Systems (AAMAS). Auckland, New Zealand, 2024. 2384-2386

[146] Chen J, Shi D W, Cai D H, Wang J Z, Zhu L L. Data-driven Bayesian optimization method for intermittent hypoxic training strategy decision. Acta Automatica Sinica, 2023, 49(8): 1667-1678 (in Chinese)
(陈婧，史大威，蔡德恒，王军政，朱玲玲. 数据驱动的间歇低氧训练贝叶斯优化决策方法. 自动化学报，2023，49(8)：1667-1678)

[147] Chalkiadakis G, Boutilier C. Coordination in multi-agent reinforcement learning: a Bayesian approach//Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). New York, USA, 2003. 709-716

[148] Foerster J, Song F, Hughes E, Burch N, Dunning I, Whiteson S, et al. Bayesian action decoder for deep multi-agent reinforcement learning//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019. 1942-1951

[149] Wang Z Y, Du Y L, Zhang Y D, Fang M, Huang B W. MACCA: offline multi-agent reinforcement learning with causal credit assignment. ArXiv, 2023, DOI: 10.48550/arXiv.2312.03644

[150] Chen D Y, Zhang Q. Context-aware bayesian network actor-critic methods for cooperative multi-agent reinforcement learning//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA, 2023. 5327-5350

[151] Xinqi D, Hechang C, Che W, et al. Robust multi-agent reinforcement learning via Bayesian distributional value estimation. Pattern Recognition, 2024, 145: 109917

[152] Eriksson H, Basu D, Alibeigi M, Dimitrakakis C. Risk-sensitive bayesian games for multi-agent reinforcement learning under policy uncertainty. ArXiv, 2022, DOI: 10.48550/arXiv.2203.10045

[153] Yang F, Vereshchaka A, Chen C, Dong W. Bayesian multi-type mean field multi-agent imitation learning//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2020. 2469-2478

[154] Chen M, Li Y, Wang E, Yang Z, Wang Z, Zhao T. Pessimism meets invariance: provably efficient offline mean-field multi-agent RL//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Online, 2021. 17913-17926

[155] Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J. Mean field multi-agent reinforcement learning//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018. 5571-5580

[156] Ding S, Du W, Ding L, Guo L, Zhang J, An B. Multi-agent dueling Q-learning with mean field and value decomposition. Pattern Recognitions, 2023, 139: 109436

[157] KurachK, RaichukA, StanczykP, ZajacM, BachemO. Google research football: a novel reinforcement learning environment//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020. 4501-4510

[158] WangW, LiuY, SrikantR, YingL. 3M-RL: multi-resolution, multi-agent, mean-field reinforcement learning for autonomous UAV routing. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(7): 8985-8996

[159] DeyS, XuH. Large-scale multi-agent system optimization with fixed final density constraints: an imbalanced mean-field game theory//American Control Conference (ACC). Toronto, Canada, 2024. 869-874

[160] Zhang Z Y, Wang C Y, Zhao W Z, Cao M C, Liu J Q, Xu K H. Path-speed decoupling planning method based on risk cooperative game for intelligent vehicles. IEEE Transactions on Transportation Electrification, 2024, 10(2): 3792-3806

[161] Zheng X J, Huang H Y, Wang J Q, Zhao X C, Xu Q. Behavioral decision-making model of the intelligent vehicle based on driving risk assessment. Computer-Aided Civil and Infrastructure Engineering, 2021, 36(7): 820-837

[162] Vimala K, Bharathi S. Cooperative solutions for TU and NTU stochastic game-theoretical approach. International Journal of Health Sciences (III), 2022. 431-455

[163] Shi Z, Cao E. Risk pooling cooperative games in contract farming. Canadian Journal of Agricultural Economics, 2021, 69(3): 117-139

[164] Campos L. Fuzzy linear programming models to solve fuzzy matrix games. Fuzzy Sets and Systems, 2020, 32(3): 275-289.

[165] Wu D R, Zeng Z G, Mo H, Wang F Y. Interval type-2 fuzzy sets and systems: overview and outlook. Acta Automatica Sinica, 2020, 46(8): 1539-1556 (in Chinese)
(伍冬睿，曾志刚，莫红，王飞跃. 区间二型模糊集和模糊系统：综述与展望. 自动化学报，2020，46(8)：1539-1556)

[166] Oliver S, David H M, Stefano B B, Stephen M M, Yang Y-D, Wang J. A game-theoretic framework for managing risk in multi-agent systems//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA. 2023. 32059-32087

[167] Yang Z S, Yang Z L, Yin J B, Qu Z H. A risk-based game model for rational inspections in port state control. Transportation Research: part E, 2018, 118: 477-495

[168] Cui Y, Quddus N, Mashuga C V. Bayesian network and game theory risk assessment model for third-party damage to oil and gas pipelines. Process Safety & Environmental Protection: Transactions of the Institution of Chemical Engineers Part B, 2020, 134: 178-188

[169] Fan H W, Lu J, Chang Z. A risk-based game theory model of navy and pirate behaviors. Ocean and Coastal Management, 2022, 225: 106200

[170] Tatarenko T, Kamgarpour M. Learning generalized nash equilibria in a class of convex games. IEEE Transactions on Automatic Control, 2019, 64(4): 1426-1439

[171] Karmakar S, Seikh M R, Castillo O. Type-2 intuitionistic fuzzy matrix games based on a new distance measure: application to biogas-plant implementation problem. Applied Soft Computing, 2021, 106: 107357

[172] Dong J Y, Wan S P. Type-2 interval-valued intuitionstic fuzzy matrix game and application to energy vehicle industry development. Expert Systems with Applications, 2024, 249: 123398

[173] Li Y, Quevedo D E, Dey S, Shi L. A game-theoretic approach to fake-acknowledgment attack on cyber-physical systems. IEEE Transactions on Signal and Information Processing over Networks, 2017, 3(1): 1-11

[174] Eisenstadt E, Moshaiov A. Novel solution approach for multi-objective attack-defense cyber games with unknown utilities of the opponent. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(1): 16-26

[175] Jakobik A, Palmieri F, Kolodziej J. Stackelberg games for modeling defense scenarios against cloud security threats. Journal of Network & Computer Applications, 2018, 110: 99-107

[176] An H, Yang R, Ma X, Zhang S, Islam S M. An evolutionary game theory model for the inter-relationships between financial regulation and financial innovation. The North American Journal of Economics and Finance, 2021, 55: 101341

[177] Li S, Wang B. Evolutionary game simulation of corporate investing and financing behavior from a risk perspective. Cluster Computing, 2019, 22: 5955-5964

[178] Huang W J, Pham V H, Haskell W B. Model and reinforcement learning for markov games with risk preferences//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020. 2022-2029

[179] Godbout M, Heuillet M, Raparthy S C, Bhati R, Durand A. A game-theoretic perspective on risk-sensitive reinforcement learning//Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI). Vancouver, Canada, 2022

[180] Wang Z F, Shen Y, Zavlanos M. Risk-averse no-regret learning in online convex games//Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, Maryland, USA, 2022. 22999-23017

[181] Ghaemi H, Kebriaei H, Moghaddam A R, Ahamdabadi M N. Risk-sensitive multi-agent reinforcement learning in network aggregative markov games//Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Auckland, New Zealand, 2024. 2282-2284

[182] Ma X, Li Y, Gao Y. Decision model of intrusion response based on markov game in fog computing environment. Wireless Networks, 2023, 29(8): 3383-3392

[183] Capponi A, Ghanadan R, Stern M. Risk-Sensitive Cooperative Games for human-machine Systems. ArXiv, 2017, DOI: 10.48550/arXiv.1705.09580

[184] Ding F, Zhang N, Li S B, Li K Q. A survey of architecture and key technologies of intelligent connected vehicle-road-cloud cooperation system. Acta Automatica Sinica, 2022, 48(12): 2863-2885 (in Chinese)
(丁飞，张楠，李升波，边有钢，童恩，李克强. 智能网联车路云协同系统架构与关键技术研究综述. 自动化学报，2022，48(12)：2863-2885)

[185] Liu Q, Liu Z, Xiong B, Xu W, Liu Y. Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function. Advanced Engineering Informatics, 2021, 49: 101360

[186] Patel B N, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digital medicine, 2019, 2: 111

[187] Shi X P, Wong Y D, Chai C, Li M Z F. An automated machine learning (AutoML) method of risk prediction for decision-making of autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(11): 7145-7154

[188] Maximilian G, Rainer T, Gemb K, Markus L. Maximum acceptable risk as criterion for decision-making in autonomous vehicle trajectory planning. IEEE Open Journal of Intelligent Transportation Systems, 2023, 4: 570-579

[189] Wang Y S, Wang C Y, Zhao W Z, Xu C. Decision-making and planning method for autonomous vehicles based on motivation and risk assessment. IEEE Transactions on Vehicular Technology, 2021, 70(1): 107-120

[190] Nguyen H D, Choi M, Han K. Risk-informed decision-making and control strategies for autonomous vehicles in emergency situations. Accident Analysis & Prevention, 2023, 193: 107305

[191] Zhou W, Chen D, Yan J, Li Z, Yin H, Ge W. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. Autonomous Intelligent Systems, 2022, 2: 5

[192] Li G F, Yang Y F, Zhang T R, Qu X D, Cao D P, Cheng B, et al. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. Transportation Research Part C: Eemerging Technologies, 2021, 122: 102820

[193] Wu T, Zhou P, Liu K, Yuan Y, Wang X, Huang H, Wu D O. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8243-8256

[194] Ali Y, Rakesh N. Risk-averse equilibria for vehicle navigation in stochastic congestion games. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18719-18735

[195] Xu Y, Yang Z, Gu W, Li M, Deng Z. Robust real-time distributed optimal control based energy management in a smart grid. IEEE Transactions on Smart Grid, 2017, 8 (4): 1568-1579

[196] Zhang Q Z, Kaveh D, Wang Z Y, Qiu F, Zhao D B. Multi-agent safe policy learning for power management of networked microgrids. IEEE Transactions on Smart Grid, 2021, 12(2): 1048-1062

[197] Shi Y F, Feng M X, Wang M R, Zhou W A, Li H Q. Multi-agent reinforcement learning with safety layer for active voltage control//Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS). London, UK, 2023. 1533-1541

[198] Ni Z, Paul S. A multistage game in smart grid security: a reinforcement learning solution. IEEE Transactions on Neural

Networks and Learning Systems, 2019, 30(9): 2684-2695

[199] Xu X H, Hou Y Z. Research status and development trend of large group risk decision-making theory and method. Journal of University of Electronic Science and Technology of China, 2021, 23(4): 1-6 (in Chinese)
(徐选华，侯宇舟. 大群体风险决策理论与方法研究现状及发展趋势. 电子科技大学学报社科版，2021，23(4)：1-6)

[200] Sun J, Chen S, Zhang C, Ma Y, Zhang J. Decision-Making with speculative opponent models. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(3):1-15

[201] Friedrich B, Michael K, Tim M. The minds of many: Opponent modeling in a stochastic game//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI). Melbourne, Australia, 2017. 3845-3851

[202] Cai J H, Jia L S, Zhou Q, Wang N N, Hu X Q. Coordination model of VMI supply chain considering mean-variance analysis. Journal of Management Sciences in China, 2023, 26(3): 20-43 (in Chinese)
蔡建湖，贾利爽，周青，王楠楠，胡晓青. 考虑均值-方差风险量化的VMI供应链协调模型. 管理科学学报，2023，26(3)：20-43)

LI Peng, Ph. D. candidate. His research interest is risk sensitive reinforcement learning and multi-agent risk decision-making.

CHEN Shao-Fei, Ph. D., associate professor. His research interests are computer game theory and intelligent decision-making.

YI Chu-Shu, Ph. D. candidate. Her research interest is optimal control.

LI Shun, Ph. D. candidate. His research interest is multi-agent reinforcement learning.

XING Jun-Liang, Ph. D., professor. His research interests are computer game theory and intelligent game interaction.

CHEN Jing, Ph. D., professor. His research interests are intelligent decision-making and computer game theory.

Background

In recent years, multi-agent systems have been widely applied in fields such as transportation, power, and logistics, significantly driving socio-economic development. However, in the real world, the deployment of multi-agent systems in open environments often faces challenges in risk decision-making due to various uncertainties. This is particularly highlighted by the current developments in large models and embodied AI, which further emphasize the need for risk decision-making in multi-agent systems operating within open environments.

This review study primarily falls within the interdisciplinary area of computer science, control science, management science, and artificial intelligence. There is a substantial body of work focused on proposing multi-agent risk decision-making methods with different characteristics from various disciplines and fields. However, there have been few attempts to provide a comprehensive overview and review of the issue of multi-agent risk decision-making. This paper reviews the relevant literature on theories and methods of multi-agent risk decision-making. The main purpose of this paper is to better understand the development and future directions of multi-agent risk decision-making theories and methods, and to further promote research on risk decision-making problems of multi-agent systems in open environments, thereby improving the quality of decision-making in multi-agent systems under risky conditions.

This paper first introduces the basics of single-agent risk decision-making, including risk decision-making theory, common risks and their mitigation methods, and risk measurement. Subsequently, it analyzes the three most popular multi-agent risk decision-making theories from the perspectives of control, learning, and game theory: optimal control theory, reinforcement learning theory, and game equilibrium theory. It then reviews the three categories of multi-agent risk decision-making methods corresponding to these theories found in the current research literature. Finally, it summarizes the applications of multi-agent risk decision-making methods in real-world domains such as human-machine collaboration and discusses meaningful topics that require attention in the future of multi-agent risk decision-making. We hope that this paper will provide an overall perspective and reference for future research, thereby enhancing the applicability of future multi-agent risk decision-making methods.

This work was supported by the National Natural Science Foundation of China (No. 62376280).