梯度的方法，这使得它能在连续动作或更高维动作空间中取合适的动作；而 Q 学习难以实现这个目标，甚至会瘫痪。相比单纯策略梯度，AC 算法应用了 Q 学习或其他策略评估方法，使得 AC 算法能进行单步更新而不是回合更新，比单纯的策略梯度的效率更高。

3.1.3　深度确定性策略梯

深度 Q 网络是一种基于价值函数的方法，难以应对大连续动作空间，无法输出离散状态动作值。Lillicrap 等[32]于 2015 年提出的深度确定性策略梯度(Deep Deterministic Policy Gradients，DDPG)是基于上述 AC 算法的，结合确定性策略梯度算法(Deterministic Policy Gradients，DPG)，在动作输出方面采用一个网络来拟合策略函数，直接输出动作，可以应对连续动作的输出以及更大的动作空间。此外，AC 模型还衍生出很多种算法，如异步优势行动评论家算法(A3C)[33]和分布式近似策略优化算法(DPPO)[34-35]。A3C 算法由 Mnih 于 2016 年提出，在 DQN 中为了破坏训练样本之间的相关性，采用了经验重放技术，即把训练样本缓存起来，每次训练时从中随机抽取一个 minibatch。在 A3C 中，利用多线程并行地去采集数据，每个线程以一个独立的智能体形式去搜索独立环境；同时，每个智能体还可以平行地利用不同的探索策略进行采样，这样每个线程得到的样本天然不相关，而且采样速度也更快。

3.2　MARL 算法进

3.2.1　可扩展

可扩展性是目前 MARL 领域的核心关注点。早期的智能体强化学习研究都是应用于小问题，在离散的动作和状态空间上对 Q 学习算法进行改进和完善。但是将算法扩展到现实的多智能体问题时，其中状态和动作空间是很大甚至连续的，Q 函数的表格存储变得不切实际或不可能。受强化学习与深度学习技术相结合的启发，研究人员把 DQN 和 DPG 等技术应用到多智能体强化学习中。近期的大部分研究集中于此。

将 DQN 泛化到 MARL 存在的最大问题是经验重放方法变得不再适用，如果不知道其他智能体的状态，那么不同情况下自身的状态转移概率会不同。Foerster 等[36]提出了两种方法对经验重放方法进行改进，使多智能体在使用经验重放技术时更稳定并更具兼容性。一方面使用一个运用重要性抽样的多智能体变量来自然衰减过时数据，将经验重放的数据解释为环境外数据[37]。由于较旧的数据往往会产生较低的重要性，因此这种方法能避免非固定重放经验数据产生的混淆。另一方面，让每个智能体通过观察其他智能体的决策来推测其他智能体的行为，这一方法能适用于更大范围的深度网络。其算法在传统多智能体算法 IQL[38-41]上做了改进，IQL 算法的全称是 independent Q-learning，顾名思义，其是对每个智能体独立地执行一次 Q-learning 算法，将其他智能体视为环境的一部分。IQL 的点在于它忽略了这样一个事实：这些智能体的策略会随着时间的推移而变化，使其自身的 Q 函数变得非平稳。而研究人员通过让每个智能体推断其他智能体的行为并学习一种策略来规避其他智能体的策略，从而改善了 IQL 的非平稳性。

Yang 等[42]提出了平均场强化学习对多智能体系统中动态行为进行建模，有效地回应了相邻智能体的平均效应，通过将多智能体问题简化为两个智能体的问题，解决了智能体数量增加导致的维数和指数增加的问题。他们没有分别考虑单个智能体对其他个体产生的不同影响，只是将领域内所有其他个体的影响用一个均值来代替，这样，对于每个个体，只需要考虑个体与这个均值的交互作用即可。应用平均场论后，学习在两个智能体之间是相互促进的：单个智能体的最优策略的学习是基于智能体群体动态的；同时，集体的动态也根据个体的策略进行更新。

还有研究人员对 DQN 在多智能体上的应用做了很多作。Tuyls 等[43]提出了 LDQN 算法，将宽容政策引入深度 Q 网络，采用宽大处理方法更新消极政策，使收敛性和稳定性都得到提高。Zheng 等[44]提出了一个多智能体深度强化学习框架，称其为加权双深 Q 网络(WDDQN)。通过利用深度神经网络和加权双估计器，WDDQN 不仅可以有效减小偏差，而且可以扩展到许多深度强化学习场景。根据经验，WDDQN 在随机合作环境中的性能和收敛效果优于现有的 DRL 算法(双 DQN)和 MARL 算法(宽松 Q 学习)。Tampuu 等[45]证明了由自主深度 Q 网络控制的智能体能够从原始感官数据中学习双人视频游戏。

将 DPG 应用到更高维的多智能体环境中时面临的最问题是，环境的不断变化进一步增大了学习的方差。Song 等[46]提出了一种新的多智能体策略梯度算法，该算法解决了通常观测到的高方差梯度估计问题，在粒子高复杂环境中可以有效优化多机器人协作控制任务。Wai 等[47]为了使基于策略的 MARL 更具可扩展性和鲁棒性，提出了一种分散的局部交换方案，其中每个智能体只通过网络与邻居通信。这是一种双重平均方案，其中每个智能体分别在空间和时间上迭代，以分别合并相邻梯度信息和本地奖励信息。Abouheaf 等[48]针对基于图交互的多智能体系统，提出了一种基于策略迭代的在线自适应强化学习方法。该方法利用降值函数求解多智能体系统的耦合贝尔曼方程，在用策略迭代方法进行更新时，考虑了降值函数以降低计算复杂度，解决了大规模优化问题。

还有一些学者开始将基于 AC 框架的算法应用到多智体强化学习中。Open AI 提出的 MADDPG 算法，实质上是 DDPG 算法的一种延伸和扩展。MADDPG 的基本架构与 DDPG 一样，每一个智能体使用自己独立的表演者，通过自己观察状态输出，确定动作，同时训练数据也只能使用自己产生的训练数据。每个智能体同时也对应一个评论家，不同的是这个评论家将同时接受所有表演者的数据，这种中心化的评论家存在多个。该算法基本解决了 DQN 经验重放不再适用和 DPG 方差过大的问题。这种以 AC 为架构的算法与多智能体强化学习的结合是未来研究的一个重要方向。

3.2.2　智能体意

人类物种的成功归功于人们对物质世界和社会环境的著适应性。人类社会智能赋予我们推理其他人心态的能力，这种心理状态推理广泛影响着我们日常生活中的决策。例如，安全驾驶要求我们推断其他驾驶员的意图并做出相应的决定。这种微妙的意图决策(心智理论)行为在人类活动中无